A Minimalist Model of the Artificial Autonomous Moral Agent (AAMA)

Don Howard and Ioan Muntean

The Reilly Center for Science, Technology, and Values.
University of Notre Dame, Notre Dame, IN 46556

{dhoward1, imuntean}@nd.edu

Abstract

This paper proposes a model for an artificial autonomous moral agent (AAMA), which is parsimonious in its ontology and minimal in its ethical assumptions. Starting from a set of moral data, this AAMA is able to learn and develop a form of moral competency. It resembles an "optimizing predictive mind," which uses moral data (describing typical behavior of humans) and a set of dispositional traits to learn how to classify different actions (given a given background knowledge) as morally right, wrong, or neutral. When confronted with a new situation, this AAMA is supposedly able to predict a behavior consistent with the training set. This paper argues that a promising computational tool that fits our model is "neuroevolution," i.e. evolving artificial neural networks.

Introduction

The present model is based on two sets of analogies: (a) the similarity between human agents and non-human agents in respect of morality, and (b) the similarities between human morality and human cognition (inspired by the "virtue ethics" literature, i.e. "skill model of virtues"). In the spirit of (a), two components are relevant: universalizability and replicability of moral agency. The present suggestion is that moral agency can be in principle extended from humans to artificial agents, and that normative agency is situated on a *continuum*, rather than displaying a sharp human/artificial distinction. (Floridi & Sanders, 2004) As analogy (b) suggests, artificial morality is part of the larger framework of artificial intelligence, and topics such as artificial moral agency, cognition, and autonomy, are situated within theories about artificial agency, cognition, autonomy. The opposite view that resists these analogies emphasizes a strong *demarcation* between cognition and morality on one hand, and between artificial and human morality on the other hand.

First, the conceptual inquiry into artificial morality, including the aforementioned analogies, should be explored within ethics, which is one of the most dynamic areas of philosophy. Extending morality beyond the human agent to non-individual, or non-human agents, is, presumably, a major challenge to mainstream ethics. Arguing for or against artificial morality challenges ethics, epistemology, and the new field of "experimental moral philosophy" (Alfano & Loeb, 2014).

Universalizability and Replicability of Morality

Analogy (a) is a timely philosophical issue. For a starter, here are some questions about moral agency: Is there normative agency outside humanity? What kind of agents can be moral—besides the mature, fully conscious, (human) individual? What stops us from extending moral agency to highly-evolved animals, legal entities, groups of people (firms, military command units, political parties, governments, nations), angels, aliens, and, last but not least, computers? Can they receive human justice, friendship, rights, etc.? And, conversely, do they qualify as "moral patients"? A relatively new question added is whether artificial moral agents are *computationally* possible: can we create (implement, simulate, program, etc.) artificial morality?

The complexity of artificial systems is growing rapidly, but it is not the one that has ethical implications: it is the *autonomy* of these systems that bothers us the most. The public is focused on the increase in the complexity of our interaction with machines, and in the autonomy they gain. The complexity of the sociotechnical system is more

important than the complexity of the machine itself.[1] The problems with technology, Wallach writes, "often arise out of the interaction of those components [specific technology, people, institutions, environment, etc.] with other elements of the sociotechnical system" (Wallach, 2015, p. 34)

When it comes to machines, both the public and philosophers are much more skeptical about delegating moral actions than delegating other, non-moral actions. Rejecting in principle artificial moral agency is probably the most straightforward attitude; a philosophical answer is to show that non-human agents cannot "navigate" the kingdom of goods or that they lack moral responsibility, which is for many a necessary condition of moral agency. Where human suffering and human existence are at stake, we prefer to make the decisions ourselves. The case of moral agency is an exceptional case of agency, and arguments are marshalled to show the conceptual impossibility of replacing human moral agents with artificial counterparts. Normative agency is, according to a pre-theoretical and pre-critical attitude called here the "no-go" stance, not suitable to be delegated to artificial agents. Freedom, intentionality, moral responsibility, motivation, and, ultimately, reason, are among the most-cited features needed for moral agency: "Neither the behavior of nature nor the behavior of machines is amenable to reason explanations, and moral agency is not possible when a reason-explanation is not possible." (Johnson, 2006) As Johnson and collaborators argued, computers are always tethered to the humans who created and deployed them, therefore they are not autonomous or actually agents, but surrogate agents. (Johnson&Miller, 2008; Johnson & Powers, 2008) For Moor, full ethical agents (as opposed to explicit, implicit or ethical-impact agents), beings like us, with "consciousness, intentionality, and free will" (Moor, 2006, p. 20)

The no-go stance sees a *delimitation*, rather than continuity, between what is too human to be artificial and what can be delegated to others, be it machines, animals, group of people, companies, etc. As Aristotle would say, we have only "accidental commonalities" with animals, when it comes to morality, justice, rights, etc.

Those who accept the continuum of normative agency raise the *universalizability* issue.[2] Similar to Kant's ethics of duty, we query the meta-ethical possibility to generalize morality beyond the biological, social and cultural limits of the human individual. One conjecture entertained here is that progress in AAMA research will shed some light on our own morality. Understanding the "other," the "different" moral agent, even only by questioning its possibility, is another way of reflecting upon ourselves. Arguing for or against the "non-human and/or non-individual" agent expands the knowledge about the intricacies of our own ethics.

The machine ethics literature goes further by asking whether computers are able to *replicate* our moral actions and be in principle *on par* with human moral agents. If normative agency exists beyond individual humans, can we design and implement agents fundamentally different from us in respect of material constitution, principles of operation, etc.? We call this the *replicability* issue. Unlike universalizability, this issue has fewer philosophical roots and belongs rather to "computationalism." Philosophers have been asking similar questions about the mind for centuries: is it fundamentally a computational process? And, if so, can we implement it? If moralisty is computable, "how-possibly" can we implement an AAMA? What is needed to have artificial moral agents, sufficiently autonomous and complex? What are the fundamental elements on which artificial moral agents should be built? Replicability is not a purely empirical question that can be solved by describing existing agents. It is a "how-possibly" question, which moves the discussion into the realm of possible moral agents, existing or not, and forces us to build computational models of non-human agency.

Most moral naturalists would reject "no-go" arguments against artificial morality as ungrounded. When faced with the *delimitation* problem in agency, the naturalist can argue that the difference is only apparent, and that the category of "essentially human agency" cannot be established on *a priori* grounds. Any strong delimitation depends on multiple factors: in this case, the level of abstraction being the most preeminent one. (Floridi & Sanders, 2004) A community of philosophers, computer scientists, psychologists, etc., who believe that normative agency is not *in principle* exclusively human initiated the field of "machine ethics" a new area at the intersection of philosophy, computer science, cognitive science and psychology (Abney, Lin, & Bekey, 2011; Danielson, 1992; Allen & Wallach, 2009; Anderson & Anderson, 2011; Wallach, 2014; Allen, Varner, & Zinser, 2000) This emergent field raises new interesting ethical issues that go beyond the ethics of emergent technologies: it enquires the ethics of our relation with technology, when humans are *not* the sole moral agents. Although humans are still the main moral decision makers, they share some moral competencies and abilities with artificial agents.

[1] We prefer to talk here about complexity of a system as a measure of its interaction with humans, the environment, etc. and not as a feature of the machine *per se*.

[2] The term of "universalizability" is used here in the context of I. Kant's practical reasoning, who in the *Groundwork of the Metaphysic of Morals* requests that reasons are "universalizable."

Parsimonious and Quietist Machine Ethics: Moral Functionalism

Analogy (b) suggests not a strong delimitation, but a *continuum* between morality and cognition. We use a common conceptual core between cognition and morality: learning and development. The main abilities of our AAMA model are moral learning, moral cognition, and moral development. More specifically, we are interested in setting the ground of a "moral machine learning" premised on learning architectures from recent AI approaches.

Our analogical model is similar to a scientific model: it is supposed to represent partially the human moral agency, through abstractions, "negligibility" assumptions, and idealizations. (Cartwright, 1989; Cartwright & Jones, 2005) Rather than building an artificial replica of the human moral agent, this model reproduces its *behavior*, without representing accurately its constitution, its causal structure. We learn about the world from minimal models in economics, physics, chemistry or biology, even when they lack a perfect isomorphism or even resemblance with reality. (Grüne-Yanoff, 2009) In model-based science, modelers postulate some theoretical entities needed to understand, explain and ultimately predict human agency.

This AAMA model is quietist in respect of its ontology and ethically parsimonious. The quietism opted here for is yet another form of abstraction, akin to a moral functionalism: the modeler takes some entities as metaphysically real, but shuns their importance in understanding, explaining or predicting the behavior of a system: we do not *need* to assume that *all* elements (e.g. consciousness, personhood, empathy) of the human agency have significant explanatory, predictive or representational powers. Similar quietism can be adopted in the case of the limits of AI in Searle's "Chinese room" argument (1980). Second, this model is parsimonious in the sense of assumptions made about moral reasons, motivations, and ultimately moral responsibility. A special regime in our model have the moral principles to be discussed here: there are a number of idealizations about the role and place of moral principles. To summarize, from the perspective of ethics, the AAMA model is closer to moral particularism, rather than moral generalism; and it is more agent-based, than action-based. Inspired by minimal models in economics, we call our project the minimalist AAMA model. Other authors in machine ethics may relate this proposal to an "ethical nihilism." (Beavers, 2011)

The model minimizes the *a priori* elements put in by *fiat*: any predefined rule-based computation; assumptions about moral intentions in general; any metaphysical load about human agency (personhood, empathy, free will, etc.). A machine that memorizes a set of principles and blindly applies them to any case is not autonomous. But nothing stops our AAMA from discovering, *a posteriori*, moral regularities and patterns in existing data. Such an AAMA may output important aspects of moral principles, motivations, freedom of human agent, etc., as *patterns* and post-rationalization elements. This minimalist model is closer to a bottom-up, rather than the top-down architecture. The ideal situation is to seek the right balance between the top-down and bottom up approaches, and produce a "hybrid model." (Allen, Smit, & Wallach, 2005; Allen & Wallach, 2009) Our hybrid model remains a data-oriented, rather than a theory-oriented model: the AAMA "reads" the moral data collected from human agents and learns patterns in the data from it. But the AAMA needs to be able to surpass simply the memorization of moral behavior of humans, avoid overfitting the data and be able to generalize from it, enough to produce new predictions about more complex behavior than the training set. The AAMA therefore, simply put, performs an inductive reading of the moral behavior of human agents.

As some virtue ethicists insist (Nussbaum, Annas, *i.a.*), a moral judgment is closer to classification of perceptions, than to reasoning from general to particular. The virtuous person makes right moral judgments on a case-by-case basis: the person of principle is prone to make bad moral decisions because she has the "tendency not to look hard enough at the details of the case before one." (Gleeson, 2007, p. 369) For the moral particularist, this entails that moral judgments need a form of moral sensibility to the context and content of each case. (Dancy, 2006) A moral expert "just sees" which features of a case are morally relevant and what action is needed. The agent cannot or does not need to follow a moral principle. The moral expert develops a perceptual-like moral competence by exploring a set of "prototypes, clearest cases, or best examples." (Dancy, 1999, p. 70) The process of categorization of moral cases is similar to the perception in which the trained observer is able to classify and categorize new information.

Another way to evade the problems of a rule-based AAMA is to rely more on semantic naturalism in ethics, and on "moral functionalism." (Horgan & Timmons, 2009; Jackson, 1998; Jackson & Pettit, 1995; Zangwill, 2000; Danielson, 1992, 1998) This model emphasizes the role of the functional and behavioral nature of the moral agent: its decision, its output state, are functional in nature, individuated by its dependence on the input, the previous output (this is nothing more than a form of "memory") and other, current or previous, moral states. As an addition to existing functionalism approaches, we add the dependence on agent's dispositions. Enter moral dispositional functionalism, deemed as more appropriate for a discussion on universalizability and replicability of normative agency.

Moral Cognition, Dispositional Virtues and Patterns in Data

The present AAMA model is premised on *some* common elements between human and artificial moral agents. We emphasize the quantifier *some*: if we "read off" the *whole* ethics of an AAMA from the constitution and function of human agents, as imperfect as they are, we risk to restrict moral agency and make it too anthropocentric. (or evolution-centric, or carbon-centric, or mind-centric, or person-centric, for that matter) As we propose here a minimalist model, we want to constrain moral agency as little as possible. Because of replicability, we want to use the right computational tools for machine ethics.

The minimalist AAMA model is different than a simulation or an emulation of a human moral agent. What matters is the "how-possibly" explanations that assess possible instances of AAMA, rather than existing instances. To get a grasp of our model, we consider a minimal core of common elements shared by AAMA and human moral agents. Inspired by recent results in neuroscience, some philosophers pushed towards a "neuroethics" approach to morality. (P. S. Churchland, 2011) The human ethics is arguably a multi-dimensional process hard to capture by definitions, but there are a couple of components from Churchland's list that we include in this AAMA model: (1) the search for moral optimality, as a specific case of problem solving in social context; and (2) the learning and development of moral practices from exemplar cases, which include mechanisms of reinforcement, imitation, trial and error, conditioning, etc.

Artificial moral cognition is a process of developing moral dispositions, instead of learning moral rules. It also presupposes that the nature of moral agency is of a dispositional nature, and not categorical properties. This takes us closer to the framework of virtue ethics. The present paper emphasizes the connection between general cognition and moral cognition as a virtue, very similar to the "skill model" of virtue ethics. (Annas, 2011, 2015) The AAMA model incorporates innovative and creative aspects of moral cognition from virtue ethics, hard to implement in other rule-based or pure utilitarian models.

The link between virtue ethics, hybrid AAMA models and connectionism has been suggested in (Danielson, 1992; Gips, 1995) and more recently by Allen and Wallach (2009, Chapter 6). Virtues of AAMA illustrate better the hybrid architecture, than models based on duties or utilitarian calculations. Nevertheless, the suggestion of pre-programming into the AAMA virtues runs in the same troubles as pre-programming principles, maxims or fixed utilitarian calculations. Another suggestion hinted to by Allen and Wallach is to link virtues to social functions or to the very process of developing moral expertise through interactions.

The present model diverges in some respects from Allen and Wallach. The type of virtue ethics used here depends on the analogy used between cognition and moral cognition. The moral learning process is practical: what we learn to do, we learn by repeating. It is a practice and the skill model of virtues emphasize this similarity: "the structure of a virtue like bravery is illuminatingly like the structure of a practical skill, like playing the piano. You need to learn it from other people, but you need to learn how to do it for yourself. It's no good just mimicking your teacher, as that would turn you into a clone of your teacher." (Annas, 2015, p. 3)

Our assumption here is that regularities in moral data, as a form of patterns, can play the role of moral norms. But they are *discovered*, rather than postulated. The agent may operate based on principles, or depending on its moral expertise, or operate based on a set of dispositional traits acquired previously. Ethical decisions are taken when "information is unclear, incomplete, confusing, and even false, where the possible results of an action cannot be predicted with any significant degree of certainty, and where conflicting values … inform the decision-making process." (Wallach, Franklin, & Allen, 2010, p. 457) Moral cognition is in fact more complicated that it might appear: lack of certainty, noise, error, ambivalence, and fuzziness are all features of the content of moral cognition.

This minimalist model assumes that numerical data represents *adequately enough* moral behavior: the moral behavior can be gathered as variables (numerical, nominal, categorical, etc.) and morality can be quantified by moral behavior. The moral decision making process is ultimately a computational process, similar to the classification of complicated patterns, playing games, or discovering strategies, creating new technologies, or advancing new hypotheses in science.

Second, it is assumed that data representing moral behavior is regular enough and that it exhibits "patterns." It is trite to say that moral data set are complex: but complexity here means a large number of degrees of patterns, not mere randomness (Shalizi & Crutchfield, 2001). Patterns, unlike rules or principles, include non-linearity, emergence, errors, noise, irrelevant data. More or less metaphorically, one can talk about a complex pattern as superposition and entanglement of symmetries and regularities. It is postulated here that patterns exist in moral data and they can be taught from data. The complexity of moral data makes us think of the AAMA as highly adaptive and highly responsive system which needs to optimize its search procedures (Holland, 1975; Ladyman, Lambert, & Wiesner, 2012; Mitchell, 2012). The ideal AMA is then characterized by a high degree of robustness under perturbation, noise, missing data, and other outer

influences. It is able to detect the right pattern in the data in the right amount of time and using the right amount of resources. A conceptual work (not reproduced here) is needed to differentiate among patterns in data, regularities and exceptionless rules (similar to a powerful generalization, versus a "law of nature" are in natural sciences). What is assumed here is that for any trustworthy and well-formed set of "moral data" there is a pattern of behavior which may or may not include rules.

Why is this model similar to the "skill model" of virtue ethics? The way we acquire skills is possibly an answer. The existence of regularities in the environment and the feedback we receive in the process of learning are two conditions of the acquisition of skills in some moral psychology theories (Kahneman, 2013). Feedback and practice are elements of moral learning from unknown and possibly very complex patterns, with element of random answers, noise and error. Moral behavior is then an unknown function, with an unknown number of variables and constraints. Although there is always missing information about the variables of moral behavior, the AAMA is able to pick relevant patterns of dependency among relevant variables.

The minimalist nature of our model couples well with particularism in ethics: it is quietist about the importance of laws, principles and. Autonomy of AAMA is understood here as independence from pre-imposed, exceptionless rules. Particularism and functionalism do not eliminate principles, but downgrade their importance in the design and understanding of the AAMA. One advantage of the present approach is the local nature of moral competence. It is not an aim of our AAMAs to be global: the model is not trained to solve all moral dilemmas or make all moral decisions. As a parsimonious model, this is far for being a "theory of everything" in artificial ethics. The minimalist AAMA model is less scalable than we, humans. The particularist designer of AAMA teaches and trains them for domain-specific moral responsibilities. Thus, ethics programming for a patient-care robot needs not include all of the kinds of competencies that would be required in a self-driving car or an autonomous weapon. We know that generalists are at great pains to show how moral principles unify a vast class of behaviors and roles the agent has to play, but this is not the case with our approach.

The generalist can always retort that there are principles of ethics which are not manifest in this or that set of data. Data may or may not unveil regularities, depending on the way we collect it: not all data can be turned into evidence for this or that scientific theory. One possible answer to the generalist is that for all practical purposes, the set of "rules" or principles are complicated, conditionalized, constrained, and too hard to be expressed in computational algorithms.

One approach that uses simple recurrent neural networks, but not evolutionary computation, is M. Guarini's (2006, 2012). He trained artificial neural networks on a set of problems similar to "X killed Y in this and this circumstances" and managed to infer (predict) moral behaviors for another set of test cases. Guarini's conclusion runs somehow against moral particularism, because some type of moral principles is needed (including some "contributory principles"), but it also shows that particularism is stronger than it seems. Unlike Guarini, this model incorporates moral functionalism and ethical particularism: principles are not impossible or useless to express, but they do not play the central role in the design of this AAMA. The focus is on the moral development and moral expertise that the successful AAMA is prone to achieve.

The Evolutionary Neural Networks and the Turing Moral Test

Given a set of variables and data about the human agent, an action of a human agent receives a moral quantifier: this is what we call here a "moral classification problem."[3] The AAMA is taught how to classify cases from data such that the results obtained is similar to a moral inductive reasoning: new classifications are inferred from previous learning processes in which the AAMA is instructed on similar classification problems, typically simpler and paradigmatic. The moral competence of any agent is context-sensitive and vary among different communities of human agents and within the cases at hand. This proposal assumes that ideally a moral competence brings in some unity of behavior, but, more pragmatically, it brings in flexibility, adaptation, and ultimately the ability of reclassification. Two thought experiments in dispositional moral functionalism are perhaps worth mentioning: a moral Turing-like test and a moral Chinese-room type of experiment. (Allen et al., 2000; Beavers, 2011; Bechtel, 1985; Sparrow, 2011) In discussing what is needed to pass a Turing-like test, Beavers concedes (Beavers, 2011, p. 340): "Though this might sound innocuous at first, excluded with this list of inessentials are not only consciousness, intentionality, and free will, but also anything intrinsically tied to them, such as conscience, (moral) responsibility, and (moral) accountability."

As Allen and Wallach suggest, for the case of artificial moral agents, we may end up with a minimal "moral Turing test" as a decision procedure about the performance of an *explicit* AAMA, given a set of moral data about human behavior. (Allen & Wallach, 2009, Chapter 1) Is a moral Turing-like test setting the bar too low? The

[3] We collect moral data by surveys, inspired by the methodology of moral psychology.

dispositional performance of any moral agent can be evaluated against a large number of testing sets, and not on their internal components or internal (sub-) mechanisms. In this setup, the moral data, produced statistically from a population of human subjects, act as the training set and codes the "folk morality" behavior. This AAMA is nothing more than a "behavioral reading", as opposed to the differing, but more popular "mindreading" hypothesis of morality. (Michael, 2014) The concept of "behavioral reading (moral) agent" is indeed part of the bare ontology of our AAMA model. The behavioral reader is any "being" able to detect and learn from behavioral patterns, able to associate them to previous experiences, categorize them, and employ those accomplishments to replicate the behavior of humans. (Monsó, 2015) The functional agent is not absolutely right or wrong, morally, but better or worse than other agents, given a moral behavior dataset. The well-trained AAMA will be able to find the right action more quickly and more efficiently than the novice AAMA, in the same way in which a trained observer is able to classify quickly an object of perception than the untrained one. The process itself is close to learning from data, but can be better described as self-building, or possibly self-discovering, patterns in data. The decision making by an AAMA is represented as a search procedure in a space of possible moral actions; it evolves in a space of possible solutions and stops when a solution "good enough" is found. The learning component warrants that the result depends on a level of moral expertise that the machine has acquired in the past, i.e. during training. The evolutionary computation is mainly designed to obtain the global extreme point of such a search procedure.

We use a computational tool for "pattern recognition structure:" the "evolving artificial neural networks," employing a population of neural networks (*NN*) and evolutionary computation (*EC*), called here *NN+EC*. We take neural networks as the natural candidates for artificial moral learning. Outside morality, neural networks are able to classify complicated patterns in the same way the brain is able to recognize and train on finding patterns in perceptual data. Philosophers hinted towards an associationism approach to human morality. The association-based model of P. Churchland de-emphasizes the rules, and reconstructs moral cognition as a classification activity based on similarity relations. Churchland writes: "What is the alternative to a rule-based account of our moral capacity? The alternative is a hierarchy of learned prototypes, for both moral perception and moral behavior, prototypes embodied in the well-tuned configuration of a neural network's synaptic weights." (P. Churchland, 1996, p. 101) For Gips, the right way of developing an ethical robot is to "confront it with a stream of different situations and *train* it as to the right actions to take." (Gips, 1995) This is strikingly similar to what we do with neural networks that are trained to recognize faces, voices, and any type of patterns in data. Others have discussed the advantages, and disadvantages, of assimilating learning and the agent-based approaches in machine ethics (Abney et al., 2011; Allen & Wallach, 2009; Tonkens, 2012).

Take the properties of neural networks as natural properties: for a given moral function, there is a (large) set of topologies, training functions, biases, architectures, etc. that may instantiate it. As the focus is on the input-output relation, moral mechanisms are multiply instantiated.

A second, albeit central, advancement of the present model is the conceptual overlapping among neural networks, evolutionary computation. The present model is inspired by "computational intelligence," a paradigm in computation usually contrasted to "hard computing." (Adeli & Siddique, 2013; Mitra, Das, & Hayashi, 2011; Zadeh, 1994) In "hard computing", imprecision and uncertainty are aspects of a system or process to be avoided at any price. Although computable, computational ethics does not belong to the domain of certainty and rigor, as there is no complete knowledge of moral or ethical content. (Allen et al., 2005)

Therefore, the EC component endows the AAMA with more moral autonomy from the initial assumptions built in the first population of NNs. If NNs depend heavily on data, successive generations of AAMA, evolved through EC, are gradually able to separate mere noisy data from evidence. At the next stage, after training is over, and the AAMAs are deemed as "good enough", the population of networks are presented with a case outside the training set. The answer is presumably the moral decision of AAMA for that case which is shared with other AAMA and with the human trainers. Depending on the complexity of the problem, the answer of the population of AAMA can or cannot be in the "comfort zone" of the trainers.

The AAMA Model and the NEAT Architecture

In the more advanced implementation of this AAMA model, we plan to move from training networks with fixed topologies to evolving topologies of neural networks by implementing one or more of the NEAT (NeuroEvolution of Augmenting Topologies) architectures: Hyper-NEAT (CPPN), MM-NEAT, and SharpNEAT (Evins, Vaidyanathan, & Burgess, 2014; Richards, 2014; Stanley, D'Ambrosio, & Gauci, 2009; Stanley & Miikkulainen, 2002).[4]

In the NEAT architecture, a population of m networks with fixed weights and constant parameters are generated. The population is divided from the second generation onwards

[4] For more information about the NEAT and its versions developed at the University of Texas at Austin, see: http://nn.cs.utexas.edu/?neat

into *species*, based on topologies. When a termination condition is met, an individual from a given species is retained as the elite performer. This is the general algorithm that a version of NEAT can use:

1. Create an initial population of NNs with a fixed topology (no hidden layers and no recursive functions) and fixed control parameters (one transfer function, one set of bias constant etc.).

2. Evaluate the fitness for each NN.

3 Evaluate the fitness of each species and of the whole population and decide which individuals reproduce and which interspecies breeding is allowed

3' CPPN: change the transfer function for a species with a composition of function (CPPN).

4. Create a new population (and species) of NN by EC: new weights, new topologies, new species

5. Repeat 2–4 till convergence fitness is obtained, or a maximum number of generations is reached, or the human supervisor stops the selection process.

This AAMA model suggests that dispositional virtues are properties of *evolved populations* of *NNs*, and that they are not represented by properties of individual *NN*s. As during the evolution process features of each species are evolved and selected, one can see that in evolutionary computation the moral competence becomes distributed within the population: the transfer functions of different NNs (which defines here the CPPN component of the Hyper-NEAT), their topologies or architectures (parts of NEAT), the weights, etc. The AAMA agent is the one able to develop a minimal and optimal set of virtues that solves a large enough number of problems, by optimizing each of them. In the *NN+EC* design, the dispositional virtue of the AAMA agent resides in the ability a population of networks to generalize from data to new input. This is codified by topologies, weights, transfer functions which are in this setup results of evolution. Moral cognition here is obtained by the EC process operating on NN, in which mutation and recombination of previous generations produce new, unexpected individuals.

This choice for NEAT has a foundational consequence for this AAMA model. Going back to the questions asked at the beginning, the moral learning and the moral behavior of the AAMA are at the end of the day a learning and training process similar enough probably to husbandry and domesticating: probabilistic processes, which can be improved up to a certain degree, but with no guaranteed success for a given individual animal trained. This AAMA is also similar enough to the learning and training process of human subjects: no learning is decisive and guaranteed. There will be always bad results and "too good to be true" results, but the conjecture here is that in the long run, such an AAMA model or something similar to it will gain "model robustness." Even for the best generation of AAMAs there is always "a kill switch," should there be any reason to worry about the emergence of morally problematic behaviors.

Some Loose Ends and a Conclusion

One line of criticism to which this model is vulnerable is its dependency on data. How reliable is moral data? How do we collect data? We simply rely here, probably uncritically, on the methodology of moral psychology or experimental philosophy, on surveys on significant samples from the adult population of a given culture. Given its EC design, the model is able to detect inconsistencies in moral data.

As any result of evolution, for very complex problems, the outcome is less accessible to deterministic, or "hard computing" algorithms. This hints towards a problem of tractability with our model, and relatedly, to modularity. In some cases, the complexity of the evolution "screens-off" the initial conditions, be them topologies or initial assumptions about the transfer functions. A delicate balance between the mechanisms of selection that decrease variation and those that increase variation (mutation) is needed. At the limit, the solution of such an algorithm may be inscrutable to humans: but the very existence of the inscrutable solution depends on the complexity of the problem, and on the very dynamics of the NN+EC schema.

Another line of criticism against this model is its ethical simplicity: one can doubt that morality of human agents can be reduced to patterns in their moral behavioral data in the same way in which visual patterns compose an image. Where is moral reasoning, where is moral responsibility in this proposal? What is ultimately ethical about this AAMA? There is no direct answer to this line of criticism, but two suggestions are worth mentioning. First, one can apply a multi-criteria optimization in which the fitness function has two outcomes: a non-moral and a moral outcome. Second, the EC component can include a penalty factor for solutions which are not moral enough, based on a selection of species of NN which prioritize the normative variables in the data over the factual variables. Another answer to this later criticism is that morality can be emergent in AAMA and not a fundamental feature of the system. In this model, morality is a statistical feature of populations of networks, rather than a property of one network.

By using evolving neural networks, a certain degree of innovation and creativity is reached by this bare model, hard to find in the other rule-based, generalist, or action-centered models.

References

Abney, K., Lin, P., & Bekey, G. A. (Eds.). 2011. *Robot Ethics: The Ethical and Social Implications of Robotics*. The MIT Press.

Adeli, H., & Siddique, N. 2013. *Computational Intelligence : Synergies of Fuzzy Logic, Neural Networks Intelligent Systems and Applications*. Somerset, NJ, USA: John Wiley & Sons.

Alfano, M., & Loeb, D. 2014. Experimental Moral Philosophy. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2014. Retrieved from http://plato.stanford.edu/

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

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

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.

Annas, J. 2011. *Intelligent virtue*. Oxford University Press.

Annas, J. 2015. Applying Virtue to Ethics. *Journal of Applied Philosophy*.

Beavers, A. F. 2011. Moral Machines and the Threat of Ethical Nihilism. In K. Abney, P. Lin, & G. A. Bekey (Eds.), *Robot Ethics: The Ethical and Social Implications of Robotics* (pp. 334–344). Cambridge, Mass: The MIT Press.

Bechtel, W. 1985. Attributing Responsibility to Computer Systems. *Metaphilosophy*, *16*(4), 296–306.

Cartwright, N., & Jones, M. (Eds.). 2005. *Idealization XII: Correcting the Model: Idealization and Abstraction in the Sciences*. Rodopi.

Churchland, P. 1996. The neural representation of the social world. In L. May, M. Friedman, & A. Clark (Eds.), *Minds and Morals* (pp. 91–108).

Churchland, P. M. 2000. Rules, know-how, and the future of moral cognition. *Canadian Journal of Philosophy*, *30*(sup1), 291–306.

Churchland, P. S. 2011. *Braintrust: What Neuroscience Tells Us about Morality*. Princeton: Princeton University Press.

Dancy, J. 1999. Can a Particularist Learn the Difference Between Right and Wrong? In Hintikka, Jaakko & E. Sosa (Eds.), *The Proceedings of the Twentieth World Congress of Philosophy* (Vol. 1, pp. 59–72). Boston, MA, USA.

Dancy, J. 2006. *Ethics without Principles*. Oxford; New York: Oxford University Press.

Danielson, P. 1992. *Artificial morality virtuous robots for virtual games*. London; New York: Routledge.

Danielson, P. (Ed.). 1998. *Modeling rationality, morality, and evolution*. New York: Oxford University Press.

Evins, R., Vaidyanathan, R., & Burgess, S. 2014. Multi-material Compositional Pattern-Producing Networks for Form Optimisation. In A. I. Esparcia-Alcázar & A. M. Mora (Eds.), *Applications of Evolutionary Computation* (pp. 189–200). Springer Berlin Heidelberg.

Floridi, L., & Sanders, J. w. 2004. On the Morality of Artificial Agents. *Minds & Machines*, *14*(3), 349–379.

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.

Gleeson, A. 2007. Moral Particularism Reconfigured. *Philosophical Investigations*, *30*(4), 363–380.

Grüne-Yanoff, T. 2009. Learning from Minimal Economic Models. *Erkenntnis (1975-)*, *70*(1), 81–99.

Guarini, M. 2006. Particularism and the Classification and Reclassification of Moral Cases. *IEEE Intelligent Systems*, *21*(4), 22–28.

Guarini, M. 2012. Conative Dimensions of Machine Ethics: A Defense of Duty. *IEEE Transactions on Affective Computing*, *3*(4), 434–442.

Holland, J. H. 1975. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. University of Michigan Press.

Horgan, T., & Timmons, M. 2009. Analytical Moral Functionalism Meets Moral Twin Earth. In I. Ravenscroft (Ed.), *Minds, Ethics, and Conditionals*. Oxford Univ Pr.

Jackson, F. 1998. *From metaphysics to ethics a defense of conceptual analysis*. Oxford: Clarendon Press ; New York.

Jackson, F., & Pettit, P. 1995. Moral Functionalism and Moral Motivation. *The Philosophical Quarterly*, *45*(178), 20–40.

Johnson, D. G. 2006. Computer systems: Moral entities but not moral agents. *Ethics and Information Technology*, *8*(4), 195–204.

Johnson, D. G., & Miller, K. W. 2008. Un-making artificial moral agents. *Ethics and Information Technology*, *10*(2-3), 123–133.

Johnson, D. G., & Powers, T. M. 2008. Computers as Surrogate Agents. In M. J. van den Joven & J. Weckert (Eds.), *Information Technology and Moral Philosophy* (p. 251). Cambridge University Press.

Kahneman, D. 2013. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Ladyman, J., Lambert, J., & Wiesner, K. 2012. What is a complex system? *European Journal for Philosophy of Science*, *3*(1), 33–67.

Michael, J. 2014. Towards a Consensus About the Role of Empathy in Interpersonal Understanding. *Topoi*, *33*(1), 157–172.

Mitchell, S. D. 2012. *Unsimple Truths: Science, Complexity, and Policy*. Chicago, Mich.: University Of Chicago Press.

Mitra, S., Das, R., & Hayashi, Y. 2011. Genetic Networks and Soft Computing. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *8*(1), 94–107.

Monsó, S. 2015. Empathy and morality in behaviour readers. *Biology & Philosophy*, *30*(5), 671–690.

Moor, J. H. 2006. The nature, importance, and difficulty of machine ethics. *Intelligent Systems, IEEE*, *21*(4), 18–21.

Richards, D. 2014. Evolving Morphologies with CPPN-NEAT and a Dynamic Substrate (pp. 255–262). School of Computing, Mathematics & Digital Technology, Manchester Metropolitan University: The MIT Press.

Searle, J. R. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(03), 417–457.

Shalizi, C. R., & Crutchfield, J. P. 2001. Computational mechanics: Pattern and prediction, structure and simplicity. *Journal of Statistical Physics*, *104*(3-4), 817–879.

Sparrow, R. 2011. Can Machines Be People? Reflections on the Turing Triage Test. In K. Abney, P. Lin, & G. A. Bekey (Eds.), *Robot Ethics: The Ethical and Social Implications of Robotics* (pp. 301–315). Cambridge, Mass: The MIT Press.

Stanley, K. O., D'Ambrosio, D. B., & Gauci, J. 2009. A hypercube-based encoding for evolving large-scale neural networks. *Artificial Life*, *15*(2), 185–212.

Stanley, K. O., & Miikkulainen, R. 2002. Evolving Neural Networks through Augmenting Topologies. *Evolutionary Computation*, *10*(2), 99–127.

Tonkens, R. 2012. Out of character: on the creation of virtuous machines. *Ethics and Information Technology*, *14*(2), 137–149.

Wallach, W. 2014. Ethics, Law, and Governance in the Development of Robots. In R. Sandler (Ed.), *Ethics and Emerging Technologies* (pp. 363–379).

Wallach, W. 2015. *A Dangerous Master: How to Keep Technology from Slipping Beyond Our Control*. Basic Books.

Wallach, W., Franklin, S., & Allen, C. 2010. A Conceptual and Computational Model of Moral Decision Making in Human and Artificial Agents. *Topics in Cognitive Science*, *2*(3), 454–485.

Zadeh, L. A. 1994. Fuzzy Logic, Neural Networks, and Soft Computing. *Commun. ACM*, *37*(3), 77–84.

Zangwill, N. 2000. Against Analytic Moral Functionalism. *Ratio: An International Journal of Analytic Philosophy*, *13*(3), 275–286.