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.

Adventures in Space Racism: Going beyond the Turing Test to determine AI moral standing

by Nicholas A. Novelli

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba in partial fulfilment of the requirements of

the degree of Master of Arts

Department of Philosophy

University of Manitoba

Winnipeg

Abstract: In pop culture, artificial intelligences (AI) are frequently portrayed as worthy of moral personhood, and failing to treat these entities as such is often treated as analogous to racism. The implicit condition for attributing moral personhood to an AI is usually passing some form of the "Turing Test", wherein an entity passes if it could be mistaken for a human. I argue that this is unfounded under any moral theory that uses the capacity for desire as the criteria for moral standing. Though the action-based theory of desire ensures that passing a rigourous enough version of the Turing Test would be sufficient for moral personhood, that theory has unacceptable results when used in moral theory. If a desire-based moral theory is to be made defensible, it must use a phenomenological account of desire, which would make the Turing Test fail to track the relevant property.

Acknowledgements: Funding from the University of Manitoba over the last two years has made this research possible. I would also like to thank the following people:

The best thesis committee a philosopher could ask for in Rhonda Martens, Rob Shaver, and Sarah Hannan. It was truly a pleasure to work with them, and their help and advice was invaluable.

The other fantastic professors in the U of M philosophy department who have helped and supported me, especially Joyce Jenkins, who has contributed a great deal to my command of ethics from the very beginning of my philosophical career, and Carl Matheson, who has always been there in the nick of time when I needed something.

My fellow students, for conversations that have given me intuitions and insights which have proven tremendously useful, and that were tremendously enjoyable as well.

My family, for their support and for listening to all this in its unrefined state, and in particular my sister Tori, for my title and for the late-night *Star Trek: Voyager* marathons that inspired it.

Table of Contents

Chapter I: Introduction

As technology advances, it becomes increasingly crucial that we explore the question of the moral status of Artificial Intelligences (AI). We already have computers with unbelievable computing power, and before long we might have machines far more "intelligent" than any human. We need to know how we ought to act towards such entities to avoid acting immorally. How advanced must machines become before we should treat them as having moral standing – or even as moral persons? Already we have machines with conversational ability nearly equal to a human's – under certain conditions, computers have already passed the Turing Test (or so it is claimed).[1] Even if we deny that they have passed quite yet, there is little doubt that they will do so soon.

Turing's test was originally proposed as a test of cognition: if a machine is indistinguishable from a human in casual conversation, it can think.[2] John Searle's Chinese Room thought experiment[3] attempts to show that even if a computer can communicate in such a way that it could convince an observer that it is human, the internal processes behind it might not be the sort that

1 See: Aron, Jacob. "Software tricks people into thinking it is human". New Scientist Magazine (September 2011), and "Computer AI passes Turing test in 'world first'". BBC News Website, June 9 (2014). http://www.bbc.com/news/technology-27762088
2 See Turing (1950).
3 See Searle (1980).

produce actual cognition. There have been many objections to Searle's argument that attempt to show that the Turing test is in fact a reliable indicator of cognition.[4]

But whether a machine can think or not gives us no moral guidance by itself. It might be relevant, if we adopt a moral theory where ability to think grants moral standing, but that is hardly universally accepted. We might instead claim, as Jeremy Bentham did about animals, that the question is not if machines can reason, but if they can suffer.[5] Without having an idea of what it would take for an AI to be deserving of moral consideration, we have no way to even begin to determine if we are acting correctly. As more and more sophisticated robots become integrated into our daily lives, we become exposed to greater degrees of moral risk. There is a chance that we may make machines with moral standing without realizing it, and proceed to harm them impermissibly without even being aware there is anyone to harm. Making robots to do difficult, dirty, and dangerous jobs in the place of humans might turn out not to be benevolent, but in fact be tantamount to slavery. Conversely, we might make machines that seem so "real" that we instinctively begin giving them moral consideration they do not deserve, to the detriment of humans and non-human animals that do have moral worth.

4 For example, see Minsky (1980) and Churchland and Churchland (1990).
5 See Bentham (1823), XVII.6 (footnote 122).

People might prefer to assist and benefit androids designed to be pleasant over humans that might be abrasive or socially inept, and might opt to do so when a choice arises. We need a way to tell whether we are acting appropriately.

Many people seem to believe that when it comes to AI, we are justified in operating on the basis of a sort of "Moral Turing Test" – if a machine can communicate and interact in a way that seems fully human, some would say that we are thereby justified in treating it exactly the same as we would a human. This seems to be the attitude we are expected to adopt for many pop culture depictions of AI – when an artificial intelligence character is present in a fiction, the other characters generally treat them as people, the same way they treat human characters.[6] In fact, this is often treated as morally required – characters that discount the personhood and moral importance of AIs are often portrayed as being insensitive bigots, and frequently an allegory for racism is present. Many of the things one might say to deny equal treatment to robots on the grounds of their material composition are taken to be direct parallels to the things one might say to deny equal treatment to other races merely on the basis of the colour of their skin.

6 For a few examples, consider Data and the holographic doctor from *Star Trek: The Next Generation* and *Voyager* (respectively), Holly and the holographic version of Rimmer from *Red Dwarf*, and Bender from *Futurama*.

In fact, positions like this are even expressed in some of the philosophical literature on Artificial Intelligence. Rob Sparrow, in criticising this use of the Turing Test for moral decisions, does not find fault with the basic approach of relying on our perceiving a machine to be human, but only says that we are not being strict enough in applying that criteria, and suggests a "Turing Triage Test" to ensure that we are adopting a high enough standard – a machine would have to seem completely human for us to be tempted to choose its "life" over an actual human's, so that should be the true test, according to Sparrow.[7] But the basic methodology of relying on the degree to which a machine outwardly seems human in its appearance and behaviour is not questioned.

However, seeming human is not the criterion for moral personhood (or any other level of moral standing) in any commonly-held moral theory. Very little of the philosophical literature on AI explicitly states which moral theory is being relied on, but it seems important to have some moral theory in mind in order to evaluate the moral arguments about Artificial Intelligences. Only then can we know which properties a machine would have to possess to be worthy of moral consideration. A complication is that many of the properties that form the basis of moral standing in the various candidate theories are

7 See Sparrow, (2004) and (2012).

not directly detectable. We might be justified in using the Moral Turing Test as a practical test for moral standing if we have some reason to believe that it will track the actual presence of the relevant moral properties. Whether that's the case depends on which properties are significant, and how we define them.

One property that occurs in a large number of moral theories is the capacity to have desires. Many theories of welfare include desire as part of the definition of well-being[8] – what is good for a person might be to get the mental states she desires, or it might be that the states of affairs that she desires are actualized. If we adopt any consequentialist moral theory based on such a theory of welfare, then to have moral standing, an entity would have to possess the capacity to have desires.[9] There are also many deontological theories where benevolence will be one of several moral duties,[10] which could take the form of an obligation to contribute to the welfare of any being capable of having welfare. These theories might use the same criteria to determine which entities merit consideration under that duty. There are of course other types of moral theory – non-welfare-based

8 See Parfit (1984) for a thorough catalogue of the candidate theories of well-being – unless we adopt an objective standard of a good life, all other options include what someone desires as an element of well-being.

9 Note that this includes not just theories where we maximize welfare, but ones where we help the least well off, perfectionist theories where we seek to have some individuals living the best possible life, and many other ways we might calculate welfare-based moral duties.

10 W. D. Ross' theory being but one example.

theories might assign moral standing only to rational agents, where being rational might be defined as requiring actions being based on some interaction of beliefs and desires. Other moral theories require us to respect the autonomy of autonomous individuals; being autonomous might involve having desires, and respecting the autonomy of others might mean not going against their desires. Capacity to have desires would then be a requirement to have moral standing in these theories as well. However, my focus will not be on those theories, and I will generally address my arguments to welfare-based consequentialist theories and deontological theories that include a welfare-based duty of benevolence, though the arguments will frequently be applicable to the other theories as well. Under the theories being considered, it is usually possible to have degrees or hierarchies of moral standing, where entities can have less moral importance than humans while still meriting consideration. I will examine whether AIs will be capable of having any degree of moral standing at all, not just whether they might reach the level of personhood.

Theories such as these that treat the capacity for desire as a criteria for moral standing are fairly popular and widely accepted, though there are viable competitors. Moral theories that do not make reference to desire would require a separate examination to determine whether the Moral Turing

Test might be justified for them, which I do not have space for here. I will not argue for any particular moral theory here, but only for the conditional that if we were to accept one of the theories that relies on desire, the Moral Turing Test would not be a reliable guide to moral standing. I will not discuss what it is about desire that makes it an intuitive candidate as a morally relevant property, but will assume that any property picked out by a theory of desire has at least some prima facie appeal for that role. For the sake of argument, assume the truth of some desire-based moral theory (the exact details are left to the discretion of the reader).

There are many controversies about desire, and I cannot examine all of these questions in detail here. For the sake of expedience I will make certain presuppositions about desire. For one, I will take for granted that desires have propositional content. Most theorists hold that desires must be desires *for* something – directed at some proposition or state of affairs. They are thus representational in the contentive sense.[11] This is widely, though not universally, accepted.[12]

Another slighty controversial assumption I will make is that desires do in fact

11 Though not in the indicative sense: desires have propositional content, but do not attempt to indicate anything about the way anything is in the world – see Schwitzgebel (1999).

12 Thagard (2006) is one of the few dissenters.

objectively exist. Some philosophers have argued that there is no justification for believing in the existence of desires in the first place – Paul and Patricia Churchland claim that belief in desires is equivalent to mere superstition.[13] If this were true, we would not be justified in ascribing moral standing even to humans on the basis of moral theories that depend on desires. If that were the case, we would have to reject all desire-based moral theories. However, Andy Clark has argued that there must be some property we are detecting, simply for evolutionary reasons – so many of our decisions are based on the assumption of desires that if we were wrong, surely we would not be so successful.[14] We will proceed under the assumption that humans have something that can be identified as desires, and we can usually trust their reports and our intuitions of when they have them. An interesting question, however, is the status of desires in non-human animals. Our intuitions are less clear and less universal in that case, but the implications of our theory of desire for the moral standing of creatures ranging from amoebae to chimpanzees should be taken into account in our evaluation of the theories' plausibility.

But what is it to have a desire? Everyone agrees that desires tend to have certain features, and there are many familiar cases where it would be

13 Churchland (1979) and Churchland (1981)
14 Clark (1987).

agreed that a desire is present, and in these cases the same types of things generally happen. A person desires to eat a sandwich: he thinks eating a sandwich would be good, so he is motivated to make a sandwich and eat it, and he feels satisfaction when he does so. But which feature is constitutive of desire, and which only tend to accompany desires? Which feature, if it were absent from a state of affairs, would lead us to conclude no desire was present? For the most part, the literature on these moral theories tends to be concerned only with humans, so questions about the nature of desire have not been very thoroughly explored in terms of their implications for moral standing. It seems like we can reliably identify desires in humans, and it is difficult to settle on a definition of desire precisely because all the features argued to be constitutive of desiring – finding something pleasant, being motivated to get it, thinking it's good (in some sense)[15] – tend to go together in humans and other animals, likely for evolutionary reasons. The things that are good for us (under normal circumstances) tend to be things we find pleasant and things we are moved to act to obtain, since an entity that was moved to cause itself pain or that was not motivated to acquire what was good for it would likely soon go extinct. In artificial intelligences, however, these features could come apart, since they could be designed and

15 Though not necessarily "all things considered". It is possible to desire something while thinking it is not good *on balance*, but it seems plausible that you must think there is something about it that is good even if that is outweighed by other factors – for instance, you desire a cigarette because you believe it will produce pleasure, which is good, while realizing that on balance it is harmful because of the health effects.

programmed any way we choose.

If we are to use desire as a criterion for consideration in our moral decision-making, there must be some empirical criteria that give us justification for making attributions of desires. Again, this is not seen as crucial when it comes to humans, since we know that humans are the kind of things that can have desires. But though the competing theories of desire agree about obvious cases, they are not empirically equivalent – there are cases where one theory will say that a human or other animal does possess a desire while another theory denies it. What we want to say about these cases will determine which theory we want to adopt for AIs. Furthermore, it is not sufficient to simply say that desires have something to do with consciousness or volition – consciousness and volition are themselves mysterious and controversial, and cannot be measured or perceived. It may well be that our intuitions are that desire requires phenomenal consciousness,[16] and if we were somehow certain that consciousness and/or volition were absent, we would always deny that desires were present. But those are not things we can test for directly, and any potential tests for them will be extremely controversial, so we must find a practical solution to determining whether the presence of desires is likely.

16 See Worley (1997) for a list of examples showing that we should never ascribe beliefs or desires in a situation where we know consciousness to be absent.

If we accept desires as the criteria for moral standing, there may be a way to justify the Moral Turing Test, if we adopt a theory of desire that makes desires something that reliably correlate with the types of behaviours that we are sensitive to in the Moral Turing Test. This is obviously the case if we accept an interpretationist theory like that of Donald Davidson[17] or like the intentional stance theory of Daniel Dennett[18] – according to this type of theory, what it means for something to have desires is that we legitimately treat it as having desires. Desires are something we project into other entities when it's predictively useful to do so. Obviously any entity that would pass the Moral Turing Test does in fact have the same desires as a human according to this definition, since by hypothesis we could make the very same predictions about it. However, adopting this theory of desire would make it unattractive to adopt a desire-based moral theory (if we want morality to be objective). If we are to make important moral decisions on the basis of desires, there must be an objective fact of the matter about what types of entities have desires and which ones they – since it is a morally relevant property, it cannot simply be a matter of interpretation or dependent on what knowledge and alternative explanation we have available at the given time.

17 Davidson (1980).
18 Dennett (1987).

There is, however, a simple and appealing theory of desire that would give us a definition of desire as a real, objectively existing property, and licence to use the Moral Turing Test. According to the action-based theory of desire, as articulated by theorists such as Michael Smith, desires are dispositions to take certain actions in certain circumstances. We can infer what dispositions humans have based on which actions they in fact take, and they tend to take actions in most situations where we would say that they have desires, so for the most part, this theory captures our intuitions about which desires humans have. And if an AI passes a demanding enough version of the Moral Turing Test, it would seem that it has done so because it has dispositions to behave very similarly to humans in response to similar circumstances, and so it would in fact have the same desires as humans according to this theory. We would therefore have reason to believe that we are correct in treating sufficiently human-seeming robots as persons.

For a defence of the Moral Turing Test to work, our theory of desire must be consistent with the pre-theoretical assumptions that led to adopting desire as part of our moral theories. Those moral theories were found plausible on the basis of our intuitions about desire, and if it turns out that "desires" are quite different from what we thought, that would be a reason to distrust the

moral intuitions that led us to adopt desire-based moral theories in the first place, not to simply plug in our new theory of desire to our existing moral views. Not everything that has to do with desire according to our naive folk-psychological conception of it is something we care about morally, so there can be some deviation, but if we come to a plausible account of desires that gives moral rulings that are deeply inconsistent with our intuitions, we would likely want to reject our desire-based moral theory and still have no way to justify the Moral Turing Test. Therefore, if this theory of desire fails either of these tests, it will be inadequate for our purposes. It must both match our intuitions about desires to a sufficient degree, and it must avoid unintuitive moral results.

I will argue that the action-based theory gives unintuitive results about desires, claiming there are no desires in some situations where it seems desires are present, and attributing desires in circumstances where it seems there are none. Furthermore, even if it is the correct theory of desire, it gives us extremely unpalatable results when used for moral decisions. It doesn't pick out all and only the morally relevant desires, and if used in moral theory, it would grant and deny moral standing incorrectly in obvious cases. Only if a theory matches our intuitions about obvious cases can it be useful for solving the difficult questions about borderline cases such as AI,

and so the action-based theory is of no use here.

An alternative to this is a phenomenological theory of desire, where desire is a particular feeling. Of course, the only way to empirically identify feelings is through introspection. We have to rely on the testimony of others to have any information about sensations other than our own. Since we each can introspect our own desires, we can assume that other humans have similar phenomenological experiences, since they are sufficiently similar to us that there is no reason to suppose there is a difference. We also must assume that we can trust people's own reports of their experiences to be truthful and accurate at least some of the time – assumptions that are not excessively controversial, I think, since under normal circumstances people would have little reason to lie. The methods we use to detect desires in other humans may be imperfect and fallible, but regardless of how accurate or inaccurate the techniques available to us are when it comes to other humans' desires, we are not obviously justified in assuming these methods are at all reliable when it comes to AIs, since AIs are dissimilar from humans in potentially relevant ways. Therefore, we need another empirical criteria to use for entities other than humans. A potential solution is to identify neurological patterns that correlate with phenomenological sensations of desire in humans, which will give us some justification for concluding that an

AI with similar states in its brain structure will also possess desires in the phenomenological sense. If we do identify those structural features, we might be able make AIs that reliably correlate their behaviour with their desires in the correct ways, and then the Moral Turing Test would provide adequate guidance for dealing with them. The point is that in the absence of that knowledge about the makeup of those machines, the Moral Turing Test provides no independent evidence of the the moral standing of AIs.

Philosophical examinations of the neurological structure of desire have been conducted by Carolyn Morillo and Tim Schroeder, which I will attempt to apply to the case of Artificial Intelligences. This analysis will show that if we adopt the phenomenological theory of desire, the Moral Turing Test will not by itself give us reliable guidance. We could, in principle, program a machine to exhibit all the outward indicators of desire while not giving it the structure that would give rise to the phenomenal state, or vice versa, the correct phenomenal states with no outward evidence of them (indeed, we may not even be in a position to know whether we've done those things, given our incomplete knowledge of the neurological basis of desire). But the phenomenological account, as I will show, is the preferable option for desire-based moral theories. Since this account leaves a lot of room for the Moral Turing Test to fail, I will argue that reliance on the Moral Turing Test in our

ascriptions of moral standing to AIs is not justified under a desire-based moral theory.

These two theories are of course not exhaustive of potential theories of desire, nor is it impossible for there to be other empirical indicators. My project, if it is successful, will have succeeded only in showing that what I take to be the most promising way of justifying the Moral Turing Test fails. The most attractive theory of desire that makes desires detectible with some form of the Turing Test is not compelling when combined with desire-based moral theories, and the phenomenological theory provides a viable alternative that is more attractive for use in a moral theory but does not justify the use of the Moral Turing Test. Therefore, we have no immediately apparent justification for relying on the Moral Turing Test if we adopt a moral theory that has desire as a prerequisite for moral standing. It is still possible that such a justification could be found, we just do not have any justification at the moment. Alternatively, we could adopt a moral theory that does not rely on ascriptions of desires. I will not argue for or against that option, I merely wish to make clear what we must commit ourselves to in order to be consistent.

Chapter II: The Disposition-To-Action Theory

In this chapter, I will present the theory of desire that seems best suited to defending the Moral Turing Test, wherein a desire is a disposition to perform actions to bring a certain state of affairs about. I will consider Michael Smith's arguments that this view is better able to attribute propositional content to desires while accounting for the uncertain epistemology of desire, and argue that these arguments fail. I will then turn to other problems for this view that bear directly on moral theory, and show that accepting the action-based theory of desire would render desire completely unsuitable as a basis for moral claims. Finally, I will examine the implications of using this theory as an empirical criteria for attributing desires, even if not a definition of desire. I will show that it is inadequate to that purpose as well.

1. The theory

Under an action-based theory of desire, having a desire is nothing more than having a disposition to do certain things in certain circumstances (generally it is claimed that the desire is the structure in the brain that gives rise to a particular disposition). A desire might happen to correlate with a

disposition to feel certain phenomenological sensations, but this is not essential to being a desire under these views; the disposition to action is what is relevant. This account of desires is usually accompanied by a dispositional account of beliefs, where a belief is also a structure that causes a disposition to act, and the combination of desires and beliefs produce particular actions. The main difference between a belief and a desire, on these theories, is "direction of fit":[19] people tend to be disposed to abandon beliefs when confronted with evidence against them, but desires persist when the world does not conform to them.[20] This theory of desires does allow for unconscious desires (i.e. desires you do not know you have). Although desires interact with beliefs, their existence does not depend on the existence of any particular belief – the posssessor of a desire must be disposed to perform actions they believe will bring about the state that is the object of the desire, but they need not know exactly which disposition causes the actions, or know what particular state their actions are directed at producing. A number of philosophers have held such a theory of desire, such as G.E.M. Anscombe[21] and Roger Stalnaker,[22] but I will focus on Michael Smith's formulation of the theory as among the clearest, as well as one of the few specifically concerned with moral theory.

---

19 See Anscombe (1957), S.32
20 Smith (1987), p. 54.
21 In Anscombe (1957).
22 In Stalnaker (1984).

2. Propositional content and epistemology of desire

Smith argues for the disposition-to-action view mainly by attempting to defeat the rival phenomenological theory, wherein desires are introspectable. His main arguments are based on cases that he claims are intuitively explained by dispositions to action, not phenomenal states ("feelings"). However, all of the cases Smith relies on can be accounted for by a phenomenological theory of desire while respecting common-sense intuitions. One claim Smith makes is that desire cannot be a feeling, because if it were, we could always tell when we had a desire, as we can with feelings. Smith claims we cannot do this – his example is of a man who goes to a certain newsstand that has mirrors behind the counter, but would deny (with all sincerity) that he goes there for that reason. Smith claims that if this man would be disposed to stop going to this newsstand if it removed the mirrors, and would be disposed to go to another newsstand if that one put up mirrors behind the counter, it is clear that he has a desire to look at himself in a mirror while he buys his morning newspaper, even if he cannot introspect any feeling of that being a desire he has.[23]

23 Smith (1987), p. 46.

However, this argument is not persuasive. For it seems that the man's desire in Smith's example is not an intrinsic desire, but an instrumental desire – not a desire for something for its own sake, but only as a means to a state that is desired for itself. And much as we are not always cogniscient of everything our beliefs entail, we need not explicitly desire every means to the satisfaction of our desires. It seems possible that the desire the man has in this case is the desire for a specific sensation, that he happens to get from seeing himself in the mirror when he buys his paper. If you asked him why he buys his paper where he does, he might not be able to give the ultimate reason, but he would likely tell you that he can introspect a feeling associated with buying it there. He may not even be able to articulate this desire, or accurately describe it, but he is introspectively aware of it. Smith has not established that it is possible for an individual to have a desire without knowing he has any desire at all, since there is always the possibility of offering this alternative story. It might still be the case that an individual always knows when she has some desire, even if she cannot say exactly what it is for. She might have any number of instrumental desires in the service of the intrinsic desire, but the intrinsic desire is what is important.

Smith's other example of the supposed failure of a phenomenological account of desires illustrates this as well – Smith describes a situation where

you search the refrigerator not knowing what you want from it, but eventually "realize what it was you wanted all along."[24] I see no reason to take that description literally. We use idiomatic expressions like that all the time while knowing they are not literally true, and on reflection we would accept a paraphrase readily. Requiring a paraphrase in this case should not be taken to be evidence that the phenomenological theory goes against common-sense intuitions. I think people would readily assent that in that case, you desired a means to a particular sensation, and evaluated which of the items in the fridge produced would lead to satisfaction of that desire. Once you formed a belief that a particular item would cause the sensation, you formed an instrumental desire for it. In everyday parlance, we say it was "what you wanted all along" because we cannot spend all day ensuring accuracy in our descriptions of desires.

The cases mentioned so far are ones where a sensation is what is desired, which makes it more plausible that we could introspect a desire without being able to articulate its exact content (sensations are frequently things we cannot explicitly describe). Handling this type of desire would be sufficient for a moral theory that adopts the view that mental states are the only things that are intrinsically desirable and that contribute to welfare. However, we would have reason to adopt the action-based account if we

24 Smith (2011), p. 46.

want to allow for intrinsic desires that do not have sensations as their object, if it is the only theory that can handle them. But that is not the case, since a phenomenological account could still handle such desires. Under phenomenological theories, the desire is a sensation, but its object/content need not be. Let us now turn to an example of an intrinsic desire for an external state of affairs rather than a mental state.

Smith addresses an argument against the phenomenological account intended to demonstrate the possibility of a person believing he possesses a certain desire, but being mistaken. In this case, a man claims to have a desire to be a musician, but he also desires not to upset his mother, who also desires that he become a musician. When his mother dies, he loses all dispositions to attempt to pursue a career as a musician. Smith claims that we should conclude that he had no fundamental desire to be a musician, and so believed himself to have a desire when he did not. However, we need not accept that the individual was mistaken about the presence of a desire, but only that we was confused about whether the desire was intrinsic or instrumental. As Smith describes the case, the man did not believe there was a desire where there was none, but correctly introspected a desire sensation, but incorrectly attributed its object. It seems quite implausible that a person could believe she had a desire but have it turn out she had

none at all, so once again Smith's objection fails to refute the phenomenological theory.

Smith, however, would take issue with this response to his cases, and claim that even if we can introspect a phenomenological sensation of desire, if we cannot accurately introspect the propositional content of desires (as we clearly cannot) we would still need to supplement the phenomenological conception with an "independent and self-standing" account of how desires get propositional content.[25] Smith claims that this reduces the motivation to adopt the phenomenological theory of desire, since he seems to think that once we find this independent account of propositional content, it by itself would give us a workable account of desires, and the phenomenological part would add no explanatory power and could be discarded as unnecessary. However, Smith does not make clear exactly how the disposition-to-action account is supposed to have a non-independent account of propositional content in a way that is relevantly different from the phenomenological account.

One way to defend such a claim would be to say that whatever brain state disposes one to take the actions associated with a desire, it itself represents the desire's propositional content, solely in virtue of its structure and

25 Smith (1987), p. 48.

function and not in virtue of being conscious or phenomenological (much as a map of North America represents North America without having any inherent phenomenological content). Desires would then be representations, with intrinsic propositional content that cannot be separated from them, though vastly different (and completely independent) from other representational states like perceptions. This would be adopting the position described by Timothy Schroeder as being a claim that "to desire that *P* is to have a mental representation that *P* which plays a certain causal role, namely, that of disposing one to bring it about that *P*", and therefore "believing that *P* involves a mental representation that *P* playing one functional role, while desiring that *P* involves a distinct representing object (token) with the same content playing a different functional role"[26] (as opposed to saying that the exact same representations can be believed, desired, both, or neither). If this were the case, the desire would have propositional content directly under the action-based account, in a way that is unlike how it would work in the phenomenological account.

It is unclear whether this is what Smith himself holds, but anyone defending an action-based account of desire on the grounds that desires themselves represent (rather than inheriting their content from beliefs or perceptions) will encounter serious problems. For it is not enough to simply assume that

26 Schroeder (2004), p. 24.

there is some particular structure that causally underlies the observable actions associated with each desire without having any idea where in the brain that might be. And the assumption that there is some easily isolated brain state associated with each desire that could carry its propositional content has proven to be unfounded. Schroeder claims that there are no such representations identifiable in the brain of humans (and so of course we would have no way to begin identifying such representations in AI). "In the whole of the cerebral cortex, there is no plausible home for the scores of mental representations required by our scores of desires",[27] and were there one, we should have found it by now with our neuroscientific understanding of the brain being at the level it is. We have already identified which parts of the brain are associated with each of the other representational states. Schroeder says that it would be implausible to claim that there are representations associated with desires that are non-localized in a way that is very neurologically demanding and completely different to how the representations that we have in fact identified function (for example: sensory/perceptual representations being found in the primary sensory cortex; the limbic association area and hippocampus being responsible for memory representations; etc.)[28] – and yet they would have to be, since the actions involved in desire are multiply realizable. To borrow a case from

27 Schroeder (2004), p. 24.
28 See Kandel, Schwartz, and Jessell (2000), p. 351.

Zenon Pylyshyn,[29] the desire of a person attempting to phone for help after witnessing a car crash cannot be associated with a structure localized in the fine motor control area that produces the hand movements that dial 911, since it might produce the action of dialing 999 if the person is (or believes herself to be) in England, or 0 to get the operator if, say, the 9 key on the keypad is broken. If the phone is not operating at all, it might involve the coarse motor control area causing her to run to find another phone, or it might involve activating the perceptual centres to obtain more information. And the case is far worse with a desire such as wanting to be economically comfortable in one's old age. It becomes implausible to say that the conjuntion or disjunction of the operation of such a large number of parts of the brain in such a massive number of ways is what represents the content of the desire.[30]

On the phenomenological view, by contrast, "desiring presupposes the prior existence of the capacity to bear the content P in a perceptual or cognitive form—in the form of it seeming to be the case that P." The representational part is "found just where the representational capacities for perception and belief are found", and the content is set by how those interact with the

29 Pylyshyn (1986), p. xiii.

30 Note that it is not impossible that AIs might be capable of having such states, since their neural structure might be vastly different from humans'. But that would be no help, since it is

"feeling" part. Therefore, "all that needs to be added to the brain in order to have desires is... neurons connecting the representational capacities" to where the desire feelings are found, "and this is a much more modest demand on brain space than that apparently called for by the representationalist version of the [action-based] theory."[31] For the phenomenological theory, the "content" or "object" of a desire is determined by the belief or set of beliefs that will affect the "desiring" sensation in the right way. This allows us to be fallible about our desires – we might find that if we were to come to believe a certain proposition, we would cease to have the relevant desire-sensations, or they might persist but demand a different interpretation, but this may be because of background beliefs that might be false and that we do not realize are playing a role. There is a sense, then, in which the propositional content is "independent and self-standing", since it already exists in independent, self-standing beliefs and perceptions. But this does not seem to be a problem, since those propositions being the content of the desire is still inextricably linked to phenomenology.

It seems that any advocate of the action-based account should adopt the parallel model of how desires would get propositional content – connection between the representational centres (perception, memory, imagination, etc.) and the action centres. Depending on how the representations are

31 Schroeder (2004), p. 29.

connected to action, we can tell whether given propositions are believed or disbelieved, and desired or not desired. The way to tell which desires a subject actually possesses will require testing which combinations of beliefs will lead to actions and which will not, which will allow us to create a theory of exactly what the desire is. But the action-based theory would thereby lose Smith's supposed advantage over the phenomenological theory for content and epistemology of desire. The way desires get their content under the phenomenological account will not be very different from the disposition-to-action theory if this is the case – they will still inherit them from perceptions and other representational states. It will be no more "independent" in one case than the other.

Once we have this account of the propositional content of desires worked out, we can see that the phenomenological theory in fact gives us a more plausible epistemology of desire than the disposition-to-action theory. Under the phenomenological theory, we are not infallible about our own desires, but we do have privileged access to them. From introspection, we get a type of information about our desires that others cannot have. This does intuitively seem to be correct. We make indirect inferences about other people's desires, but these are frequently subject to revision based on the testimony of the possessor of the desire, who has this special access to a

different kind of information. It seems this is how we generally conduct ourselves in our daily lives. We do frequently make mistakes about other peoples' desires, and we tend to accept their claims when they inform us that we are mistaken (unless we have a clear explanation of why they are making a mistake in a particular case). Under Smith's view, the only reason we might be more reliable about our own desires than other people is that we have more opportunity to observe our own behaviour. If there were someone were observing all my actions, then it seems that if there was any disagreement between us about my desires I would have no more reason to believe I was correct than that the observer was, since we have exactly the same type of information. But except in very unusual cases, it seems that the possessor of a desire is the more reliable guide to its content, or at least has some special access to it, in a way that is not explained simply by their having observed more of their own actions.

Thus, the action-based theory does not have any advantage when it comes to combining propositional content with a plausible epistemology of desire.

3. Problems as an element of a moral theory

Disposition-to-action theories have been criticized on the grounds that they do not provide adequate explanation of actions. Agnes Gellen Callard, for example, says that defining hunger (i.e. the desire to eat) as a disposition to eat is useless, since the mere fact that someone has a tendency to eat when hungry gives us no understanding of why they eat, it only pushes the explanation back – we can still ask why they have the disposition to eat.[32] Similarly, Nagel argues that it might be the case that a dispositional desire is necessary for action without desire providing any explanation or reason for action.[33] However, this type of objection is of no concern for the purposes of my project. Desire need not have any explanatory power to be morally significant – or rather, it needs to explain moral facts, but not any empirical facts. There might be some deeper property that grounds the dispositions, but it doesn't matter what that property is for our present purposes. The explanatory role of desires might be important for other purposes, but a property need not fill that role to fill the role of "desire" as it appears in moral theories. Thus it is not "trivial" that a desire must be present for beliefs to motivate (as Nagel claims), as long as that fact has moral importance.

And indeed, it has been argued that desires in the sense of a disposition to

32 Callard (2008), p. 109.
33 Nagel (1970), p. 30.

action do play a role in morality. Aaron Simmons has argued that we should adopt an action-based account of desires in order to get the correct results when it comes to non-human animals.[34] It is generally assumed that animals such as pigs and cows do not have beliefs about life and death, in virtue of not having concepts of life and death (though it is difficult to be certain due to the present impossibility of adequate communication with them). If they can be said to consciously entertain beliefs at all, they have, at most, very simple beliefs, about eating and avoiding predators and such. Some theories of desire would therefore say that they desire food, and desire safety, but do not desire life. If we accept this, and also adopt a theory of well-being that depends on desire, we would have to say that we do not harm such animals merely by killing them. Simmons points out that such theories attempt to avoid this unintuitive result by claiming that killing animals harms them by inflicting pain, or by depriving them of satisfaction of future desires, but Simmons maintains that it is intuitively obvious that the killing itself harms them. We would therefore have reason to prefer a theory of desire that can accomodate that intuition. The disposition-to-action theory can do so, because pigs are disposed to take actions that will result in a states of affairs in which they continue to live. Therefore, it can be claimed that they desire life. This is true whether or not they are aware that such a state is "life", have any concept of "life" or "death", or are aware they have such a desire

34 Simmons, 2009.

at all. Behaviour such as fleeing from predators, locating and consuming food, etc. are actions, and as such are motivated by desire. Furthermore, this disposition is (presumably) responsive to beliefs – counterfactually, if a pig did have beliefs about life and death, they would affect its behaviours. Thus, if we adopt a dispositional theory of desire, we can say that killing animals harms them – a conclusion which it seems might appear intuitively plausible even to those who think we are permitted to kill animals (obviously, any number of factors could outweigh welfare considerations depending on which moral theory we adopt).[35]

However, there are problems for using this theory of desire in our moral deliberations. The action-based theory does not make the correct distinctions between the states that matter in a desire-based moral theory and those that do not. The property that it picks out is not the correct basis for making moral claims, since it includes habitual actions that are not

35 Note that this type of view is also compatible with non-welfarist deontological moral theories. Smith's own moral theory is that what we are morally obligated to do are things we would do if we were ideally rational, in the sense of internally coherent. Smith argues that to avoid being self-thwarting, an individual must necessarily have certain desires (in the sense of being disposed to do certain things), since these desires are a requirement to be instrumentally rational and coherent, and will desire that others have their rational capacities developed as highly as possible as well. This is defined solely in terms of dispositions to action. Therefore, it is possible for desires as conceived in the dispositional theory to be the foundation for a deontological moral theory, wherein entities that we have obligations towards are those with the capacity for belief-desire rationality in the procedural sense (Smith, 2011). Thus, this type of desire could be relevant even to deontological theories that give no moral importance to welfare – the entities that we have moral standing are still those that have the capacity for desires. I will not explicitly address this theory, however, and will focus on more widely-held moral views.

"desires" in the sense moral theories mean when they say that desires are morally important, and excludes wishes that do have moral importance and should count as genuine desires. If we did adopt the action-based theory of desire, that would only lead us to conclude that desires are not what we should base our moral theory around, and so we would still be left with no way to justify the Moral Turing Test, as will become clear.

Smith claims that we frequently act on the basis of desires without having any real feeling of desire – we cross the road while feeling completely dispassionate about it, but it would be absurd to say that therefore "I cross the road... even though I do not want to!"[36] However, feelings of desire come in degrees, and in that case it is simply that the desire is so faint as to be barely noticeable. For this reason, we could say that frustrating that desire has less moral weight than frustrating stronger desires, an intuition that the action-based theory is less equipped to deal with. Indeed, if there was genuinely no feeling at all that the possessor of the desire was aware of, and if preventing that outcome caused no frustrations of any actually felt desire (including the desire not to be mildly, momentarily irritated, say) it seems plausible to claim that there was no morally-relevant desire at play. Cases like that seem more like someone who crosses the street out of habit, because she has done so many times before. If she were actually going

36 Smith (1987), p. 49.

somewhere different this time, the phenomenological theory of desire allows us to actually say that she crossed the street even though she did not want to, which in fact seems the correct description, contrary to Smith. The action-based theory must say that she simply had an irrational desire to go the way she usually goes, when it seems like she had no such desire at all. Smith is therefore correct that the phenomenological account precludes subconscious desires, but it seems that actual cases of truly subconscious motivation to action are often not correctly described as desires. At the very least, it seems like those "desires" do not have moral importance in the same way that other desires do, and it seems like an advantage for the phenomenological theory to be able to rule out such motivations from the class of morally relevant states. Even if there was no conflicting desire to make such a "desire" irrational, it seems absurd to say that satisfaction or frustration of it would have an effect on well-being. It would not be morally impermissible to interfere with that type of action unless it prevented the satisfaction a real, felt desire, either at the moment or in the future.[37]

The phenomenological theory is better able to handle the sort of desires that

37 In addition, if one wanted to adopt a rationality-based moral theory, it seems having "desires" of that kind does not make an entity a rational being. And if we want an autonomy-based theory, those do not seem to be expressions of an individual's autonomy and it is plausible to say that interfering with them is not an infringement of autonomy in the way that interfering with the satisfaction of a felt desire is, so it seems phenomenological theories of desire fare better on that front as well. There may be ways to make those claims compatible with action-based theories of desire for that type of moral theory, but I will not explore this in greater detail here.

could never lead to action, which is a difficult case for a theory that defines desire as a disposition to action. Take, for instance, the desire that God not exist. It seems no actually possible set of coherent beliefs could dispose someone to take any action in service of that desire, but it could certainly have phenomenological effects. One could say that there is an impossible counterfactual that makes it the case that these are still dispositions to action, but there would be no way to determine if such a disposition exists, and it seems strange to say that any such disposition is what such a desire actually consists of. However, it seems perfectly possible to have beliefs one way or the other, and our feelings caused by those beliefs are what makes them the content of desires. Even for propositions where no change in our belief is possible, we might still determine the content by entertaining different beliefs and their negations, resulting in some modification to the desire feelings. No such solution is available to the action-based theory. The phenomenological account might have difficulty handling desires such as wanting a round square to exist – it is impossible to believe that without being conceptually confused, but it might be possible to desire it. If so, it might be possible to imagine a round square existing in some sense, even if not a very accurate one, and that might be enough to effect the phenomenological changes that are relevant. At the very least, that is more plausible than a non-conceptually confused individual taking any action in

service of that end.

Timothy Schroeder suggests that an adherent of a dispositional account could say that these aren't "desires" in the sense they are concerned with, but are mere "hopes" or "wishes", belonging to a separate category.[38] However, it seems they are desires in the sense we are concerned with in the realm of moral theory – most would say they have the same moral import as desires that could lead to action and should belong to the same category. It would be unintuitive to claim that the satisfaction or frustration of these "wishes" does not contribute to well-being, for example. If someone had strong "hopes" or "wishes" about God, and it turned out God did not exist, many moral theories would hold that this person's life has gone much worse than if God did exist, even if her "hopes" about God weren't the kind of thing that affected her actions. Similarly, Galen Strawson proposed the thought experiment of the "weather watchers", beings that have no capability to act and thus have not evolved structures that dispose them to act, but that have beliefs about the weather and hopes about how the weather will turn out that involve feelings.[39] It seems their welfare is increased if their "wishes" are satisfied rather than frustrated, and it seems that if anyone else was in a position to take action to satisfy those "wishes",

38 Schroeder (2004), p. 20.
39 Strawson (1994), ch. 9.

they would have a moral reason to do so (*ceteris paribus*). Thus from the point of view of morality, for those who wish to adopt desire-based moral theories, there is no distinction to be made between "desires" that lead to action and "wishes" and "hopes" that do not.

Thus, using the action-based theory of desires could lead us to grant and deny moral standing to the wrong entities. For one thing, it would grant moral standing to beings capable only of the type of habitual, instinctive actions that are irrelevant to morality (the unthinking street-crossing kind of behaviour). It is ill-suited to distinguish complex, sophisticated animals (including humans) from even extremely simple organisms, such as insects, paramecia, and protozoa, since those entities do have dispositions to act in that sense. Also, it seems that such a theory will grant moral standing to practically *any* machine. A home computer has "beliefs" and "desires" in the dispositional sense, even a robot vacuum seems to, but it is obvious they have no moral standing. The advocate of the action-based theory might say that there is a threshold of dispositions, wherein a certain level of complexity is required before actual desires are present. However, this is not persuasive. This modification correctly excludes modern robot vacuums, but there is no reason we could not make such a vacuum with more memory and many more dispositions without significantly altering the way they are

programmed. The reason we do not ascribe desires to them is not that there are not enough ways their behaviour is affected by different stimuli, but that we know how they are made. Conversely, a human would not fail to have desires in virtue of having very few dispositions. It is possible to imagine such a person who has extremely simple dispositions, perhaps of the same number and complexity as a robot vacuum, but if these desires were associated with the same phenomenological sensations as in normal humans, such a person would still have desires – and certainly would not fail to have moral standing.

An action-based theory of desire will likely require that the dispositions must be responsive to some beliefs – humans have a disposition to convert carbohydrates into lipids and then adipose tissue when they are introduced to their digestive system, but that does not mean that people "desire" to turn ice cream sandwiches into fat. This is because there are no beliefs that could affect that disposition in any way – there is no information anyone could learn that would prevent them from performing that function. This seems necessary if we are to use this theory morally, as Simmons attempted to – if dispositions to do things that keep an organism alive are by themselves sufficient for having a desire for life, then the ham in your ham sandwich is no more morally problematic than the wheat used to make the

bread – both had a desire for life, and had moral standing on that basis. It might seem that responsiveness to beliefs could be used to make the relevant distinction between the cases – wheat, protozoa and robot vacuums do not have genuine "beliefs", and as such cannot have desires. However, to make this distinction do the necessary work to rule out all the problem cases, a dispositional account of desire would need a far more robust account of belief than a mere disposition. Protozoa, robot vacuums, even wheat do process information, and treat it in a way that tends to conform to the way the world is, and are disposed to respond on that basis. It is unclear what definition of "belief" we could use to distinguish those cases from actual beliefs, but it seems difficult to do so in a way that allows us to reliably identify beliefs without the benefit of introspection. For creatures with vastly different sensory apparatus than us, it will be impossible to tell whether a certain processing of input is akin to our eyes taking in photons and processing it into images (which involves belief), or to our stomachs taking in food and processing it into energy (which does not). How are we to tell which takes place when ants "communicate" by vomiting chemicals into each others' mouths, or when computers interpret the pattern of electrical impulses caused by the input of binary data? This abandons a significant advantage of having a purely action-based theory of desire. Furthermore, even if we have this robust conception of belief, there can still be

counterexamples: Stampe provides a case where a tennis player believes that serving in a certain way will cause him to fault, and his nervousness causes him to do exactly that – he has a disposition to take whatever action he believes will cause him to fault, but he clearly does not desire to fault.[40] Furthermore, even if we have a robust conception of beliefs, the dispositional theory of desires makes it far too easy to get from having beliefs to having moral standing, since all it takes to have desires in the full sense is performing some behaviours that could be affected by beliefs. In essence, it makes beliefs the more important criteria for having moral standing, which does not seem independently plausible.

4. Empirical detectability and practical usefulness

A significant practical advantage of the action-based theory is that it would make it (relatively) easy and straightforward to tell if an entity possesses desires. Science generally proceeds on the basis of identifying dispositions of one kind or another, so if desire is nothing more than a certain kind of disposition that responds to certain other dispositions in certain ways, we can identify it in the same way we identify any other entity accepted by science. We simply look at which

40 Stampe (1986).

conditions lead to which results.[41] And it is easy enough to observe actions. Of course, it will not always be easy to determine exactly which desires a being has – in a certain situation, the dispositions might be blocked by various factors, such that (for example) a desire might be present but never lead to action, because other desires always override it, or because relevant beliefs are absent or false beliefs are present. There will be various equally empirically supported theories about which set of beliefs and desires led to the data that was observed. But it will generally be possible to tell if the being is the type of thing that has desires – this will be the case if it has certain dispositions that disappear in the face of perceptions of contradictory information about the world (beliefs), and if these dispositions affect other dispositions that persist in the face of contradictory information, and result in attempts to force the world to conform to them (which would be the desires). This test could be applied to any entity we choose, even those radically different from humans.

Indeed, even if we do not accept the action-based theory as a definition of desire, it might be useful as a test of the presence of desires for pragmatic reasons. If desire involves phenomenal states, we will have knowledge of

41 There are some general problems with dispositions, of course, but these can be put aside for present purposes, since, as we shall see, any theory of desire will likely have to make some use of dispositions of one kind or another.

our own desires, but no direct way to tell if other beings have desires. If the morally relevant desires reliably correlate with a pattern of actions where we can identify a disposition, then even if desires aren't dispositions, we can use them as the basis for making correct assignments of moral standing.

However, this is only useful if there are not a large number of cases where the dispositional theory clearly gives a result that is at odds with obvious moral intuitions. It seems obvious that dispositions that meet all the criteria of the action-based theory can exist where there are clearly no desires (and certainly where we would likely want to say there is no moral standing).

Therefore, as we have seen, the dispositional account of desire is of no help in justifying the Moral Turing Test. It is no more plausible as a theory of desire than the phenomenological theory, but even if it were, it picks out the wrong properties for a moral theory, and it would assign moral standing to the wrong entities.

Section III: Phenomenological Theories

Based on the arguments in the previous section, we have two options

available: find a new, non-desire-based moral theory, or find a new theory of desire. If we take seriously the intuition that desire matters morally and could form the basis for moral standing, then it is a constraint on a theory of desire that it give results that are consistent with moral intuitions. In this section, I will argue that there is a plausible theory of desire that is compatible with desire-based moral theories, and that it does not justify the use of the Moral Turing Test. Thus, we can continue to maintain a moral theory that uses desire as a basis for moral standing, but if we do so, it will require deeper investigation than social interaction and observation of behaviour to determine which entities deserve moral consideration.

I will begin this section by examining what I take to be the naive phenomenological theory of desire, the hedonic theory,[42] and identifying the problems with it. I will then present a more sophisticated theory proposed by Carolyn Morillo, and identify some lingering problems for it and how to address them. I will thereby arrive at the theory that I believe is the most attractive if we adopt a desire-based moral theory.

42 Though Smith accuses his critics of implicitly adopting a phenomenological theory of desire, very few have explicitly articulated and argued for any such theory. Thus, I am not sure how widely held any version of a phenomenological theory actually is. Schroeder (2004) treats the hedonic theory as one of the main contenders in his examination of desire, but does not provide examples of anyone who explicitly advocates it.

1. The hedonic theory of desire and its faults

An initially attractive statement of a phenomenological theory of desire is what is sometimes called the hedonic theory of desire – to desire a state of affairs is to take pleasure when it seems that state of affairs has obtained, and to take displeasure when it seems that state of affairs has failed to obtain. However, the hedonic theory of desire has some difficulty with theories of pleasure. According to some theories, what makes a state pleasurable is that it satisfies a desire. It is circular and uninformative at best to define desire in terms of pleasure if pleasure is to be defined in terms of desire. Thus, we can only say that to have a desire for a state of affairs is to have a feeling of pleasure when it seems that state of affairs is actualized if pleasure is a distinct sensation. Some have argued that pleasure is such a sensation, like seeing the colour blue, that sometimes forms a part of our conscious experience and it is present whenever we are in a pleasurable state. This theory of pleasure is relatively unpopular, however, since most deny that it is possible to introspect a common feeling that remains the same across the various and diverse pleasurable experiences – there is nothing similar, it is claimed, between the feeling of eating a delicious meal and of apprehending a clever mathematical proof, though both might be pleasurable. To deny that undermines the reason for

adopting many of the desire-based moral theories – the lack of such a sensation is often a motivation for saying that what is good for a person is to have the mental states she desires, rather than saying that a simple sensation of pleasure is good. Also, some people adhere to a philosophy of asceticism, and desire to not feel the simple sensation of pleasure. Others are masochistic, and desire the sensation of pain. These are paradigm reasons for including desire in our theory of welfare, and we should not define desire in a way that prevents it from handling these cases.[43] We should say only that there is a sensation involved in desiring.

A potential problem for phenomenological theories of desire such as this one is that they might not define desire in a way that is acceptable to everyone who uses desire in their moral theory. Part of our goal is to ensure that we can agree on the term, to ensure that people aren't simply talking past each other, and if our definition of desire rules out a number of desire-based moral theories and pushes us to a particular view, then it seems like we are proposing a new concept rather than offering an interpretation of a term common to a number of theories. However, despite how it may seem at first glance, this theory is compatible with most theories of welfare, and we can

43 Of course, there have been various attempts to handle such cases within the context of non-desire-based welfare hedonism, and I do not wish to argue against that theory here. I only wish to point out that many people have been convinced by these types of arguments, and the issue is contentious enough that it is desirable to be able to remain neutral about theories of pleasure.

remain neutral about how various terms such as pleasure and pain relate to this concept of desire. If we wish to claim that what is good for a person is that her desires be satisfied, and desire satisfaction is a type of pleasure, this is still not necessarily equivalent to saying that what is good for a person is pleasure, and does not commit us to welfare hedonism. It might also be the case that disposition to pleasurable feelings is what establishes a person's desires, but what is good for them is that the state of affairs obtains that would dispose them to that pleasure if they knew of it, not that the pleasure is in fact felt. We are still able to hold a view where what is good for a person is something other than a mental state. Alternatively, it might not be the case that all episodes of pleasure involve desire satisfaction, depending on how we define "pleasure". Then only a special kind of pleasure would be good for a person, the desire-satisfaction kind, which is relevantly distinct from welfare hedonism. And of course, welfare hedonism is still a viable option as well. This theory is therefore consistent with a wide range of desire-based moral theories.

The hedonic theory of desire, however, has a significant disadvantage compared to other phenomenological theories, in that it does not posit a phenomenological sensation until it seems either that the state of affairs has obtained or that it has failed to obtain, and so it would give us no way to

introspect a desire until we have reason to believe things have turned out one way or the other. This seems false – before my sports team begins to play a game, I can clearly have a strong feeling of desire that they win, but this feeling is neither pleasure that they win, nor displeasure that they lose, since I do not believe either proposition at that point. Under the hedonic theory, we would have to infer whether we desire something based on expecting that we will feel pleasure if it comes to pass, and displeasure if it fails to occur. But this is unintuitive – if I were to have some novel food described to me, I might expect (correctly) that I would derive pleasure from eating it, and yet it still seems entirely possible for me to not desire it. Furthermore, it also seems possible for me to desire to eat it and yet not derive pleasure from it when I do. It would not be a case of being wrong that I desired to eat it all along – I desired that, but ceased to do so once I found out the results weren't pleasurable. Pleasure and pain can cause our desires, and cause us to revise our desires, but taking pleasure in something is not what it takes for that thing to be desired.

Therefore, desire is a sensation that tends to produces pleasure and displeasure under certain conditions, but is not merely a disposition to feel pleasure or displeasure when it seems to be satisfied or unsatisfied. We can still easily accomodate preference-satisfaction theories of welfare under this

theory – a belief that the state of affairs obtains would remove the desiring sensation. That is what determines the content of the desire. We do not necessarily need to come to believe it for the desire to be fulfilled. We could still maintain that there is some proposition that, were it known, would alleviate the desire – and if that proposition is true, welfare has increased, whether it is discovered or not.

2. Morillo's theory of desire and the amendments it requires

Carolyn Morillo proposes a more sophisticated theory of desire that is still fundamentally phenomenological in nature. The basis of desire, according to Morillo, is a "reward event", an experience that drives all our motivations.[44] Morillo points to empirical evidence that suggests that this experience is always present when we are motivated, and concludes that all our desires are ultimately for this state. One might worry that it would have significant moral implications if Morillo's theory were true – it would rule out the possibility of claiming that some external state of affairs could be good for a person, since they could never desire such a state of affairs intrinsically. It would also mean that all that people ultimately ever desired intrinsically was their own pleasure, ruling out altruistic desires. However, I believe that this

44 Morillo (1990).

conclusion is unwarranted, and in making the necessary modifications to Morillo's theory, we will avoid committing to psychological hedonism or ethical hedonism.

Unlike in the hedonic theory, a feeling must precede the motivation for Morillo's theory to be correct. Morillo explicitly rules out "desires operating prior to, and independent of, any associated reward event"[45] like in the food case mentioned previously. Rather, the feeling is prior, and is what sustains individuals' motivation. This avoids that problem for the hedonic theory and allows that individuals can introspect the presence of their own desires.

Morillo wishes to claim that the reward-event is the only object of desire, but realizes she must answer the following question: "Even if something like [the reward-event theory] is true, why does that not merely tell us more about the mechanism of motivation, about why we have the many different objects of motivation we do have? Why should that mechanism itself count as the only ultimate object?"[46] She claims that the focus is on these outward objects for evolutionary reasons, since obtaining them is what is in our best interests, but "the reward event would still be the aspect of these more complex experiences which is what we are motivated to obtain."[47] But there

45 Morillo (1990), p. 179.
46 Morillo (1990), p. 177.
47 Morillo (1990), p. 177.

is no reason to divide up the experiences like that to isolate which one particular aspect is desired and claim the rest are not. The presence of the reward-sensation might be what causes the whole state of affairs to be desired, yes – but the other features might be what causes the sensation to be present, which should lead us to say that those aspects are desired. Desires can clearly have propositional content that involves things far more complex than pleasure or the basic things that lead directly to pleasure (as Morillo admits), and merely because these desires are likely derived from or explained by basic drives that involve pleasure does not show that pleasure is all we intrinsically desire. Morillo grants that "one might, for different purposes, wish to emphasize the differences among the physiological states (so there would be many different motives), or the differences among the external behaviours and objects (so there would be many different objects of motivation)."[48] Articulating and explaining how desires fit into moral theories seems like a purpose where we would want to do so. Thus, Morillo makes an error in how she connects desire to motivation. The problem might be an ambiguity in claiming that we are "motivated by" a reward-event. That could mean that the experience is what causes us to be motivated, or it could mean that we are motivated to get that experience. The former does not entail the latter, and it would be a mistake to conflate the two senses.

48 Morillo (1990), p. 182.

If Morillo is correct that there is a phenomenal state that is necessary for us to be motivated, it seems far more plausible to say that it *is* desire, rather than that this mental state is the only thing we desire. There is a clear distinction between saying that pleasure is all we desire, and saying that desires are the only things that can motivate. The second, Humean point is far more plausible and widely accepted, and still allows us to make the common-sense attributions of desires that most moral theories depend on. Morillo should say that what she describes as the reward-event is the desire, not that it is the object of our desires.

Morillo grants that desires can sometimes be aimed at things we have never experienced.[49] This cannot be explained by a reward and reinforcement system, and it would undermine Morillo's theory to say that motivation precedes the reward and we are motivated simply by the belief that we will get the sensation in the future. Morillo instead explains these cases with "the hypothesis that we have the ability to envisage future, or possible, or merely imaginary states of affairs, and that such envisagement, particularly when vivid, can link directly to the reward event."[50] Presumably we then assume that we will get more reward if the state of affairs we are imagining actually comes to pass. But there would be no reason to take any action to

49 Morillo (1990), p. 179.
50 Morillo (1990), p. 180.

accomplish things that we know we will never personally experience occuring no matter what we do (for example, things that will happen after our death). Many desires are like that, as Morillo admits – desires for success of one's children after one's death might even exist even in some non-human animals. Perhaps this could be explained by claiming that the more we do to accomplish those goals, and the more they seem likely to occur, the more vividly we can experience the reward-event. This model, however, is problematic in light of cases where no pleasurable "reward" sensation is present. In many cases, our desires involve only unpleasant sensations, and contemplating them is the antithesis of "reward".

Morillo says that her theory "anchors all *positive* motivation in the reward event", but admits that "of course creatures can be, and undoubtedly are, motivated to avoid some things". She presumes "that all such aversion-based learning is also internally anchored, in some aversion event (or events)."[51] Since contemplating these states of affairs will cause us to experience aversion-events instead of reward, the expected effect would be to discourage certain actions rather than promoting them. Morillo is correct that some cases have a positive phenomenal character while others have a negative one, but makes a mistake in apparently claiming that only the positive sensations (the "desires" involving the "reward-event") motivate

51 Morillo (1990), p. 176.

attempts to bring certain states of affairs about, while the negative "aversions" only aim at avoiding certain outcomes by preventing action. In fact, either one might motivate to action, and a desire for a certain state of affairs might produce no pleasant sensations, but only negative sensations when frustrated. A person might desire that some criminal be brought to justice, though the criminal being punished gives him no pleasure, it is just that her escaping punishment upsets him greatly. The difference, then, is not between desire for something to come to pass as opposed to aversion to it, nor is there a difference in motivation to action – pleasurable sensations might be found in the idleness of avoiding doing some difficult or unpleasant work, and desires associated with negative feelings could be just as strong a motivation to action as positive desire. Morillo should not endorse a distinction where something being "aversive" means it "diminished and eliminated operant behavior"[52] as opposed to a real "desire" that promotes action. Both are desires, just with different characters. In fact, it seems an advantage of a phenomenological theory that it can distinguish positive and negative desires based solely on the character of the sensation rather than its effects. But if that's the case, and desire can occur with either positive or negative feelings associated, then Morillo's reward-reinforcement model is flawed. Just contemplating certain states of affairs can produce the relevant

52 Morillo (1990), p. 177 (footnote). Morillo seems to be endorsing that definition, though it's not clear.

sensations, which are not always pleasurable and rewarding but still motivate to action. The sensation is therefore not a reward-event that is the only thing we desire.

This mistake aside, I believe that things are essentially as Morillo describes them. There is a phenomenal state that causes individuals to be motivated, though it can be present even when there is no disposition to action. It is possible for people to do things that are not motivated by this state, but all cases like that seem to be part of a separate class of behaviour that should be treated differently both theoretically and morally, grouped together with (and treated like) things such as digestion and blinking. Morillo admits that it is possible that sometimes perceptions and beliefs could lead directly to behaviour in the absence of this desire sensation, but these cases of "Kantian" motivation[53] are not actually deliberate and rational actions, but "reflex responses, such as the toad's zapping small moving objects ("bugs") with its tongue when they are detected visually."[54] These are not genuine desires, and certainly do not have moral importance – interfering with them would not be morally problematic without an independent reason why we

53 It is interesting to note that both Morillo and Michael Smith, in the papers in which they articulate their views on desire, have as their aim defending the Humean theory of motivation. Humeanism about motivation says that we only act to do what we desire to do, as opposed to a Kantian theory of motivation, wherein beliefs alone can lead people to action. Smith and Morillo each argue for Humeanism on the basis of their theories of desire, despite their theories being wildly different – about as close to opposites as theories can be.

54 Morillo (1990), p. 178.

should not do so. And desire-based moral theories would likely want to deny moral standing to beings capable only of that kind of behaviour. Thus, we arrive at the theory of desire I wish to endorse as the most plausible if desire is to be the basis of a moral theory – desire is a phenomenal sensation that precedes motivation to action, but is not a feeling of pleasure.

3. Problems for the theory

There yet remain a number of objections to the phenomenological theory of desire that must be addressed. The first, most obvious problem is the issue of standing rather than occurent desires. It seems there are many desires we are not aware of at a given moment, creating difficulty for the phenomenological account. A man might desire that his children be successful – when he considers his childrens' success, he might find himself feeling a sensation of desire for it. He might be made satisfied when he learns of increases to his childrens' success, and made dissatisfied when he learns of things that will hinder his children from becoming more successful. However, when he is occupied with tasks that require a great deal of concentration – perhaps a sporting competition, a game of chess, or attempting to solve a difficult mathematical or philosophical problem – it

seems likely that a desire for his children's welfare is completely absent from his consciousness. It seems unintuitive to say that at those moments, he does not possess a desire for his childrens' welfare. Certainly for many moral theories we would want to be able to say that he is made worse off if something negatively affects his children during those times, and contravening a decision he made in the interests of that goal would still constitute interfering with his autonomy during those times. It clearly still contributes to his rationality, as well – if he were to wager his childrens' college fund on a poker game, we would not say he wasn't acting irrationally merely because he wasn't feeling any desire about his children at that moment.

Smith claims that though the phenomenological theory cannot handle these cases, the action-based theory can accomodate them easily, since the disposition to action is present even when it is not triggered. However, Smith errs when he rephrases the claim that "desires are states that have phenomenological content essentially" to "if there is nothing that it is like to have a desire, at a time, then it is not being had at that time."[55] The former does not entail the latter, since it is possible to have a theory where a desire can be had in some sense even if it is not being felt at a particular time, and yet still have desires be defined by their phenomenological content. A desire

55 Smith (1987), p. 48.

might involve a disposition to feel certain sensations when the relevant things are brought to mind, which would still make it the case that phenomenological content is essential to desire. Surely we must assume that even at the moment, the man in question has a disposition to feel the right sorts of sensations when the prospect of his childrens' success or failure is brought to mind. If we say that a standing desire is just a disposition to have an occurent desire when thinking about the right things (or a structure that disposes one to have the occurent desire sensations when thinking about those things), desire is still essentially phenomenological. The differences between the two types of desire could justify treating them differently in various ways, allowing us to accomodate different moral theories. And we still exclude fully unconscious desires, desires that could never become conscious, thus avoiding a problem with the action-based account of desire.

Another potential difficulty for the phenomenological theory of desire is the results it would yield about the changing strength of desires due to states such as depression. Timothy Schroeder claims that a person suffering depression might feel his desires far less acutely than when he was not depressed (which certainly seems to be a standard effect of depression), and that we would not therefore want to say that he had come to desire things

to a lessened degree, as the phenomenological theorist must say.[56] A possible response would be to say that the depressed person's abnormal mental state blocks the feelings, and the person's desire levels should be set by what he would feel under "normal" conditions, i.e. what he would feel if he were not depressed. Schroeder argues that this will not work, since we cannot avoid the unintuitive results by excluding changes resulting from depression without also excluding some genuine cases of altered desires due to "abnormal" brain states – for example, an 89-year-old woman whose syphilis caused her to experience an increase in sexual desire, which she chose not to "cure", considering it to be a real desire.[57] However, this response is not necessary, since it seems perfectly plausible to say that the depressed person's desires are diminished. Schroeder claims that a depressed man who fails to have strong sensations associated with his wife receiving a promotion would say that he does not care any less than if he were not depressed. "'Of course I still want you to succeed, I'm just having a bad patch' is the sort of thing the moderately depressed husband might say to his wife, after being criticized for failing to show happiness upon learning that she has been promoted, and he is likely to be believed."[58] But the real reason the husband has an excuse is that his desire has not diminished *relative to his other desires*. His wife's success might still be one of the most

56 Schroeder (2004), p. 32.
57 Schroeder (2004), p. 32.
58 Schroeder (2004), p. 31.

important things to him, he just feels less desire for everything in his life. This might explain both why people become depressed, and why depression is bad – people subjected to sufficient misfortune sink into depression as an evolved defence mechanism to prevent severe loss of well-being, but it also prevents them from deriving significant changes to their welfare from things that should make them better off.

Thus, the modified form of the phenomenological theory is capable of handling the objections that have been leveled against such theories.

4. Empirical detectability and practical usefulness

But how do we identify the capacity for phenomenological desires? Desires must have propositional content. Under the theory I have proposed, the content of desires comes from sensitivity to beliefs. Introspectively, we can tell *that* we have desires, and *when* we have them. We are not infallible about our desires, since we can be mistaken about their content, but not about their presence. Our desires are responsive to our beliefs, and we have a multitude of beliefs at any given moment, and when we gain beliefs we usually gain many beliefs at once due to things being entailed by other facts

– conjunctions and disjunctions of beliefs, etc. It is not always a simple matter to tell which exact belief affected our feelings. But this does not show that we are not feeling some desire at those times. It is not so easy for entities other than ourselves, but still the representational power of desires can be explained by the representational power of beliefs. A good deal of neuroscientific research has been conducted and the representational states associated with beliefs have been isolated.[59] These states may not by themselves be enough for their possessor to have true "beliefs", but it is enough for our purposes to identify the representational aspect of belief, since that is what factors into desire. It might be claimed that we need a more complex theory of belief to avoid unintuitive results, but this is a separate theoretical issue. It may be that this theory will grant the capacity to have beliefs far too easily – present-day computers have representational states, and it does not even seem to be very difficult to design a computer that has representational states of the same structure as those associated with belief in human brains. It may seem counter-intuitive to say that these machines thereby have beliefs, properly speaking, but though this is a theoretical problem, it is not morally problematic, since the presence of beliefs by itself does not entail any moral facts. This issue, then, can be set

59 See Kandel, Schwartz, and Jessell (2000). We can have a believing-attitude towards the propositions represented in our perceptions or memories, though what it takes to have such an attitude is admittedly somewhat more mysterious.

aside for our present purposes.[60]

We also need to identify the neural state associated with the phenomenological component of desire. Schroeder provides an overview of various candidate structures – the activity of the anterior cingulate cortex, in particular the perigenual region, correlates to certain phenomenological sensations, and modifications to it alter these sensations and attitudes towards certain states (it is unclear the degree to which this is directly correlated with pleasure and displeasure).[61] The ventral tegmental area, and the substantia nigra pars compacta, correlate with reward and punishment, and seem to cause some sensations, though not always pleasure and displeasure.[62] The circuit connecting the nucleus accumbens, ventral pallidum, and brainstem parabrachial nucleus is another candidate, and Kent Berridge claims that it is this structure that is the seat of "liking", which is very similar to desiring.[63] Some of the reasons given to reject some of these structures as the seat of any phenomenological sensations of desire are directed specifically at the hedonic theory, and Schroeder admits they do not

60 Note that this is a problem for some moral theories that claim that knowledge is intrinsically valuable – if the value is not simply that a proposition is known by someone, but value is added for each person who learns a proposition, then it may be that if relatively simple machines can have beliefs, we morally ought to add as much memory capacity as possible to all of them, so we can make them "know" a huge number of facts irrelevant to their purpose – a clearly counterintuitive result.

61 Schroeder (2004), p. 78.

62 Schroeder (2004), p. 81.

63 Berridge (2003).

apply to anything other than "what people commonly denote by ‘pleasure’ and ‘displeasure’."[64] We have identified such structures in humans, and have identified the parallel structures in other animals relatively similar to humans. It becomes controversial with animals with nervous systems more dissimilar to human biology – there has been a great deal of debate about the degree to which fish, for example, possess the capacity to feel anything like desires. But this result puts the borderline where we would expect, where our intuitions about desire and morality are unclear, giving evidence that we are tracking the correct properties. Though more scientific research remains to be done, this is at least a viable research project.

As for synthetic artificial intelligences, there remain some questions about what it would take for them to have desires in this sense. Plausibly, neurons and synapses are just one instantiation of this structure, and the phenomenological properties could be reproduced by similar structures made from silicon and metal. But it would have to be very complex to truly approximate the structure of the brain of even the simplest creatures that seem to have the desire sensations. The structure might be instantiable as a program, but again, it would have to be an incredibly complex program, requiring an unimaginably powerful computer to run, far beyond the capacity of any currently existing machine. According to Anders Sandberg and Nick

64 Schroeder (2004), p. 82.

Bostrom, although we are close to being able to run programs that simulate the arrangement of neurons present in the brains of animals such as rats and cats, current simulations run on immensely powerful supercomputers, and even then "most are a hundredfold to a thousandfold slower than biology,"[65] and "achieving the performance needed for real-time emulation appears to be a... serious computational problem."[66] We are not in danger of assigning desires to entities that obviously fail to have them, like modern desktop computers and smartphones. But even accomplishing this emulation would only produce a structure that might potentially have the capacity to instantiate desires, not necessarily something that actually possesses desires. We can assume "that this is the appropriate level of description of the brain, and that we [will] find ways of accurately simulating the subsystems that occur on this level," but ultimately "we are still largely ignorant of the networks that make up the brains of even modestly complex organisms."[67] In fact, these types of emulations might never be enough, and some of the functions of the human brain might be impossible for any computer program to truly emulate.[68] There remain a number of questions about what it would take for us to conclude that a machine likely possesses desires.

---

65 Sandberg and Bostrom (2008), p. 72.
66 Sandberg and Bostrom (2008), p. 81.
67 Sandberg and Bostrom (2008), p. 83.
68 As argued by Lucas (1961), Dreyfus (1972), and Penrose (1994), among others.

Another possible criterion, suggested by Schroeder, is that actual physical alteration and reinforcement to neural patterns through learning is crucial to desire. If that correlates with desire experiences, it might be the case that the substance of the "brain" would have to be adaptable to possess desires. An artificial brain might pass if nanites could modify it in the right ways, for example. That might be sufficiently similar to what takes place in an organic brain that is capable of desiring. It might instead be argued that requirments like that show that conscious experiences such as desires can only truly be instantiated in organic matter, as Searle claims.

But structures like these are clearly neither necessary nor sufficient conditions for having any particular dispositions to action, and not even for passing the Turing Test. As the performance of modern chatbots suggests, machines with structures completely dissimilar to anything that might possess desires could even pass the Turing Test. And conversely, we could in principle design a machine, made out of whatever substance we like, that has the desire structures but has no dispositions to actions that we would find appropriate to the desires, and thus have desires that are not detectable by the Moral Turing Test. We could create an entity like Strawson's "weather watchers", feeling desire but never reacting or moving at all, deserving

moral standing but not inspiring sympathy from others. This means that the Moral Turing Test is of no use on its own in determining whether machines ought to be granted moral standing. It could only be contingently reliable, in cases where we already know we have created a robot where the presence of actual desires correlates with the behaviour normally associated with them. It could never be the answer to how we determine whether a machine deserves moral consideration, it could only provide a guide of to how act in particular circumstances once we have answered that question already. But we are far from being able to do so. And even if we did, it might be dangerous to become conditioned to complacently rely on assuming that all and only robots that seem like they have desires actually do, when there is always the potential for someone to make a machine with a mismatch between outward behaviour and inner feeling.

Section IV: Some Practical Conclusions

We have seen that if capacity for desires is the correct basis for assigning moral standing, then under the most plausible theory of desire compatible with moral theory, it is possible for a machine to have desires and not show it. It is also possible for a machine to pass the Turing Test and seem fully

human, yet still fail to have desires. But we are not equipped to detect the relevant properties directly, and so it is very tempting to continue to rely on the Moral Turing Test. After all, relying on our ability to recognize desires in other beings has served us well for millions of years of evolution. It will not be easy to simply ignore our impulse to believe the evidence of our senses. It seems, then, that we should avoid creating structures that might be capable of instantiating desires if we are not certain that they in fact do – and also avoid creating human-seeming robots until we are better able to reliably determine the presence of specific phenomenal states. Otherwise, we will be exposed to situations where we might make wrong decisions about whether to consider these entities in our moral deliberations. Some might say that we could simply adopt a policy of "better safe than sorry" – we should just treat all machines that seem human as though they are human, and avoid harming them or doing anything impermissible to them. If they do have desires, we avoid acting immorally, while if they don't, no harm done. However, this is not an attractive option, since it is not always costless to avoid treating machines in ways that would be impermissible if they had desires. Often we must make choices about which of several entities would receive treatment that would be harmful, or which will receive (and which will be deprived of) what would be a benefit, if they were capable of being harmed and benefitted. We cannot simply treat robots as people "just to be

safe" without risking causing senseless harm to humans if we are wrong.

Perhaps then we should just ensure that the androids we make do not have desires, but we can still make them look as human as we wish. As long as people are aware that these entities do not have genuine desires, they will know how to act appropriately, one might argue. However, I do not think this information would be effective at preventing people from acting wrongly, and I think we should avoid making robots that appear too human. Even with relatively simple robots, people develop strong emotional connections – Matthias Scheutz tells of soldiers who form strong attachments to ordinance disposal robots, worryingly, since this seems like exactly the sort of situation that could lead to dangerous consequences from incorrect moral deliberation.[69] He also tells of other similar situations with social and personal care robots. It seems that we can remind ourselves that they are unthinking automata only with difficultly. If a machine were to be made that was nearly indistinguishable from a human, I doubt we could prevent ourselves from thinking of it as human. Rob Sparrow argues that an immediate, primitive moral reaction is required before we ought to treat a machine as a person. However, contrary to Sparrow's arguments, such a response is just as inevitable towards a robot that acts completely human as it is towards an actual human, and yet that response is undeserved and

69 Scheutz (2012).

might lead to disastrous consequences if directed towards a robot that did not possess desires.

Even those who manage to resist the urge to treat such machines as persons and succeed in reminding themselves that these machines are not worthy of moral standing might be socially pressured into doing so. The science-fiction examples that make attitudes towards artificial intelligences an allegory for racism show how severe this might be – if it has been continuously reinforced that certain claims one might make about the correct treatment of machines are analogous to ones that might be made about treatment of other races, and that discrimination against robots is tantamount to racism, that would likely produce severe stigma for anyone who would deny moral personhood to completely convincing AIs. People quite rightly do not want to be racist and treat other races unequally, and merely due to perceived parallels they might treat AIs as equals just to avoid being labeled as some kind of future sci-fi space racist. Therefore, if capacity for desires is the criteria for moral standing, we should avoid making human-seeming robots until we can reliably determine whether or not a given entity has the relevant phenomenal states.

However, this course of action might not be costless either. There are great

advantages to having robots that can be interacted with socially, for example as therapeutic robots to help the elderly and individuals with developmental disorders.[70] Even in cases where social ability is not an obvious requirement of a machine's function, it is a significant advantage in terms of ease-of-use to be able to issue commands and have them confirmed in a natural, familiar way – this is why computer interactivity has moved from punch-cards and entering lines of code towards programs like Apple's Siri. The temptation might be to go further, but I believe we should take a careful look at the actual gains we would acheive from making machines that are even more convincing human analogues and social beings compared to the increased risk of making incorrect moral attributions and the consequences that would arise from that. I suspect that we will reach a point of diminishing marginal utility. Another solution might be to ensure that the robots we cannot help but treat as having moral standing are in fact deserving of that status. We will then, of course, have to determine how to avoid causing harm to them and the costs associated with that.

In fact, giving robots desires might sometimes be advantageous even when they are not required to interact with humans. It is quite plausible that desires are a large part of what makes humans and other sophisticated

70 See Robins, Dautenhahn, Boekhorst, and Billard (2005), and Kidd, Taggart, and Turkle (2006).

organisms successful and effective. Robots that are required to learn and adapt might benefit from having real desires. This raises complex questions for cases such as automated bomb detection/disposal robots whose function is inherently dangerous. Even if they have moral standing, we might be able to minimize bad consequences if we are careful not to imbue them with desires that are likely to be frustrated by injury and death the way most humans' desires are. But it is far from obvious that we are in any position to be sure of which desires we are giving them and how to avoid giving them the unwanted ones. This highlights another potential problem: we are not presently equipped to reliably determine when an AI would have desires, so as we give robots more and more complex structures to fulfill complex roles, we might be in danger of giving them desires without realizing it, and harming them without knowing. It is not desirable to avoid giving any structure that might potentially lead to the capacity for desires, since that might rule out useful and necessary features. As computer techology advances, we have an increasingly urgent need for a reliable way to identify the features that give rise to the relevant phenomenological states.

We could, of course, adopt a different moral theory, that does not make reference to desires. In that case, we need to determine what those theories will claim are the properties are necessary for moral standing, and how we

could tell when an entity possessed them.

Bibliography

Anscombe, G.E.M. (1957). *Intention*. Harvard University Press.

Bentham, Jeremy (1823). *Introduction to the Principles of Morals and Legislation*, second edition. Hafner.

Berridge, Kent (2003). "Pleasures of the brain". Brain and Cognition 52, pp. 106–128.

Callard, Agnes Gellen (2008). *An Incomparabilist Account of Akrasia*. Proquest.

Churchland, Paul (1979). *Scientific Realism and the Plasticity of Mind*. Cambridge University Press.

Churchland, Paul (1981). "Eliminative materialism and the propositional attitudes". Journal of Philosophy 78, pp. 67–90.

Churchland, Paul and Churchland, Patricia (1990). "Could a machine think?". Scientific American 262:1 (January 1990).

Clark, Andy (1987). "From Folk Psychology to Naive Psychology". Cognitive Science 11, pp. 139-154.

Davidson, Donald (1980). *Essays on Actions and Events*. Oxford University Press.

Dennett, Daniel (1993). *The Intentional Stance*. MIT Press.

Dreyfus, Hubert (1972). *What Computers Can't Do*. MIT Press.

Kidd, Cory; Taggart, Will; and Turkle, Sherry (2006). "A sociable robot to encourage social interaction among the elderly". Proceedings of the 2006 IEEE International Conference on Robotics and Automation, pp. 1050–4729.

Lucas, J.R. (1961). "Minds, machines, and Godel". Philosophy 36:112-27.

Minsky, Marvin (1980). "Decentralized Minds". Behavioral and Brain Sciences 3:3.

Morillo, Carolyn (1990). "The Reward Event and Motivation". The Journal of Philosophy, 87:4 (April 1990), pp. 169-18.

Nagel, Thomas (1970). *The Possibility of Altruism*. Oxford University Press.

Parfit, Derek (1984). *Reasons and Persons*. Oxford.

Penrose, Roger (1994). *Shadows of the Mind*. Oxford.

Pylyshyn, Zenon (1986). *Computation and Cognition*. MIT Press.

Robins, Ben; Dautenhahn, Kerstin; Boekhorst, R.T.; and Billard, Aude (2005). "Robotic assistants in therapy and education of children with autism: Can a small humanoid robot help encourage social interaction skills?" Universal Access in the Information Society 4:2, pp. 105–120.

Ross, W. D. (1930). *The Right and the Good*. Oxford University Press.

Sandberg, Anders and Bostrom, Nick (2008). *Whole Brain Emulation: A Roadmap.* Technical Report #2008-3, Future of Humanity Institute, Oxford University.

Scheutz, Matthias (2012). "The Inherent Dangers of Unidirectional Emotional Bonds Between Humans and Social Robots". In *Robot Ethics*, ed. Patrick Lin, Keith Abney, and George A. Bekey. MIT Press.

Schroeder, Timothy (2004). *Three Faces of Desire.* Oxford University Press.

Schwitzgebel, Eric (1999). "Representation and desire: A philosophical error with consequences for theory-of-mind research." Philosophical Psychology 12:2, pp. 157–180.

Searle, John (1980). "Minds, Brains and Programs". Behavioral and Brain Sciences 3:3.

Simmons, Aaron (2009). "Do Animals Have an Interest in Continued Life? A Desire-Based Approach". Environmental Ethics 31 (Winter 2009), pp. 375-392.

Smith, Michael (1987). "The Humean Theory of Motivation". Mind 96, pp. 36–61.

Smith, Michael (2011). "Deontological Moral Obligations and Non-Welfarist Agent-Relative Values". Ratio 24:4 (December 2011), pp. 351-363.

Sparrow, Rob (2004). "The Turing Triage Test". Ethics and Information Technology 6.

Sparrow, Rob (2012). "Reflections on the Turing Triage Test". In *Robot Ethics*, ed. Patrick Lin, Keith Abney, and George A. Bekey. MIT Press.

Stalnaker, Roger (1984). *Inquiry.* MIT Press.

Stampe, Dennis (1986). "Defining desire". In *The Ways of Desire*, ed. J. Marks. Precedent.

Strawson, Galen (1994). *Mental Reality*. MIT Press.

Thagard, Paul (2006). "Desires are not propositional attitudes". Dialogue 45, pp. 151–156.

Turing, Alan (1950). "Computing Machinery and Intelligence". Mind 49:236 (October 1950).

Worley, Sara (1997). "Belief and Consciousness". Philosophical Psychology 10:1, pp. 41-55.

Ethics Inf Technol (2013) 15:239–241
DOI 10.1007/s10676-012-9308-8

BOOK REVIEW

Mark Coeckelbergh: Growing moral relations: critique of moral status ascription

Palgrave Macmillan, New York, 2012, 239 pp, ISBN: 978-1-137-02595-1

David J. Gunkel

Published online: 28 February 2013

In the *Structure of Scientific Revolutions*, Thomas Kuhn (1962) famously distinguished between what he called "normal science" and those rare but necessary instances of paradigm shift, when there is a fundamental transformation in the very rules of the game. I do not think it is hyperbole to say that Mark Coeckelbergh's *Growing Moral Relations: Critique of Moral Status Ascription* is a book that is situated at, and contributes to what can only be described as, a significant paradigm shift in moral thinking. It is, in other words, a real game changer.

The subject of Coeckelbergh's book is moral status ascription, or more precisely the problem of deciding *who* is morally significant and *what* is not. These two small, seemingly simple words, as Jacques Derrida (2005, p. 80) has reminded us, are not unimportant. They make all the difference, for they distinguish between persons who have moral standing and what are considered to be mere things. This decision (quite literally a cut that is made within the very fabric of existence) is typically enacted and justified on the basis of the intrinsic properties of the entity in question. Coeckelbergh calls this transaction "the properties approach to moral status ascription," and the book begins with a penetrating analysis and critique of this "normal science."

The properties approach is rather straight forward and intuitive: "identify one or more morally relevant properties and then find out if the entity in question has them" (p. 14). But as Coeckelbergh insightfully points out, there are at least two persistent problems with this undertaking. First, how does one ascertain which properties are sufficient for moral status? In other words, which one, or ones, count? The history of moral philosophy can, in fact, be read as something of an on-going debate and competition over this matter with different properties—rationality, speech, consciousness, sentience, suffering, etc.—vying for attention at different times. Second, once the morally significant property has been identified, how can one be certain that a particular entity possesses it, and actually possesses it instead of merely simulating it? This is tricky business, especially because most of the properties that are considered morally relevant are internal mental states that are not immediately accessible or directly observable from the outside. In other words, even if it were possible to decide, once and for all, on the right property or mix of properties for moral standing, we would still be confronted and need to contend with a variant of the "other minds problem."

The history of moral philosophy can be interpreted as an ongoing (and perhaps even an inconclusive) struggle to respond to and resolve these two problems. And significant developments in the field are often the result of identifying some new criteria of inclusion. This is evident, for example, in animal rights philosophy, which, following the suggestions of Jeremy Bentham (2005), shifted attention from an active ability, like speech or conscious thought, to the passive capability of "Can they suffer?" (p. 283). Similarly recent efforts to address the moral status of machines, like artificial intelligence systems and robots, often propose clever reconfigurations of the Turing Test (see Allen et al. 2000; Sparrow 2004) to help ascertain whether an artificial entity actually possesses a particular property or not.

Coeckelbergh, although recognizing the importance of these issues and debates, does not engage the question of moral status at this level. He realizes that this endeavor, although having the weight of tradition behind it, can only produce what G. W. F. Hegel called *schlechte Unendlichkeit*, a bad or spurious infinity that simply turns

D. J. Gunkel (✉)
Dekalb, IL, USA
e-mail: dgunkel@niu.edu

out more of the same. Consequently, Coeckelbergh seeks to move moral thinking beyond this intellectual *cul-de-sac* by proposing an alternative paradigm that proceeds otherwise—where the word "other" indicates another method that is able to accommodate other forms of otherness. Coeckelbergh, therefore, addresses the problem of moral status ascription not by playing according to the rules of the game—what Kuhn would call moral philosophy's normal science—but by challenging and even changing the terms and conditions of the game itself.

Coeckelbergh's alternative approach to moral status ascription can be described by three terms: relational, phenomenological, and transcendental. By "relational," he emphasizes the way moral standing is always and already social, meaning that moral status is not something that is located in the inner recesses of an individual entity but transpires through the interactions and relationships situated between entities. This "relational turn," which Coeckelbergh skillfully develops by capitalizing on innovations in ecophilosophy, Marxism, and the work of Bruno Latour, Tim Ingold, and others, does not get bogged down trying to resolve the epistemological problems associated with the standard properties approach. Instead it recognizes the way that moral status is socially constructed and operationalized. But Coeckelbergh is not content simply to turn things around. Like Friedrich Nietzsche, he knows that simple inversions (in this case, swapping the relation for the relata) changes little or nothing. So he takes things one step further. Quoting the environmental ethicist Baird Callicott (1989), Coeckelbergh insists that the "relations are prior to the things related" (p. 110). This almost Levinasian gesture is crucial insofar as it undermines the usual way of thinking. It is an anti-Cartesian and postmodern (in the best sense of the word) "intervention." In Cartesian modernism the individual subject had to be certain of his (and at this time the subject was always gendered male) own being and his own intrinsic properties prior to engaging with others. Coeckelbergh reverses this standard approach. He argues that it is the social that comes first and that the individual subject (an identity construction that is literally thrown under or behind), only coalesces out of the relationship and the assignments of rights and responsibilities that it makes possible.

This relational turn in moral thinking is clearly a game changer. But like all conceptual innovations of this type, it is necessarily and unavoidably exposed to reappropriation and domestication by the dominant system it confronts and contests. And Coeckelbergh knows this. In fact, he explicitly recognizes and points out that the relational approach always runs the risk of "ontologicalization," where the relation is turned into just another kind of relata. For this reason, he insists on the second descriptive term, "phenomenological." Coeckelbergh's "relationalism" (a word that he uses to characterize his own position) is phenomenological insofar as it is concerned with what appears within the space and time of social reality and not with the usual metaphysical suspects, that is, those essential features that are commonly assumed to subsist "behind" or "underneath" the phenomena. "What matters, morally speaking, is not what the entity in question *is*, but how the entity appears" (p. 24) and how it has been situated and functions within a particular social ecology. Coeckelbergh, therefore, develops an approach to moral status that is not interested in probing the profound metaphysical depths of the *Ding-an-sich*. Instead his approach is and can be radically superficial. Understood in this way, moral status is not rooted in the ontology of an individual entity. It is something that is encountered and decided on the basis of "how we experience and construct the entity, how it appears to our consciousness and how we give it reality, meaning and status" (p. 25).

As innovative as this appears, Coeckelbergh recognizes that this strategy cannot be credited as a new moral ontology. In fact, as he argues, "the relational approach can only constitute an attractive alternative paradigm if it is not understood as an alternative moral ontology, in the sense of a better description of moral reality" (p. 6). In an effort to avoid the pitfalls of this kind of "dogmatic interpretation" that would effectively eviscerate the innovations of the moral-relational project, Coeckelbergh adds a third term to his characterization, "transcendental." Coeckelbergh's project is transcendental, in the strict Kantian sense of the term. He is, therefore, not interested in making indubitable claims about the true nature of moral reality but is concerned with tracking and exhibiting the condition for possibility of moral status ascription. For this reason, the operative question is not "What is moral status?" but "What are the conditions for an entity to appear as having a certain moral status?" or "What are the conditions for moral status ascription/construction?" (p. 7). And in the second half of the book, Coeckelbergh takes up and investigates a number of contributing factors: language and the way moral status not only has a particular linguistic form but is already partially given in and by language, socio-cultural structures that already inform and influence the way we encounter and contend with others, material and bodily conditions that place very real restrictions on what is possible and constitute the very matter of things, moral geographic patterns of domestication and alienation endemic to the colonial and postcolonial legacy of Western thought, and spiritual conditions whether formulated in religious or secular terms. In pursuing these transcendental conditions of possibility, Coeckelbergh seeks to account for the diverse constellation of forces that both make moral status ascription possible and also limit the range and reach of moral discourse.

Coeckelbergh tackles all of these items with remarkable insight, knowledge, and dexterity. The analyses he provides are insightful and detailed without losing sight of the big picture. The sequence and structure of the chapters is logical and organized without ever becoming formulaic or risking predictability. And the list of scholars he mobilizes in this effort is wide-ranging and includes major figures from both the analytic and continental traditions. If there is one thing that I could say by way of criticism of this thoroughly impressive work, it would only be a "sin of omission." Despite the fact that Coeckelbergh is able to marshal an inspiring list of thinkers—Diogenes, Plato, Kant, Marx, Heidegger, Searle, Wittgenstein, Latour, etc.—one name is absent, Emmanuel Levinas. This is unfortunate, because so much of what Coeckelbergh develops under the rubric of his relational/phenomenological approach resonates with Levinas's ethics of otherness. The pivotal Levinasian (1969) claim that "ethics proceeds ontology" is, for instance, remarkably close to and shares important points of contact with Coeckelbergh's insistance that "the relations are 'prior' to the relata" (p. 45).

This is not to say that Levinas somehow trumps Coeckelbergh. In fact, Levinas's philosophy has its own set of problems, not the least of which is its virtually unquestioned allegiance to humanism and the residue of a human exceptionalism that has the unfortunate effect of excluding both animals and machines from moral consideration. Unlike Levinas, Coeckelbergh's *Grown Moral Relations* is, in fact, able to accommodate and make a place for these others. For this reason, Levinasian thought not only can be used to bolster Coeckelbergh's relationalism by supplying additional evidence and perspective, but, and perhaps more importantly, Coeckelbergh's text can provide critical perspective on and a useful corrective to Levinas's humanism.

In conclusion, Mark Coeckelbergh has provided a penetrating analysis of moral status that does not simply seek to challenge or even expand the circle of inclusion. Instead his book critiques the standard operating presumptions of moral status ascription and proposes an innovative alternative that is able to circumvent many of the problems that have plagued the individual properties approach typically employed to decide these matters. For this reason, *Growing Moral Relations* is a book that should be of interest to a wide and diverse audience. It clearly has a great deal to contribute to recent debates concerning the contested ethical status of others, especially those other kinds of others, like animals, the environment, and the increasingly intelligent machines of our own making. But it also provides an opportunity for perceptive critical reflection on the history of moral philosophy, the standard approach to moral status ascription, and the problems endemic to this legacy system.

References

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

Bentham, J. (2005). In: J. H. Burns & H. L. Hart (Eds.), *An introduction to the principles of morals and legislation*. Oxford: Oxford University Press.

Callicott, J. B. (1989). *In defense of the land ethic: Essays in environmental philosophy*. Albany, NY: State University of New York Press.

Derrida, J. (2005). *Paper machine*. (R. Bowlby, Trans.). Stanford, CA: Stanford University Press.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Levinas, E. (1969). *Totality and infinity* (A. Lingis, Trans.). Pittsburgh, PA: Duquesne University Press.

Sparrow, R. (2004). The turing triage test. *Ethics and Information Technology, 6*(4), 203–213.

Machines and the Moral Community

Article *in* Philosophy & Technology · March 2014

DOI: 10.1007/s13347-013-0114-y

1 author:

Erica Neely
Ohio Northern University
10 PUBLICATIONS **7** CITATIONS

Machines and the Moral Community

Erica L. Neely

Abstract: A key distinction in ethics is between members and non-members of the moral community. Over time, our notion of this community has expanded as we have moved from a rationality criterion to a sentience criterion for membership. I argue that a sentience criterion is insufficient to accommodate all members of the moral community; the true underlying criterion can be understood in terms of whether a being has interests. This may be extended to conscious, self-aware machines, as well as to any autonomous intelligent machines. Such machines exhibit an ability to formulate desires for the course of their own existence; this gives them basic moral standing. While not all machines display autonomy, those which do must be treated as moral patients; to ignore their claims to moral recognition is to repeat past errors. I thus urge moral generosity with respect to the ethical claims of intelligent machines.

1. Introduction

A key distinction in ethics is between members and non-members of the moral community; this is the foundation for understanding how we should treat the entities we encounter in the world. Over time our notion of this community has expanded; those we take as non-members have changed, and the criteria used to make that distinction have also altered. Historically, as surveyed by Lorraine Code (1991), Charles Mills (1999) and Naomi Zack (2002), criteria such as intellect and rationality were used to separate white men from women and non-whites. Taken to be governed primarily by emotion rather than rationality, these people were seen as moral inferiors, deserving of lesser or no moral consideration.

Even upon conceding that rationality was not the exclusive preserve of white men, and so including women and non-whites as members of the moral community, many continue to deny moral standing to animals. Both contemporary thinkers (Scruton 2006) and earlier philosophers (Kant 1786/1996) see humans as having the moral high ground of rationality and consciousness. However, rationality criteria raise questions as to how rational a being must be to receive moral standing – there is a serious risk of excluding certain humans (such as infants) from the moral community which, as discussed by Peter Singer (2002), is unpalatable to many thinkers. Furthermore, our understanding of the biological similarities between humans and other animals makes it difficult to maintain a sharp distinction between them; various other animals seem to possess degrees of rationality and consciousness as well.[1] Such reasoning has caused many (Bentham 1823/1996; Taylor 1996; Singer 2002) to move to sentience as the criterion for moral standing: if something can feel pain, it is wrong to take intentional action to make it suffer unnecessarily.[2]

This is a large expansion to the moral community, yet of course many things continue to lack moral standing; an object such as a table or chair is not a member of the moral community, for instance, because it is not possible to cause moral harm to the object itself. Unless the object belongs to someone else, I can do what I wish to it; the only kind of moral harm that can be

[1] For instance, the National Institute of Health (2013) has recently designated chimpanzees as inappropriate for most forms of animal research, since they are our closest relatives and "are capable of exhibiting a wide range of emotions; expressing personality; and demonstrating individual needs, desires, and preferences." The sort of clear distinction between human and non-human animals once thought to exist is increasingly being challenged, giving rise to new ethical implications.

[2] Obviously there is clarification required to specify what constitutes unnecessary suffering and exactly how much moral standing animals have. However, sentience suffices to give them a foot in the door of the moral community, so to speak.

caused in this situation is harm to a person or persons who have a claim to that object.[3] As such, there is currently a strong ethical divide between living beings and non-living things. This has serious implications for the ethical issues pertaining to intelligent machines, since they too are inanimate objects. There is a strong temptation to classify them as similarly undeserving of any moral standing.

I will argue that the relevant criterion for membership in the moral community should not be understood as whether one can feel pain but rather whether something has interests. While sentience is certainly one way of having interests, it is not the only one. Using this criterion, I will show that certain kinds of machines are, in fact, members of the moral community; specifically, I will argue that they are moral patients.[4] Thus while we are correct in extending membership in the moral community to encompass humans and animals, we must further extend it further to include these machines.

2. Ethics and the Prevention of Harm

When conversing with people, one informal objection that frequently occurs to granting moral standing to a machine is the claim that you cannot "hurt" a machine. In essence, this is an internalization (and over-simplification) of the sentience criterion for moral standing. Ethics is often taken to involve the prevention of harm; as David Gunkel (2012) notes, the central question of moral patiency often is phrased as whether something can suffer. Hence if something cannot be harmed, many are reluctant to offer moral standing to the thing in question.

For humans, the harm generally involves some kind of pain. However, the ability to feel physical pain cannot be the only criterion for membership in the moral community. Consider a person with congenital analgesia, i.e., one who is unable to register physical pain. If someone were to step on his foot, he would not be able to feel any pain from the action. Yet it would surely be wrong if one stepped on him simply because one took a kind of perverse amusement in his inability to feel it. Stepping on his foot intentionally, without his permission, and without some kind of greater justification strikes us as wrong.[5]

This is not because the action caused pain (since, by design, it does not); sentience as construed to involve purely physical sensations is not sufficient to render this action wrong. Furthermore, even if we extend sentience to include mental or emotional pain, it is still insufficient; it is wrong to cause harm even if the victim is emotionally unmoved, such as when we see emotional dissociation in child soldiers or victims of abuse. The wrongness in our case stems from two key points. First, the action could cause damage, even if it does not cause pain. Second, since we have specified that the person does not give permission for the action, deliberately stepping on his foot violates his desire to remain unmolested; moreover, there is little justification for this violation.

---

[3] The ownership of an object could be the community as a whole, such as with public art installations. If someone were to destroy the Vietnam Veteran's Memorial, one could argue that it would cause harm to the public (which has a claim on the memorial) and is thus morally wrong. It would be odd to say that you had morally wronged the monument itself, however.

[4] I am concerned in this paper with what it takes for a machine to be deserving of rights and hence be a moral patient. I leave open the question of what it would take for a machine to have moral responsibilities and thus be a moral agent.

[5] This action might be justified if it were done out of a different motivation. Even if I lack his consent, deliberately stepping on his foot might be acceptable if it prevented a greater harm (such as stepping into the path of a vehicle.) However, this is a rather different case than interfering with another's body simply because it entertains me.

What is necessary for moral standing is not sentience per se but having interests; the person in our congenital analgesia example lacks sensation, but he retains interests. As it is possible to harm those interests, it is possible to harm him. To expand, consider John Basl's definition of interests as "those things the satisfaction of which contributes to [an individual's] welfare." (Basl 2012) This implies that a being can have interests without being aware of them. For instance, an ant may have an interest in not being stepped on and killed even if the ant is unaware of that interest; similarly for a person in a persistent vegetative state.[6] In each case, their welfare can be threatened regardless of whether they are aware of that threat. Of course, many times we are aware of our interests – we have ideas about how we wish to run our lives, and thus have some interest in those plans being followed; we may be harmed if our desires are simply ignored.

The notion of harm that is relevant for morality, therefore, moves beyond physical pain and hinges on the idea of disrespecting the integrity and autonomy of the individual. The possibility of the action causing damage, even if it does not cause pain, raises the idea of bodily integrity. At a minimum, beings have an interest in retaining sufficient bodily integrity for continued existence; anything which damages one's body threatens this interest. This interest can certainly be outweighed by other factors – I may consent to having my appendix removed because that particular violation of bodily integrity actually promotes my continuation under certain circumstances. Frequently in medicine we consent to actions which are extremely damaging to our bodies (such as chemotherapy) if the alternatives are worse.[7]

In addition to these dramatic cases, we consent to small violations of bodily integrity on a regular basis; it is clearly possible to overstate our commitment to it, since most people trim their fingernails or their hair or will pick open the occasional scab. Yet people are unlikely to see those actions as presenting any serious threat to continued existence. Hence one might argue that a minor harm, such as stepping on a person's foot, cannot truly be objected to on this basis alone. Indeed, I believe that the emphasis on bodily integrity dovetails with the desire to remain unmolested mentioned above; together they highlight the fact that we have certain wishes about the shapes of our lives.

By ignoring the person's desire not to be trod upon, the aggressor's action violates his autonomy. In much of ethics, autonomy is emphasized as an important good.[8] To cast it aside for no reason other than to satisfy one's own sadistic desires is to jeopardize the interest of the injured person in governing the course of his own life. Such an action may not cause physical pain, but it clearly causes harm to that person – it treats him as incapable or unworthy of directing his own actions, and views his desires as irrelevant and something that may simply be

[6] This is why it would, for instance, be wrong to take pornographic photos of a person in a persistent vegetative state; we believe that a person can be harmed even if he or she is unaware of it.

[7] One could also justify suicide this way for some cases, since my interest in bodily integrity could be outweighed by an interest in avoiding large amounts of suffering from a terminal disease, say. While we have an interest in bodily integrity, it is not the only interest that matters.

[8] We see this both in Kant (1786/1996) with the view of rational beings as ends-in-themselves and in Mill (1859/1993) with the emphasis on individual liberty.

ignored.[9] Although it is clear that sometimes a person's desires must, ethically, be overridden, we surely cannot ignore another's wishes completely. [10]

Hence while sentience certainly leads to having interests, it is not necessary for them: the joint properties of consciousness and self-awareness will also suffice.[11] Once a being is self-aware and conscious, it is aware of its self, can desire continuation of that self, and can formulate ideas about how to live its life.[12] It is possible to harm such a being by ignoring or thwarting those desires; one should not act against the being's wishes, therefore, without some overriding reason. The requirement of such a reason, however, is equivalent to granting the being at least minimal moral standing; one does not need to have a reason to destroy a chair, but one must provide such a reason to destroy a human. This holds true for intelligent machines just as much as for a person with congenital analgesia; they both have interests and desires, hence they both have basic moral standing.

One could object at this point that we have moved too quickly from consciousness to ascribing desires to a machine. Basl discusses the possibility of a machine which is conscious only insofar as it has the ability to experience colours. However, it has no emotional or cognitive responses to those experiences – it may experience blue, but it does not care. Basl claims that it would not be wronging such a machine if we were to shut it down. Similarly, suppose a being existed which could feel pain but had no aversive reaction towards it; furthermore, this is the only conscious experience the being has. In this case, the being would presumably have no interest in avoiding pain, and Basl believes that it would not be wrong to cause pain to it. In each of these cases, Basl argues, there is consciousness without moral patiency. Consciousness, understood as the ability to have sensory experiences, is not sufficient for having interests – instead one must have the capacity for attitudes towards those experiences. (Basl 2012)

Basl's view of consciousness is extremely limited, however. Steve Torrance (2012) notes that "To think of a creature as having conscious experience is to think of it as capable of experiencing things in either a positively or negatively valenced way – to think of it as having desires, needs, goals, and states of satisfaction and dissatisfaction or suffering." I believe that this robust notion of conscious experience is more in line with what we mean when we consider conscious machines. Basl is correct in arguing that the liminal cases he describes are likely not instances of moral patiency. Furthermore, there is importance to considering these instances, since it seems probable that the first conscious machines will be more akin to the machine which can experience colour than any others. However, the first machines we recognize as conscious will likely be ones which exhibit consciousness of the sort Torrance describes, and these machines

[9] While I will not rehearse the arguments for each ethical theory in detail, note that ignoring a person's desires for his life will fail to calculate the utility/disutility generated by particular actions, will treat the person as a means to an end, is certainly not something rational people are likely to consent to from behind a veil of ignorance, and demonstrates a lack of care, compassion, and benevolence. None of these ethical theories will condone simply ignoring the desires of a person, although they will almost certainly allow us to take actions counter to those desires in many cases.

[10] This is one reason why advance directives are important, even if fraught with complications: they allow a person to express her wishes in advance to cover circumstances (such as being in a coma) where she cannot do so directly.

[11] An interesting discussion of the connection between self-awareness and moral standing (or personhood, as she puts it) can be found in Mary Anne Warren's discussion of personhood and abortion (Warren 1973) as well as in Scruton (2006).

[12] It might be that consciousness is also unnecessary for having interests, particularly if we consider an Objective-List view of welfare, as Basl (2012), notes. Hence the category of moral patients may extend slightly further than I argue for here; any machine with interests will count, although I am only arguing here that conscious and self-aware machines have interests.

clearly would have the necessary propositional attitudes in order to have interests.[13] As such, my previous argument stands, and such machines would have moral patiency.[14]

3. Intelligence and Autonomy

Thus far I have argued that having interests is what is necessary for moral standing. Since conscious, self-aware machines have interests, they also are moral patients. In general, however, the question of moral standing for machines is raised in the context of artificial intelligence – would an intelligent machine have moral standing? To provide an answer to this general question, we must ask whether we can assume that intelligent machines are conscious and self-aware. If so, we have addressed the moral standing of all intelligent machines; if not, then further work is necessary to clarify the status of the remaining machines.

To respond to this, we must consider what is meant by an intelligent machine. Shane Legg and Marcus Hutter have gathered many of our informal definitions of intelligence and used them to devise a working account of machine intelligence. (Legg and Hutter 2006a, 2006b, 2007) Informally, their definition of intelligence is "Intelligence measures an agent's ability to achieve goals in a wide range of environments."[15] (Legg and Hutter 2007) One key question that emerges from this definition is who determines the goals of the agent. There are two possibilities: one, the agent's goals are always determined by an outside source or, two, the agent's goals are not always determined by an outside source.

Consider the case where the agent's goals are always established by an outside source. In this case, the goals are communicated to the agent in some fashion, and the agent simply uses its resources to accomplish whatever goals it has been given. For instance, my computer takes actions based on user input and the commands dictated by its programming; its actions are always ultimately determined by a human. Such an agent lacks autonomy.[16] Since the agent lacks self-awareness and lacks the ability to formulate goals for itself, the argument for moral standing does not apply; it will not have a desire for continuation or any wishes as to how to live its life. As such, it is in the same category with chairs and tables mentioned above and lacks moral standing; it is not clear how one could harm or benefit such an entity.[17]

---

[13] I believe that we are more likely to recognize as conscious a machine which has a robust consciousness since that consciousness is more like our own and thus more apt to display behaviors which match up with the conscious behaviors of humans. It is far from clear how we would ever determine that machine had an awareness of colors if that were the full extent of its consciousness. Hence while we may create such limited machines, I suspect we will not realize we have done so.

[14] Marie-des-Neiges Ruffo (2012) would likely object to this conclusion as she believes that machines are not things which are capable of well-being or ill-being because they lack human feelings. I find this unconvincing for two reasons. First, I believe you could create a case which paralleled the congenital analgesia example and argue that it is still wrong to harm such a person even if she lacked emotion. Second, it is not clear to me why she assumes that we will never be able to create machines which have emotions. It is true that we cannot currently do so, but there was a time when everyone was certain a machine would never be able to play chess. This has, of course, proven false; as such, I find our current capabilities to be poor predictors of future ability.

[15] They provide a formal definition (Legg and Hutter 2007), however space does not permit the detailed exposition required to fully explicate this definition.

[16] I am using "autonomy" in the sense typical of ethics, meaning something akin to "being able to make one's decisions free of external influence or control;" the term is (confusingly) used somewhat differently at times in robotics.

[17] Presumably the machine is not sentient, or we could have had a much shorter argument for moral standing; as such, it cannot gain moral rights through an appeal to sentience. One might try to argue that such a being has rationality and thus, on some views of morality at least, must be granted moral standing. I am not convinced this is

Contrariwise, consider the case where the agent's goals are not always determined by an outside source, i.e., where the agent is capable of determining its own goals at least some of the time[18]. In this case, the agent is expressing a basic capacity for autonomy, which implies that these goals must be chosen by the agent itself; they cannot simply be chosen by following an algorithm or program.[19] As such, the agent must be deciding for itself what it desires to do. Once an agent is capable of exhibiting desires, however, we may collapse this into my previous argument concerning moral standing; while the agent's desires may be overridden, they may not simply be ignored.[20]

One could object that this argument is prejudiced by the use of the word "desires" – perhaps the machine is choosing what to do, one might argue, but that does not imply that the machine desires that course of action. Yet, it is not at all clear what such a choice would mean, in this case, since it cannot be determined by an algorithm or program. The machine would need to be making a decision which was in some sense its own; it could not be purely the result of an outside influence or program. Where else would the choice stem from if not from the machine's own wishes? If we have eliminated any external factors or internal compulsion, what remains is the machine's own will.[21]

One point worth noting is that moral questions are not black-and white; both autonomy and moral standing exist on a continuum.[22] The more autonomous the machine, the more duty we will have to respect its wishes; the less autonomy, the more we are permitted to act as its guardian. This is akin to how we treat children and the severely mentally disabled; they are not viewed as capable of making decisions in as many areas as fully-functioning adults, hence we do not see their desires as binding to the same extent. They still have moral standing, of course, in that it is wrong to harm them without just cause. Nevertheless, they are not granted as much governance over the course of their own lives, and we do not view overriding their wishes as comparable to overriding the wishes of other adults. In a similar fashion, a machine with greater

---

the case; while Kant sees morality as shared by rational beings, he makes it clear that the kinds of beings he is discussing have a will – the machines, as I have described them, do not. (Kant 1786/1996) In general, I believe that the rationality criterion for moral standing is more complex than simple intelligence, and machines with bare intelligence will likely not satisfy it.

[18] It is not clear whether such a machine currently exists; I suspect it does not yet, although the evolution of drone technology seems to be heading us in this direction.

[19] While the choices may be influenced by the programming of the machine, human choices are also influenced by upbringing, societal pressure, brain chemistry, and so forth. Since moral theorizing generally views human autonomy as worth preserving despite these factors, machine autonomy likewise has worth.

[20] One might also make the argument that autonomy itself is sufficient for granting something moral standing. If we view autonomy as a good, then the fact that such machines exhibit autonomy suffices to grant them at least some consideration. We may place limits on the expression of their autonomy, just as we do for people, but we likely could not simply ignore it.

[21] Note that this argument is separate from the argument of whether such machines could exist. Ruffo (2012) believes that a machine cannot deliberate; any choice it makes would be a result of programming. As such, she would argue that no machine could determine its own goals. While I am unconvinced, it is not necessary for our present purposes.

[22] A machine which is programmed to learn based on past interactions will be somewhere along this continuum, depending on the complexity of its programming; a simple program will likely result in a machine with little autonomy, but a complex program may approach the situation we have with humans. Since we also learn and adapt as a result of our interactions – following social norms, rules we have been taught, biological imperatives, and so forth – a sufficiently complex set of instructions for a machine may model this; if we consider ourselves to be at least somewhat autonomous, we must consider the machine to be as well.

autonomy likely has more claim on us to respect that autonomy, and it will be a greater moral fault if we ignore its wishes.[23]

In summary, I believe that autonomy implies that the agent has desires. My previous argument fails to apply only to intelligent machines which both lack self-awareness and consciousness and also which are not capable of setting their own goals. Such machines lack moral standing because they have no self-concept and no desires; it is implausible to hold that they could desire existence or have goals for that existence. Determining whether and to what extent a machine is autonomous will likely be difficult, however, and those who oppose granting moral standing to machines might well use this as an excuse to deny their moral worth. This is a dangerous move to make, though, since the long-standing philosophical dilemma of other minds demonstrates that it is also hard to ensure that other people have minds and are not cleverly programmed automata which simply deceive us into thinking they are conscious humans.

The problem of how to determine whether machines are conscious or autonomous is difficult. Torrance (2012) seems fairly optimistic about the prospect; assuming that consciousness is not simply some mysterious fact about the universe, then it likely hinges on facts about the structure of our brains and is exhibited in our behaviour. Hence, in general, we assume that a person is conscious because she acts in particular ways and because she has certain biological similarities to ourselves; if we take ourselves to be conscious, then a creature which acts like us and is built like us seems likely to be as well. Yet, as Basl points out, this is challenging to extend to machines because they are not like us biologically. Even if we build machines with biological components, they will not share the same evolutionary history as we do; hence it is more difficult to argue that their minds have developed similarly (and thus must also have given rise to consciousness.) (Basl 2012) Perhaps as our knowledge of what is physically necessary for consciousness in humans progresses will be able to recreate it in an artificial setting; for now, it leaves us in a bit of a quandary.[24]

In general, it is wise to err on the side of caution – if something acts sufficiently like me in a wide range of situations, then I should extend moral standing to it.[25] Joanna Bryson (2010) has argued that there is danger in being overly generous and extending rights to machines because we may waste energy and resources on entities which are undeserving of them; furthermore, this diverts our attention from the human problems which should be our concern.[26] However, I believe she is too hasty in arguing that we simply can avoid the problem by not designing robots which deserve moral concern. While she is correct that we design and build robots, it should be clear to anyone who interacts with computers or software that we do not always correctly predict

[23] The analogy is somewhat imperfect, since we tend to take children to be beings who will increase in autonomy over time; they have the potential for as much autonomy as fully-functioning adults, whereas we generally are not as optimistic about the prospects of the severely mentally disabled. However, I can see the potential for both sorts of machines: there may be some whose autonomy only ever reaches a low level and others whose autonomy develops over time. Hence the two prongs of this analogy are both useful, since I believe our treatment of those machines ought to parallel our treatment of similar humans.

[24] For that matter, we could likely repeat this argument when addressing the question of whether a machine can have a mind, since again such a machine will not share our evolutionary history and so forth.

[25] Think of this as the moral equivalent of the Turing Test: if the machine's behaviour is indistinguishable from a human's behaviour in most situations, then there is a prima facie case for treating it similarly. This argument is used by Peter Singer (2002) to argue for our assumptions of sentience both in other people and in animals. A similar line of thought has been developed by Rob Sparrow (2004, 2012) in trying to determine when we would view a machine as similar enough to a human to warrant the same moral standing.

[26] This concern has been echoed by Torrance (2012), although he seems more sympathetic to the dangers of mistakenly denying rights to machines which deserve them.

the results of our creations; moreover, there will almost certainly be someone who tries to design a self-aware autonomous machine simply because he can – because it would be interesting.[27] As such, it is overly optimistic to believe we can simply avoid the question in the manner she suggests.

Once we acknowledge that someone is likely to try to create such machines, or perhaps has even done so**,** we cannot ignore the question of appropriate moral standing. At least two pertinent objections must be acknowledged. First, there is the concern that by extending moral standing too widely with respect to machines we might unjustly limit the rights of the creators or purported owners of said machines: if, in fact, those machines are not autonomous or self-aware, then we have denied the property claims of their owners. Second, there is Bryson's concern that we may waste resources by extending rights to machines that are not autonomous. In each of these cases, however, I see the moral fault in being overly conservative is much larger than the risk of being overly generous.

The risk of losing a piece of property is trivial compared to denying moral standing to a being. However, it is much more difficult to dismiss the larger concern that, as a society, we may divert resources inappropriately; we have difficulty using our resources to aid the humans we know have moral standing, and the problem only magnifies when we consider the case of machines. It is certainly reasonable to recognize that our resources are limited and we may not be able to help all persons. Yet surely it is morally repugnant to allow someone to freeze to death because we had diverted our energy to power my toaster. While we may must sometimes make difficult choices in situations where we cannot help all people, even those who are ultimately unaided must be given ethical consideration; we cannot simply ignore them.[28] Yet, of course, there is great uncertainty in determining whether machines are moral persons; how then should we address the worry of diverting resources inappropriately?

I believe the best approach is a probabilistic one. We generally believe that other humans are sufficiently like us in various respects that we see as relevant to having moral status. While the problem of other minds raises the possibility that we are deceived, most of us regard it as relatively unlikely; we view the probability of error as too small to risk denying moral standing based on that possibility. While we cannot directly know what it is like to be another human, we use our best understanding to draw parallels with our own experiences; we then make decisions based on that understanding.

We do the same thing when considering the moral status of non-human animals, albeit with a higher probability of error. Hence we may examine the behaviours of chimpanzees, the brain activity in various animals, and so forth. We then compare this to our criteria for moral standing and ask how likely the entity is to satisfy those criteria; if an animal whimpers when you step on its paw, what is the probability that its whimper is a sign of pain? If an animal appears to exhibit emotions, does that make it psychologically like us? We assign moral statuses depending on our answers to these questions, and we reassign statuses when new data shows us that our previous beliefs were mistaken; this is why we have moved away from using chimpanzees in medical research, for instance. (National Institute of Health 2013)

---

[27] There are already many researchers involved in trying to create intelligent machines, for instance via The Mind Machine Project at MIT. Furthermore, there has been a great deal of discussion about what consciousness or self-awareness in a machine would entail. For a number of optimistic outlooks on the matter see Long and Kelley (2010), O'Regan (2012), Gorbenko et al. (2012).

[28] This is why presumably no matter what decision one makes in the trolley case, one is acting unethically if she fails to consider the humanity of all of the people involves. Simply ignoring the personhood of any of the individuals involved is not an ethical move, no matter how much simpler it would make the scenario.

Our application of moral criteria to beings other than ourselves always rests on a kind of estimation because it is not possible to have first-hand experience of others' situations. In the case of non-human animals, as well as in the case of other humans, we are able to find biological similarities to ourselves. However, we will clearly have some evidence in the machine case as well: the question will then be how likely we believe that its behaviours stem from consciousness and self-awareness as opposed to mere programming. Whether we acknowledge it as a moral patient will depend on our answer; if it seems highly likely, then we are more apt to divert resources to it. If it seems less likely, then we may only divert those resources if they are not needed for others.

We will undoubtedly be mistaken in our estimates at times. A failure to acknowledge the moral standing of machines does not imply that they actually lack moral standing; we are simply being unjust in such cases, as we have frequently been before. I am inclined to be generous about moral standing, however, because history suggests that humans naturally tend to underestimate the moral status of those who are different. We have seen women and children treated as property; even today many victims of human trafficking are still treated this way. Under the auspices of colonialism, entire existing civilizations of people of colour were dismissed as inferior to those of white Europeans. Animals remain a source of contention, despite the fact that they seem to suffer. I believe that we are already very sceptical about the status of others; as such, I am less worried that we will be overly generous to machines and more worried that we will completely ignore their standing. I see the risk of diverting resources inappropriately away from machines as far less likely than the risk of enslaving moral persons simply because they are physically unlike us.[29]

4. Moral Standing and Rights

4.1 Rights of Machines

The moral standing of intelligent autonomous machines is a natural extension of the sentience-based criteria for moral standing.[30] Intelligent, self-aware machines are beings which have interests and therefore have the capacity to be harmed. Hence, they have at a minimum moral claims to self-preservation and autonomy, subject to the usual limits necessary to guarantee the rights of other community members.

It is difficult to specify what moral entitlements said machines will have until we know the nature of those machines. For instance, Kevin Warwick (2012) discusses the possibility of conscious robots with biological brains. If those brains contain a sufficient number of human neurons, then they deserve the same kinds of protections we give to other beings with such neural complexity; we would be committing a moral wrong if we treated them as simply a thing in a laboratory. However, since machines (whether a biological hybrid or not) are physically rather different than humans, some rights will need to be "translated." A basic human right to sustenance will take a rather different form for machines, for instance, since they are unlikely to

[29] Some claim that this argument could be used to extend rights to a foetus. However, I think it clear that a foetus does not *at the time it is a foetus* act like me in a wide range of situations; we weigh the probability of its personhood as less than that of an adult human, although how much less will depend on the individual.

[30] It is probably possible also to defend granting moral standing to such machines on a rationality-based understanding of the moral community, however as I am sympathetic to the criticisms of such theories, I shall not attempt to do so here.

need food and water; they might well have an similar need for access to electricity, however. Similarly, just as humans have a need for medical care of various kinds, intelligent machines might require certain kinds of preventative maintenance or repairs.

Even such basic rights raise issues concerning what it means for these machines to exist or to cease to exist. In order to have a right to self-preservation, we must understand what that means with respect to these beings. At present, I can create a copy of a file which is functionally identical to the original. If I copy it on to two separate computers, we generally say that the same file is on each computer. What happens if this is possible to do with a virtual consciousness? Does the entity survive so long as at least one copy remains? If there are multiple copies, does that mean that there are now multiple copies of the same entity? Or are each separate entities with separate identities? Can we destroy an intelligent machine as long as we have copied all of its files onto another machine? What is life and death to a machine?[31]

Moving beyond basic needs for survival, consider rights on a larger socio-political scale, such as the basic human rights espoused in the United Nations' *Declaration of Human Rights* (1948). It is not immediately obvious how some of these will be handled, such as the claim that everyone has the right to a nationality. For humans, we determine that nationality based on the arbitrary criterion of birthplace (or parental nationality); it is then theoretically possible to change affiliation by undergoing certain processes.[32] One might suggest, therefore, that we could grant machines a starting nationality based on where they were first "switched on."

4.2 Rights of Virtual Entities

This answer is further complicated if we extend moral consideration from machines to entities which are not embodied and have only a virtual presence.[33] My argument could fairly easily be expanded to include these entities, since they could also display autonomy or self-awareness. The main adjustment needed is to devise an understanding of what their existence consists in, since it cannot be linked easily to embodiment. We do not have much experience with non-corporeal existence, hence there are metaphysical questions that would need to be addressed before we can determine how best to understand the rights of these beings.

For instance, the human sense of self is frequently tied to our physical embodiment; we see our bodies as part of who we are, which is why people who undergo procedures such as mastectomies often struggle to see themselves as the same person. (Piot-Ziegler et al. 2010) This strong connection to our bodies makes it hard for us to comprehend what sort of identity a disembodied being would have. Clearly such a being should be able to have an identity, however, since even after an amputation or a mastectomy a person retains some sense of self, even if somewhat altered. As such, a specific embodied form is not a requirement for identity and self-awareness. Similarly, the desires of many people not to be kept alive in persistent

[31] This touches on questions relevant to moral agency as well, since as people have noted (Asaro 2012), having legal responsibility would require us to be able to punish a machine which failed in its legal responsibilities; this requires us to know whether and how it is possible to do so.

[32] I say "theoretically" since, in practice, the change of nationality is fairly difficult; most people are pragmatically limited to the nationality of their birth, regardless of having a human right to change it.

[33] One could object that, speaking precisely, such entities will likely not be wholly virtual. Rather, they may well require the existence of physical objects in the same way that computer viruses require physical machines on which to reside; their existence is not independent of physical objects. However, the identity of the virus or the machine is quite distinct from the physical object(s) they depend on in a way unlike our experience of other identities; if they are embodied, it is in a very different sense than we currently understand.

vegetative states highlights the fact that for many the important component of identity is not the body. Together, this implies that a virtual entity could have an identity. Yet clearly this sort of entity will complicate questions such as nationality: how do you attach a nationality to something which does not have a physical presence per se? Is there any benefit to trying to do so? What would it mean if they existed outside the current borders of our political structures?

One possible avenue for investigation is to consider how we treat the moral status of other non-biological entities, such as corporations; they too have an existence which is not directly tied to a particular physical instantiation. A number of philosophers (Wallach and Allen 2009, Asaro 2012) have noted that we have granted legal rights and responsibilities to corporations, effectively treating them like persons. Furthermore, corporations can commit moral wrongs, such as outsourcing garment production to places which lack reasonable safety precautions for workers; corporations which do this are unethically placing profits ahead of human life. These kinds of issues particularly occur with multi-national corporations; we are struggling with how to apply the notions of rights and responsibilities to entities which are not tied to a single location and physical entity. While machines will differ from corporations in a variety of ways, the parallel highlights the fact that we have made this kind of extension of morality before; there may be no easy answers, but we are not left entirely without precedent.[34]

In addition to translating current human rights into forms which are more applicable to machines, it will likely be necessary to consider new problems which these machines generate. For instance, at the moment, it is not possible to replicate the contents of my mind. As such, my identity is fairly solidly unique and not in need of protection. However, as artificial brains become possible, we must ask whether a person has some kind of uniqueness rights; if we copy a virtual entity on to another machine without its permission, have we wronged it?[35] These questions will be necessary to consider as we move forward with artificial intelligence. Hence while it is clear that conscious, self-aware machines have moral standing, it is much more difficult to say exactly what that standing grants them; much depends on how our technologies evolve.

5. Conclusion

I have argued that the sentience criterion for moral standing is, in fact, insufficient to cover all humans; it does not explain what is morally wrong about one's action in the congenital analgesia case. Rather than seeing sentience as necessary for moral standing, therefore, I have suggested a move to an interest-based account: if a being has interests, then it is wrong to ignore those interests or to harm them in the absence of some suitable overriding reason. This view of moral patiency, however, may be extended to machines. If a machine has interests, then it may be harmed or benefitted; it deserves some moral consideration.

Furthermore, I have argued that there are several ways that a being may have interests in addition to being sentient. A self-aware conscious being will have interests; so will an autonomous intelligent machine. In general, if a being is capable of desiring its own continuance and of forming wishes about its future, then we have some prima facie obligation to respect those desires.[36] Determining the details of machines' moral standing is difficult, particularly since the

[34] There is, of course, debate about whether this is a good precedent to have set. The point remains, however, that we have dealt with non-human persons in the law before; it is not entirely new territory.
[35] This is similar to questions raised by cloning a person without permission.
[36] As with any other member of the moral community, those rights may be overridden if necessary.

relevant machines do not yet exist (or at least are not acknowledged to exist); some moral theorizing may need to wait until we have a better idea of what they are like. However, we have some precedent for thinking about the moral standing of non-biological entities by considering the moral status of corporations.

The battle for recognition of machines' moral standing will not be easy. We do not acknowledge the claims of others readily, even when the only difference between ourselves and those people is skin colour or gender; this difficulty will be magnified for intelligent machines. One key problem is the need for others to acknowledge the autonomy and/or consciousness of those machines. Philosophers have been arguing over the problem of other minds for millennia with respect to humans; the problem will likely magnify for machines, since we do not have a clear set of criteria that all will accept as sufficient for consciousness or autonomy.[37] The risk of attributing incorrect moral standing to machines is one which will likely plague discussions of machine ethics for some time to come.

Although I acknowledge that we make mistakes in our attribution, I believe that we are more likely to err on the side of conservatism than that of excess. Not only do we have a long history of doing so to other humans, given our past experiences with colonialism, but there will likely be intense financial pressure not to recognize machines as moral persons. We depend upon machines to do many tasks for us, and we do not currently pay machines or worry about their needs (beyond perhaps basic maintenance). One of the rights enshrined by the United Nations (1948) is the right to remuneration for work, meaning that the financial pressure to avoid recognizing any moral standing for intelligent machines will likely rival the push to avoid acknowledging African-Americans as full persons in the Confederate South. However, we cannot ethically deny someone moral standing simply because it is convenient.

The time to start thinking about these issues is now, before we are quite at the position of having such beings to contend with. If we do not face these questions as a society, we will likely perpetrate injustices on many who, in fact, deserve to be regarded as members of the moral community. I urge moral generosity when considering the moral claims of machines; we need to counter our legacy of sluggishness in recognizing as moral persons those who are physically unlike us.

References

Asaro, P. (2012) A Body to Kick, but Still No Soul to Damn. In P. Lin, K. Abney & G.A. Bekey (Eds.), *Robot Ethics: The Ethical and Social Implications of Robotics.* MIT P, Cambridge, USA.

Basl, J. (2012) Machines as Moral Patients We Shouldn't Care About (Yet): The Interests and Welfare of Current Machines. In D. J. Gunkel, J. J. Bryson, and S. Torrance (Eds.), *Proceedings of the AISB/IACAP World Congress 2012: The Machine Question: AI, Ethics and Moral Responsibility*. Birmingham, England.

Bentham, J. (1996) *An Introduction to the Principles of Morals and Legislation*. J.H. Burns and H.L.A. Hart (Eds.) Oxford UP, New York, USA.

Bringsjord, S. (2010) Meeting Floridi's Challenge to Artificial Intelligence from the Knowledge-Game Test for Self-Consciousness. *Metaphilosophy*, 41, 292-312.

[37] See Floridi's presentation of this conundrum (Floridi 2005) and an attempt to devise a test for self-consciousness in response (Bringsjord 2010).

Bryson, J. (2010) Robots Should be Slaves. In Y. Wilks (Ed.), *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues.* John Benjamins, USA.

Code, L. (1991) Is the Sex of the Knower Epistemologically Significant? In: *What Can She Know?: Feminist Theory and the Construction of Knowledge.* Cornell UP, Ithaca, USA, 1-26.

Floridi, L. (2005) Consciousness, Agents and the Knowledge Game. *Minds and Machines*, 15, 415-444.

Gorbenko, A., Popov, V., & Sheka, A. (2012) Robot Self Awareness: Exploration of Internal States. Applied Mathematical Sciences, 6: 675-688.

Gunkel, D. J. (2012) A Vindication of the Rights of Machines. In D. J. Gunkel, J. J. Bryson, and S. Torrance (Eds.), *Proceedings of the AISB/IACAP World Congress 2012: The Machine Question: AI, Ethics and Moral Responsibility*. Birmingham, England.

Kant, I. (1996) Groundwork of The Metaphysics of Morals. In Gregor, M. (Ed.), *Practical Philosophy*. Cambridge UP, Cambridge, U.K.

Legg, S. & Hutter, M. (2006a) A Collection of Definitions of Intelligence. In Goertzel, B. (Ed.), *Proc. 1st Annual artificial general intelligence workshop.*

Legg, S. & Hutter, M. (2006b) A Formal Measure of Machine Intelligence. In *Proc. Annual machine learning conference of Belgium and The Netherlands*. Ghent, Belgium.

Legg, S. & Hutter, M. (2007) Universal Intelligence: A Definition of Machine Intelligence. *Minds and Machines*, 17, 391-444.

Long, L. N. & Kelley, T.D. (2010) Review of Consciousness and the Possibility of Conscious Robots. Journal of Aerospace Computing, Information, and Communication, 7: 68-84.

Mill, J.S. (1993) *On Liberty and Utilitarianism*. Bantam, NY, USA.

Mills, C. (1999) *The Racial Contract*. Cornell UP, Ithaca, USA.

National Institute of Health. (2013) Council of Councils Working Group on the Use of Chimpanzees in NIH-Supported Research Report. http://dpcpsi.nih.gov/council/pdf/FNL_Report_WG_Chimpanzees.pdf. Accessed 6 March 2013.

O'Regan, J. K. (2012) How to Build a Robot that is Conscious and Feels. *Minds and Machines*, 22: 117-136.

Piot-Ziegler, C. et al. (2010) Mastectomy, body deconstruction, and impact on identity: A qualitative study. *British Journal of Health Psychology*, 15: 479–510.

Ruffo, M. (2012) The robot, a stranger to ethics. In D. J. Gunkel, J. J. Bryson, and S. Torrance (Eds.), *Proceedings of the AISB/IACAP World Congress 2012: The Machine Question: AI, Ethics and Moral Responsibility*. Birmingham, England.

Scruton, R. (2006) *Animal Rights and Wrongs.* Continuum, London, U.K.

Singer, P. (2002) *Animal Liberation*. Ecco, USA.

Sparrow, R. (2004) The Turing Triage Test. *Ethics and Information Technology*, 6, 203-213.

Sparrow, R. (2012) Can Machines Be People? In P. Lin, K. Abney & G.A. Bekey (Eds.), *Robot Ethics: The Ethical and Social Implications of Robotics.* MIT P, Cambridge, USA.

Taylor, A. (1996) Nasty, brutish, and short: The illiberal intuition that animals don't count. *The Journal of Value Inquiry*, 30: 265-277.

Torrance, S. (2012) The centrality of machine consciousness to machine ethics: Between realism and social-relationism. In D. J. Gunkel, J. J. Bryson, and S. Torrance (Eds.), *Proceedings of the AISB/IACAP World Congress 2012: The Machine Question: AI, Ethics and Moral Responsibility*. Birmingham, England.

United Nations. (1948) The Universal Declaration of Human Rights. http://www.un.org/en/documents/udhr/. Accessed 3 January 2013.
Wallach, W. & Allen, C. (2009) *Moral Machines: Teaching Robots Right from Wrong*. Oxford UP, Oxford, U.K.
Warren, M. A. (1973) On the Moral and Legal Status of Abortion. *Monist*, 57, 43-61
Warwick, K. (2012) Robots with Biological Brains. In P. Lin, K. Abney & G.A. Bekey (Eds.), *Robot Ethics: The Ethical and Social Implications of Robotics.* MIT P, Cambridge, USA.
Zack, N. (2002) *The Philosophy of Science and Race*. Routledge, New York, USA.

Beyond the skin bag: on the moral responsibility of extended agencies

by F. Allan Hanson

2009

Hanson, F. Allan "Beyond the Skin Bag: On the Moral Responsibility of Extended Agencies," Ethics and Information Technology 11:91-99, 2009

Published version: http://dx.doi.org/10.1007/s10676-009-9184-z

Beyond the Skin Bag: On the Moral Responsibility of Extended Agencies

Acknowledgements footnote: I am grateful to Richard Cole, Richard DeGeorge, Anthony Genova, Louise Hanson, Deborah Johnson, Rex Martin, and Evan Selinger for stimulating criticisms and suggestions as I prepared this essay.

F. Allan Hanson
University of Kansas
November 29, 2008

The growing prominence of computers in contemporary life, often seemingly with minds of their own, invites rethinking the question of moral responsibility. Eurotransplant is a highly computerized system that generates priority lists of recipients for organs on the basis of compatibility, age, waiting time, distance between donor and recipient, and balance among the several participating countries (Tufts 1996:1326). Ethical considerations are clearly involved, for Eurotransplant seeks to achieve "an optimal proportion between justice and efficiency—the medical ethical criteria," and it is generally thought that it realizes these objectives better than previous procedures that relied entirely on human evaluations (De Meester, Persijn, Claas and Frei 2000:333). Where does the moral responsibility for the priorities generated by the computerized system lie: with the human programmers and users alone, or also with the databases and computer hardware and software?

It does not seem controversial to say that the moral responsibility for an act lies with the subject that carried it out. Thus our question about Eurotransplant raises the further question, what is the subject? This question has no absolute, definitive answer because the subject is a social construct that varies cross-culturally and historically. For some time the subject has been understood in our society to be the human individual. From that perspective, moral responsibility involved in Eurotransplant, or anything else, is limited to the humans involved. "After all," it might be said, "if something goes wrong, we don't punish the computers." True

enough. But we very well may *blame* them, and blaming, no less than punishing, is generally taken to be a marker of moral responsibility. This gives a glimpse of the possibility that moral responsibility might include things beyond human individuals. But if there is to be such a thing, it must be grounded in a different concept of the subject. Recent social theory has proposed a way of thinking about the subject that is indeed conducive to a broader concept of moral responsibility. The purpose of this essay is to review that alternative view of the subject, to imagine what a concept of responsibility derived from it might look like, and to consider what advantages it might bring and the objections that are sure to be raised against it.[1]

Individualism and extended agency

The view of the subject as only the human individual is known as methodological individualism. This theory holds that subjects are human beings entirely contained in their "skin bags" (Clark 2003), that maintain their identity and integrity over time and in their various dealings with other individuals and things of all sorts. Computers, other machines, tools, and animals on this view are understood as no more than objects that people encounter and manipulate in the course of their actions (Flew 1995, Jones 2000). An alternative view that has recently come on the theoretical scene goes under names such as extended agency, actor-network theory, distributed cognition, and cyborg. It maintains that most actions are undertaken by subjects that extend beyond the human individual to include other human beings and any number or kind of nonhuman entities (Clark 2003; Hanson 2004, 2007; Hutchins 1995, Haraway 1991, Law 1999; Selinger and Engström 2007).

A student looks up a book in a library's online catalog. The methodological individualist insists that the subject in this case—the agent that carries out the task—is the student alone. Extended agency theory holds that the deed is accomplished by the combined entity consisting of

the student, the database of library holdings, and the automated hardware and software that put the two in contact. The basic reasoning behind this extension of agency beyond the individual is that if an action can be accomplished only with the collusion of a variety of human and nonhuman participants, then the subject or agency that carries out the action cannot be limited to the human component but must consist of all of them. The essential participation of nonhuman elements in high-tech applications such as Eurotransplant, robotic assembly lines or MRI procedures is especially clear. However, it is obvious that many activities—herding sheep with dogs, mowing a lawn, even driving a nail—require the participation of nonhuman beings or tools as well as people. Hence, although the growing importance of computers has been the most important factor in the recent development of extended agency theory, that theory applies to actions of all sorts.

What does this have to do with moral responsibility? For methodological individualists the issue is simple. Every action is ultimately attributable to human individuals, and whatever role computers, robots, dogs, lawnmowers, or hammers may play is ancillary. Obviously, if only human individuals act, then only human individuals can be responsible for those acts, and that's the end of it. This view may be called *moral individualism*, the ethical twin of methodological individualism. When methodological individualist assumptions are deeply rooted, the mind is closed to any alternative to moral individualism.

If, however, one is open to extended agency theory's redefinition of the nature of acting subjects, then one can at least entertain the idea that responsibility may apply to nonhuman as well as human beings. This view may be called *joint responsibility*. Given the traditional dominance of individualism, both methodological and moral, extended agency as a theory of action requires something of a paradigm shift in conventional ways of thinking. And the notion

of joint responsibility demands even more, for many people find it counterintuitive to attribute moral responsibility to objects, animals, and other nonhuman things. To entertain the plausibility of such a paradigm shift requires closer consideration of the evidence and how it might be construed. Because moral individualism and joint responsibility are corollaries of the competing moral individualist and extended agency theories of action, first we will examine the argument for the subject as extended agency.

It is no ineluctable fact of nature that the subject is and must be the human individual. The subject, as I have said, is a social construct, and a great deal of evidence from various times and places indicates that the acting agent is frequently understood to be something other than the human individual. One common notion is that the things people do are often not of their own making but result from external forces working through them. Possession by demons and words from the deceased spoken by spirit mediums are obvious examples. The Sioux Indian Black Elk, who had powers of healing and clairvoyance, said: " Of course it was not I who cured. It was the power from the outer world, and the visions and ceremonies had only made me like a hole through which the power could come to the two-leggeds. If I thought that I was doing it myself, the hole would close up and no power could come through" (Neihardt 1988:205-205).

Explanations of action and responsibility for it are also often shifted from individuals as wholes to some of their constituent parts. Contact-period New Zealand Maori often explained how their "dark intestines" would not allow them to rest until they had taken revenge for the death of a kinsman (Hanson and Hanson 1983:128-29). Abraham Lincoln (1839) recounted the anecdote of "a witty Irish soldier, who was always boasting of his bravery when no danger was near, but who invariably retreated without orders at the first charge of an engagement, being asked by his captain why he did so, replied: 'Captain, I have as brave a heart as Julius

Caesar ever had; but, somehow or other, whenever danger approaches, my cowardly legs will run away with it.'" Closer to home, a Holy Grail in today's biomedical science is to look within individuals to identify the genetic causes of behaviors and qualities such as attention deficit disorder, autism, intelligence, and many others.

Of greater interest here is consideration of agencies that include but extend beyond the individual. A widespread example is the corporate group. In many simpler societies this takes the form of a kinship group such as a lineage or clan. Moral responsibility lies with the group as a whole, making it appropriate to retaliate for grievances or injuries against any member of the offender's group. The famous feud between the Hatfields and the McCoys is a further case in point. More common in our society is the corporation, a legal individual that is held responsible for what it does to or for other corporations, clients, and employees.

The idea of the individual as subject is only a few centuries old. Erich Fromm, following Jacob Burckhardt (1954 [1860]:100-101), held that the individual was born in the Renaissance. "Medieval society," he wrote, "did not deprive the individual of his freedom, because the 'individual' did not yet exist….[Man] did not yet conceive of himself as an individual except through the medium of his social…role" (Fromm 1941:43). It is also a peculiarly Western idea (Ess 2006:222). Confucius held that ''persons are not perceived as superordinated individuals—as agents who stand independent of their actions—but are rather ongoing 'events' defined functionally by constitutive roles and relationships as they are performed within the context of their specific families and communities'' (Ames and Rosemont 1998:20). The same perspective is found in Zen Buddhism: "seeking after and grasping at a 'coherent self' that is non-existent from the outset only leads to a 'suffering.' The Buddhist idea of 'codependent arising' maintains that all things under the sun arise in a codependent relationship with each other. Nothing in the

world exists in complete independence and isolation from others. There is no such a thing as a solid basis that exists autonomously" (Nishigaki 2006:240). And finally, one prominent contemporary theorist has suggested that the conflation of the subject with the individual may be on the verge of disappearing, even in the West, "like a face, drawn in sand, at the edge of the sea" (Foucault 1970:387).

To think of the subject as something other than the human individual, then, is by no means unusual in human experience. They have widely been considered to be extended agencies consisting of pluralities of human and nonhuman beings and constituted by the contexts and events in which they participate. One can think of them as much as verbs as nouns, the doings of activities. It is a notion consistent with physicist David Bohm's view of the world as informed by relativity and quantum theory, in which everything is an unbroken flow of movement and supposedly durable things such as observer and observed are only momentarily stabilized forms of movement that form wholes for a time and then flow apart to join in new configurations (Bohm 1980:xi, 47).

To bring this into the practical realm, consider the well known phrase "guns don't kill people; people kill people." That is an absurd thing to say because it falsely implies that the possession of a gun is not pertinent to killing, or to the subject that possesses it. Of course guns don't go around shooting people all by themselves, but everybody knows that a person with a gun is a far greater threat to kill someone than a person without one. Selinger and Engström (2007:575-76), following Ihde (1990, 2002), explain that human beings are changed when they use certain technologies: a man-with-a-gun is a different being or subject than the same man without a gun. Similarly, it is commonly said that some people are transformed when they get behind the wheel of a car. Or consider the saying, "if all you have is a hammer, everything looks

like a nail." That is, the possibilities for action depend not just on human beings, but on the available means of action as determined by the relationship of humans with technology and/or other components of extended agencies. The entity that acts—the subject—is the extended agency (person/gun, person/hammer), not just the human individual.

The varieties of responsibility

If moral responsibility for an act lies with the subject that undertakes it, and if the subject includes nonhuman as well as human beings, then so may moral responsibility. Two versions of this point of view are distinguishable. One holds that at least some nonhuman entities may have moral responsibility in their own right. The other is that moral responsibility belongs to extended agencies as wholes rather than specifically to any human or nonhuman parts of them.

As for the first position, several authors who address the moral implications of automation ask whether automated agents such as computers and robots can be morally responsible (Coleman 2004, Friedman and Kahn 1992, Schick 1997, Hall 2000, Sparrow 2004). It is widely acknowledged that nonhumans participate in activities with moral import (e.g., Eurotransplant's computers) but the salient question here is whether they possess, all by themselves, the mental qualities generally accepted as necessary for moral responsibility. Prominent among these are intentionality, the capacity to act voluntarily, and awareness of the consequences of what they do. Dennet (1997) and those cognitive scientists represented in Dietrich (1994) are more willing than most to credit automated agents with mental qualities such as these, and are thus more open to the notion that they are, or eventually will be, morally responsible. But most thinkers are unwilling to go this far, and they resort to locutions about technological objects being quasi-responsible, conducive to moral behavior, implicated in it, and

the like (Johnson and Powers 2005, Johnson 2006, Floridi and Sanders 2004, Stahl 2006, Verbeek 2006).

Possibly future development of automated systems and new ways of thinking about responsibility will spawn plausible arguments for the moral responsibility of nonhuman agents. For the present, however, questions about the mental qualities of robots and computers make it unwise to go this far. Moreover, this perspective's focus on nonhuman agents in their own right actually shares the moral individualist tendency to separate them from humans. Extended agency and joint responsibility theory aims precisely to overcome that separation. Peter-Paul Verbeek anticipates that when it is overcome, "ethics can move beyond the fear that nonhuman objects will…suffocate human subjects and direct its attention to the moral quality of associations of subjects and objects" (Verbeek 2009:255). This enables recognition that "technologies play a fundamentally mediating role in human practices and experiences, and for this reason it can be argued that moral agency is distributed over both humans and technological artifacts" (Verbeek 2008:24, see also Verbeek 2009:257).

However, while it is relatively straightforward to attribute actions to extended agencies, other knotty issues come into play when it is claimed that they are also morally responsible. Most important among these are the notions that 1) moral responsibility is associated with deserts and 2) it requires awareness of consequences and freedom of action. These matters are conventionally restricted to human individuals. But a case for joint responsibility can dismiss the first of these as irrelevant, and, as for the second, can argue that the joint responsibility of extended agencies is not only plausible but that it accords even better with certain everyday notions than does moral individualism.

Deserts

A standard objection to joint responsibility by proponents of moral individualism is that, as animals, machines, and tools are neither punished nor rewarded for the activities in which they participate, they do not share moral responsibility for them. The first response to this objection that a proponent of joint responsibility might make is to point out that deserts are not a criterion for attributing responsibility because the question of deserts comes up only *after* a determination about responsibility has been made. Especially in cases potentially involving capital punishment, criminal trials are often divided into the determination of guilt phase and the penalty phase. The first phase is limited to the question of whether the defendant is guilty of (i.e., responsible for) the crime. Only after that question has been decided in the affirmative is the second phase of the trial convened to determine the punishment. The punishment (desert), that is to say, is a *consequence* of an already-made decision about responsibility, not a factor in making that decision.

The same pattern characterizes virtually all applications of deserts, be it in formal situations or in the give and take of daily life. We consider punishment or rewards only after we have decided that what was done was bad or good. In behavioral psychology, rewards and punishment are known as positive and negative reinforcements. They are applied after a person or experimental animal has behaved in a certain way, positive reinforcement if the aim is to encourage future repetition of that behavior and negative if the aim is to extinguish it. In all these cases, the decision about whether a certain form of behavior is good or bad, desirable or undesirable always comes before a decision to apply deserts of one sort or another.

It might still be maintained that even if deserts are after-the-fact responses to decisions about moral responsibility rather than criteria for making them, deserts are still unique markers of responsibility. That is, all applications of deserts occur when moral responsibility is involved,

and only then. The rejoinder to that would be, in the first place, that it is simply not true, and in the second, it is not evidence for moral individualism because deserts are applied not only to human individuals, but to other entities as well.

Deserts are often used in response to behaviors that have no moral implications at all, as when psychologists give treats or electric shocks to rats that push buttons they are being trained to push or avoid. They are also used in cases where the behavior itself may be morally pertinent but the subject performing it is not considered to be morally responsible. Children are routinely rewarded and punished for desirable and undesirable behavior before they are old enough to be responsible. Owners constantly praise or chastise their pets ("Good dog!" "Bad dog!"), give them treats to reward appropriate behavior, and sometimes subject them to corporal punishment such as spanking. In the Middle Ages the animal as well as the human participant could be hanged if found guilty of bestiality, and bells that were rung to summon crowds to an uprising were flogged or destroyed (Ihde 2006:273-74). People criticize tools, machines and other inanimate objects that are badly designed or made, blame them for poor performance, and praise things of all sorts that function well. Yet, with the possible exception of medieval notions about beasts and bells, none of these—be it a small child, an animal, or an inanimate object—is thought to be morally responsible for what it does.[2]

Moreover, the notion that the application of deserts supports moral individualism cannot stand because they are often applied to subjects other than human individuals. It is possible to view punishments such as cutting off the hand of a thief or castrating a rapist as directed specifically against a part of the body rather than the persons as a whole. In a similar vein, sharing much with Lincoln's "cowardly legs" anecdote recounted above, Jesus said "if your eye causes you to sin, pluck it out; it is better for you to enter the kingdom of God with one eye than

with two eyes to be thrown into hell." The sentiment is important enough to be repeated almost verbatim in two gospels: Mark 9:47, quoted here, and Matthew 18:9. In both of these places Jesus also expresses similar attitudes about one's hand or foot.

As for subjects that extend beyond the individual, it has already been noted that corporations are held to be morally responsible, legally as well as in public opinion, for their acts. They are praised and given citations, sued and fined or even (the corporate equivalent of capital punishment) dissolved for their misdeeds. It is the same with those informal extended agencies that are ubiquitous in everyday life, which form spontaneously to undertake particular activities and then dissolve when the task is finished and their components reform in other combinations for other activities. This is especially clear with certain kinds of negative sanctions that have the goal of breaking up an offending extended agency. Conviction of reckless driving or DUI often includes suspension of one's driver's license. When a child misbehaves with a toy by using it destructively or refusing to share it, a typical punishment is to take the toy away. The conventional view is that the child or criminal is being punished. But the problem emerges from the *conjunction of* child and toy, driver and vehicle. If either component were lacking— the child *or* the toy, the driver *or* the car—the offense would not have occurred and there would be nothing to punish. Thus it is reasonable to understand the punishment as directed against an offending relationship rather than a particular part of it. While we tend to think that the toy is being taken away from the child, for large toys with fixed positions such as a sandbox or a jungle gym, the child is removed from the toy. But what is really going on is that the offending toy-child relationship is being terminated.

The same principles apply, in mirror image, for positive sanctions. What are appreciated, praised, and encouraged are, at bottom, beneficial deeds and their results. These are often the

products of extended agencies rather than individuals acting alone. Even when individuals are honored they often acknowledge that the achievements were possible only because of the contribution of other people, education, and other circumstances. Frequently rewards are conferred explicitly upon extended agencies as wholes. This is obvious in the case of championship trophies and prizes awarded to debate and athletic teams, but there are many other examples. If a company decides to make a special investment in a given department or project, or a foundation makes a grant for an innovative educational program or for research, the award typically includes bonuses or salaries for the leading individuals, financial support for lower level personnel, and enhanced resources for office, classroom or laboratory space and equipment. The recipient of the reward is, in such a case, the department, school, or project team, i.e., the extended agency itself.

To summarize, 1) the issue of deserts comes up only after decisions about moral responsibility have already been made, 2) recipients of deserts may be considered to be either morally responsible or not morally responsible, and 3) those recipients may be animals, inanimate objects, and extended agencies as well as human individuals. For these reasons, joint responsibility theory holds that deserts are not decisive factors in assigning moral responsibility and certainly do not constitute evidence for moral individualism.

Freedom of action and awareness of consequences

Two people collude in a murder. One provokes the victim to chase him into the street and the other runs the victim down with a car. Neither conspirator could have accomplished this deed alone. Everyone would agree that they are both responsible for the crime. But it is also true that it could not have been done without the car. Does it share in the moral responsibility? Moral individualists would say no, because the car has no choice of action and no awareness of

the consequences of what it did. The reason for this is the presumption that only human beings can act of their own volition and know (or, to include negligence, should know) the consequences of what they are doing. The criteria or awareness and free choice are so strong that when they are not met, even adult humans are released from responsibility. People are not held morally responsible for acts they are forced to do, for consequences of their deeds that were completely unforeseeable, or if they are unable to evaluate the consequences of their acts (as with the McNaghten Rule that a defendant who was unable to distinguish between right and wrong is not guilty by reason of insanity).

Joint responsibility theory is in basic agreement with these propositions, and makes no claim that the car has moral responsibility by itself. It diverges from moral individualism in that it attributes responsibility to the extended agency as a whole, including both the human perpetrators and the car. [3]

Awareness of consequences and freedom of choice, which indeed are limited to the human conspirators, are necessary but not sufficient conditions for moral responsibility. It is also necessary for the deed actually to have been done. If the victim in our murder example had not been run down, there could be no responsibility for running him down. This was done by the extended agency that included the two humans and the car. Given that moral responsibility cannot exist but for the action of the extended agency, it lies with the extended agency as a whole and should not be limited to any part of it.

This argument can be run in a positive as well as a negative sense. Two people cooperate to rescue a toddler who had fallen into a well. The one descends into the well, a rope tied around her waist, to seize the toddler. The other, at the upper end of the rope, pulls his partner and the toddler out of the well. Everyone would agree that both human rescuers are responsible for the

good deed. But it is also true that the rescue could not have occurred without the rope. Absent the rope there would have been no rescue and therefore nothing to be responsible for. Hence the extended agency of rescuers-and-rope is responsible for the deed.

Obviously, the term "responsibility" takes on a different meaning here from its conventional usage. It is not, however, that the conventional sense of the word is being stretched to include extended agencies. The conventional sense of "responsibility" is defined by moral individualism, which is an artifact of the underlying theory of methodological individualism. Extended agency is an alternative theory to methodological individualism, and the concept of joint responsibility is equally an artifact of extended agency theory. This is the paradigm shift mentioned earlier. To entertain the possibility of joint responsibility entails recognizing extended agency as a distinct theory of action.

Joint responsibility can also be anticipatory or forward-looking (see Johnson and Powers 2005:100). A father hands the car keys to his sixteen year old daughter who has just gotten her driver's license. He admonishes her that, as the driver of a car, she is assuming new and greater responsibilities. In this usage, "responsibility" refers not to what has happened but to what *might* happen: she might hit a pedestrian or another car, cause injury to people riding in her car, get a speeding ticket, and so on. She is responsible to take proper precautions to avoid such events. But she can assume those particular responsibilities only when she gets behind the wheel of a car. Here again, if the action of an extended agency is necessary for the potential occurrence of an event or circumstance with moral import, then the anticipatory moral responsibility should be attributed to the extended agency as a whole rather than limited to the human part of it.

Intentions

Extended agency theory generates an argument for intentions that parallels the one for joint responsibility. I will not develop a detailed analysis here, but simply indicate its main lineaments. Intending, like forward-looking responsibility, is a before-the-fact concept that concerns what may happen rather than what has already happened. To intend involves willing, but it is limited to activities that have a reasonable probability of actually being attempted, as well as a reasonable chance of being achieved. It makes sense, for example, to say that one would like to jump across the Grand Canyon, or wishes one could do so, but no sense to say that one intends to do so. In Salman Rushdie's novel *Shalimar the Clown,* Shalimar is a highly accomplished Kashmiri tightrope walker. His dream is to become so adept that he can dispense with the rope entirely. He would like to do this, he wants to do it, and at one point in the novel he even does it, escaping from prison by running along the top of a wall and continuing to run through the air after the wall comes to an end. But the reader can countenance that episode only by suspending disbelief, for in the real world it is not possible for a human being to run through the air supported by nothing. Therefore, in the real world, Shalimar could not intend to do it because it cannot be done.

Just as moral responsibility requires awareness of consequences and freedom of action, intending requires will. Inanimate objects and animals can no more intend to do something than they can be aware of the consequences of what they do or are free to decide whether or not to do it. Most important for our purposes, however, is that intention also shares the requirement with forward-looking responsibility that the act in question could actually take place. If it is not possible, it cannot be intended. If it is an act that can be accomplished only by an extended agency that includes nonhuman as well as human components, such as walking on a tightrope, the intention to do it lies with the extended agency as a whole. It must include a human

component because that is the locus of the will, which is necessary to intentions. But it must also include the nonhuman components, because without them the act could not be done, and therefore it could not be intended.

Moral individualism and causal responsibility

If pressed, a proponent of moral individualism might accept the proposition that responsibility extends beyond the human individual, but still insist that human and nonhuman entities have different *kinds* of responsibility. The responsibility of the car in our murder example or the rope in our rescue example is purely causal (as, for example, high wind might be responsible for blowing down a tree) while that of the human perpetrators, who alone are aware of what they are doing and have the freedom to act otherwise, is both causal *and* moral.

It is possible to frame parallel arguments for intending, but I will limit the response to responsibility. First, assigning different kinds of responsibility to different components of an extended agency depends on the ability to distinguish unequivocally between what the various components do. Sometimes this is not difficult. In our rescue example, the exact roles played by each of the two human rescuers and the rope are easy to identify. But with extended agencies that include automated expert systems featuring decisions made by computers, differentiating tasks is much more difficult. As Johnson and Powers point out, "the distribution of tasks to computer systems integrates computer system behavior and human behavior in a way that makes it impossible to disaggregate in ascribing moral responsibility" (2005:106).

It is especially difficult to call upon initial human programming to explain the behavior of artificial intelligent systems that are the result of CAD-CAM (computer-assisted design, computer-assisted manufacture) and that can learn. After a time such systems may develop into something quite different from their beginnings, independently of any further human input.

Consider Tierra, a computer program designed by biologist Thomas Rey to simulate evolution by natural selection. He began with a self-replicating digital creature consisting of 80 instructions. Its progeny were designed to replicate themselves with fewer instructions, being selected for by using less CPU time in an environment where that is a scarce resource. Over the generations and with no further human intervention, the system produced new, more efficient creatures. Small parasitic creatures emerged that co-opted the features of larger ones. In response, some of the host creatures developed immunity to the first generation of parasites, following which new parasites capable of penetrating the hosts' defenses were replicated. Rey, the original programmer, could not predict the developments that were taking place (Turkle 1998:321). If any system with clear moral significance, such as Eurotransplant, incorporates machine learning, it would eventually become impossible to sort out how the various parts of the extended agency influence each other and what part of an action each of them does. At that point the only recourse is to assign responsibility to the extended agency as a whole.

It might be argued, however, that the difficulty in distinguishing what humans do from what computers do in cases such as this is purely a practical problem. As such, it has no bearing on the theoretical or in-principle issue of where moral responsibility lies. But practical issues often determine how various situations are regarded, and this is as true of individualism as it is of extended agency theory. It is commonly recognized that an individual's behavior is determined by multiple factors, including genetic make-up, childhood experience, formal training, and momentary impulse. It should be possible in principle to sort out the particular contribution of each factor in a given situation, but this is seldom done because the task would be extremely complex, time consuming, and highly controversial. Instead, moral individualists adopt the practical option of attributing responsibility to the person as an undifferentiated whole. Cases

where the contributions of humans and computers to tasks such as predicting the weather or allocating donor organs are difficult to disentangle are no different.

Second, joint responsibility actually accords better in certain ways with general understandings of responsibility than does moral individualism. Recall the father admonishing his teenage daughter about her new responsibility, now that she will be driving a car. It is possible that he articulated similar warnings several years earlier, when she learned to ride a bicycle. "Be careful not to run into people or things, don't crash your bike or hurt yourself, and especially don't ride into the street without looking." Her responsibility with the bicycle is, however, considerably less momentous than that with the car. Imagine now that she later becomes President of the United States and assumes the heaviest responsibility that any human being can have: to order a nuclear attack. Here her responsibility is infinitely greater than that associated with riding a bike or driving a car.

We appreciate the different degrees of responsibility in these examples immediately and intuitively, but they are more readily explained by joint responsibility than by moral individualism. If moral responsibility is restricted to human beings, then differences in responsibility must be explained in terms of human differences. But in this case the woman, who continues to ride a bicycle and drive a car after she has become President, is a constant. She is expected to exercise the most thoughtful prudence in all of these contexts. Hence moral individualism gives no satisfactory explanation for the obvious differences in her responsibilities. Using the concept of joint responsibility, the difference is explained by recognizing that there is not one subject in these scenarios, but three. More damage can be done by the extended agency woman-driving-a-car than woman-riding-a-bike, and a great deal more damage still can be done by woman-with-her-finger-on-the-nuclear-trigger. When we recognize the responsible parties in

these three cases as three quite different extended agencies, we can readily explain the differences in the moral responsibilities they carry.

Finally, the joint responsibility perspective encourages constructive, moral behavior in all contexts. Under moral individualism people are isolated in their skin bags, independent of other things. They of course have obligations to others, but the others remain, precisely, Other, ultimately alien from the Self. In contrast, extended agency theory emphasizes the multiple connections between humans and nonhumans of all descriptions in systems of action ranging in scope from the immediate all the way to the global. This is more consistent with recent emphases on ecological thinking. When the subject is perceived more as a verb than a noun—as a way of combining different entities in different ways to engage in various activities—the distinction between Self and Other loses both clarity and significance. When human individuals realize that they do not act alone but together with other people and things in extended agencies, they are more likely to appreciate the mutual dependency of all the participants for their common wellbeing. The notion of joint responsibility associated with this frame of mind is more conducive than moral individualism to constructive engagement with other people, with technology, and with the environment in general.

REFERENCES CITED

Ames, R. and H. Rosemont

1998 The Analects of Confucius: A Philosophical Translation. Ballantine Books, New York.

Bohm, David

1980 Wholeness and the Implicit Order. London: Routledge & Kegan Paul.

Burckhardt, Jacob

1954 [1860] The Civilization of the Renaissance in Italy. New York: Modern Library.

Clark, Andy

2003 Natural-Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence. New York: Oxford University Press.

Coleman, Kari Gwen

2004 Computing and Moral Responsibility. *In* Stanford Encyclopedia of Philosophy, Fall edition. E. N. Zalta, ed Stanford Encyclopedia of Philosophy.

De Meester, Johan, Guido G. Persijn, Frans H. J. Claas and Ulrich Frei

2000 "In the Queue for a Cadaver Donor Kidney Transplant: New Rules and Concepts in the Eurotransplant International Foundation." Nephrology Dialysis Transplantion 15: 333-338.

Dennett, D. C.

1997 When HAL Kills, Who's to Blame? Computer Ethics. *In* HAL's Legacy: 2001's Computer As Dream and Reality. D. G. Stork, ed . Cambridge, MA: MIT Press.

Dietrich, Eric, ed.

1994 Thinking Computers and Virtual Persons: Essays on the Intentionality of Machines. San Diego: Academic Press.

Ess, Charles

2006 Ethical Pluralism and Global Information Ethics. Ethics and Information Technology 8:215-226.

Flew, Anthony

1995 Thinking About Social Thinking. 2nd edition. Amherst, NY: Prometheus Books.

Floridi, Luciano, and J.W. Sanders

2004 On the Morality of Artificial Agents. Minds and Machines 14:349-379.

Foucault, Michel

1970 The Order of Things: An Archaeology of the Human Sciences. New York: Vintage.

Fromm, Erich

1941 Escape From Freedom. New York: Rinehart & Company.

Giere, Ronald N.

2008 Human Moral Responsibility Is Moral Responsibility Enough: Comment on F. Allan Hanson's "The Anachronism of Moral Individualism and the Responsibility of Extended Agency". Phenomenology and the Cognitive Sciences 7:425-427.

Hall, J. Stors

2000 Ethics for machines, Discuss.Foresight.Org/~Josh/Ethics.Html, August 16, 2004.

Hanson, F. Allan

2004 The New Superorganic. Current Anthropology 45:467-482.

2007 The Trouble with Culture: How Computers are Calming the Culture Wars. Albany: State University of New York Press.

2008 The Anachronism of Moral Individualism and the Responsibility of Extended Agency. Phenomenology and the Cognitive Sciences7:415-424.

Hanson, F. Allan and Louise Hanson

1983 Counterpoint in Maori Culture. London : Routledge & Kegan Paul.

Haraway, Donna J.

1991 Simians, Cyborgs, and Women: The Reinvention of Nature. London: Free Association Books.

Hutchins, Edwin

1995 Cognition in the Wild. Cambridge, MA: MIT Press.

Ihde, Don

1990 Technology and the Lifeworld. Bloomington: Indiana University Press.

2002 Bodies in Technology. Minneapolis: University of Minnesota Press.

2006 Forty Years in the Wilderness. *In* Postphenomenology: A Critical Companion to Ihde. E. Selinger, ed. Pp. 267-290. Albany: SUNY Press.

Johnson, Deborah G.

2006 Computer Systems: Moral Entities But Not Moral Agents. Ethics and Information Technology 8:195-204.

Johnson, Deborah G., and Thomas M. Powers

2005 Computer Systems and Responsibility: A Normative Look At Technological Complexity. Ethics and Information Technology 7:99-107.

Jones, Richard H.

2000 Reductionism: Analysis and the Fullness of Reality. Lewisburg, PA: Bucknell University Press.

Law, John

1999 After ANT: Complexity, Naming and Topology. *In* Actor Network Theory and After. J. Law, and J. Hassard, eds. Pp. 1-14. Oxford: Blackwell.

Lincoln, Abraham

1839 Speech on National Bank delivered to the Illinois House of Representatives, December 20, 1839. The Writings of Abraham Lincoln, vol. 1. http://www.classicreader.com/book/3237/

Neihardt, John

1988 Black Elk Speaks. Lincoln: University of Nebraska Press.

Nishigaki, Toru

2006 The Ethics in Japanese Information Society: Consideration on Francisco Varela's *The Embodied Mind* From the Perspective of Fundamental Informatics. Ethics and Information Technology 8:237-242.

Schick, Theodore, Jr.

Winter 1997 Can a Robot Have Moral Rights? Free Inquiry 18:42-44.

Selinger, Evan, and Timothy Engström

2007 On Naturally Embodied Cyborgs: Identities, Metaphors, and Models. Janus Head 9:553-584.

Sparrow, Robert

2004 The Turing Triage Test,. Ethics and Information Technology 6:203-213.

Stahl, Bernd Carsten

2006 Responsible Computers? A Case for Ascribing Quasi-Responsibility to Computers Independent of Personhood or Agency. Ethics and Information Technology 8:205-213.

Tufts Anette

1996 "Eurotransplant to Allocate Kidneys by Computer." Lancet 347(9011):1326

Turkle, Sherry

1998 "Cyborg Babies and Cy-Dough-Plasm: Ideas about Self and Life in the Culture of Simulation." *In* Cyborg Babies: From Techno-Sex to Techno-Tots. R. Davis-Floyd and J. Dumit, eds. Pp. 317-329. New York: Routledge.

Verbeek, Peter-Paul

2006 The Morality of Things: A Postphenomenological Inquiry. *In* Postphenomenology: A Critical Companion to Ihde. E. Selinger, ed. Pp. 117-128. Albany: SUNY Press.

2008 Obstetric Ultrasound and Technological Mediation of Morality: a Postphenomenological Analysis. Human Studies 31:11-26.

2009 Cultivating Humanity: Toward a Non-Humanist Ethics of Technology. In New Waves in Philosophy of Technology. J. K. Berg Olsen, E. Selinger, S. Riis, eds. Pp. 241-263. Hampshire: Palgrave Macmillan.

[1] This essay represents my second attempt to deal with these issues. For the first, see Hanson (2008) and, for a rejoinder, Giere (2008).

[2] It is important to bear in mind that the concept of joint responsibility under discussion here does not advocate the moral responsibility of nonhuman things in their own rights, but only the responsibility of extended agencies consisting of humans and nonhumans taken together.

[3] Strictly speaking the agencies involved in these events include many components in addition to those specified here, such as gasoline to fuel the car, the weather, the conditions of the road, and so on. There comes a point of diminishing returns, where the participation of some elements is sufficiently inconsequential (the make and year of the car) or so constant (the effect of gravity, the presence of oxygen) that they can be disregarded. As this is a discussion of general concepts rather than exact allocation of responsibility for particular events, it considers only the most obvious variables of any extended agency.

Chapter 20
The Functional Morality of Robots

Linda Johansson
Royal Institute of Technology, Sweden

ABSTRACT

It is often argued that a robot cannot be held morally responsible for its actions. The author suggests that one should use the same criteria for robots as for humans, regarding the ascription of moral responsibility. When deciding whether humans are moral agents one should look at their behaviour and listen to the reasons they give for their judgments in order to determine that they understood the situation properly. The author suggests that this should be done for robots as well. In this regard, if a robot passes a moral version of the Turing Test—a Moral Turing Test (MTT) we should hold the robot morally responsible for its actions. This is supported by the impossibility of deciding who actually has (semantic or only syntactic) understanding of a moral situation, and by two examples: the transferring of a human mind into a computer, and aliens who actually are robots.

INTRODUCTION

Technoethics focuses on the ethical aspects of technology in society, and attempts to devise principles to guide technological development in particular in relation to emerging new technologies that give rise to new ethical issues. Increasingly autonomous and intelligent robots represent one of these new technologies. Autonomy in robots raises questions about robot morality, and about how we can make sure that the autonomous robots behave ethically, in some sense (Bekey, 2005; Allen, Smit, & Wallach, 2006; Andersson, 2008; Andersson, Anderson, & Armen, 2004; Allen, Varner, & Zinser, 2000).

This is relevant for ongoing technological development. Today there are robotic research programs where internal "ethical governors" and

DOI: 10.4018/978-1-4666-1773-5.ch020

"guilt systems" are being developed, and there are discussions on potential "pull the trigger-autonomy" of unmanned aerial vehicles in war (Arkin, 2009).

It has sometimes been argued that a robot can never be held responsible for its actions. No matter how advanced it is it can never be autonomous; since it is always programmed by a human there is no question of a robot having alternative possibilities. Robots also seem to lack mental states, which are considered necessary in order to be an agent.

This paper argues that we do not need to know what goes on in a robot, in terms of being programmed or possessing mental states. If a robot can pass a so called Moral Turing Test, than we can hold it morally responsible for its actions.

An objection to the whole idea of trying to find criteria for whether a robot might be morally responsible, is the "what would be the point"-objection, that is, that it would be pointless to hold a machine responsible. We cannot punish a robot; it would be useless to send it to prison, for instance. If it misbehaves, we would simply turn it off or destroy it. But in a longer perspective, with robots becoming more advanced, this issue cannot be ignored. The potential responsibility of robots might have an impact on liability when something goes wrong. If the robot is considered responsible, we may not be able to punish it, but its responsibility may have implications for the responsibility of others, such as programmers. The ascription of responsibility to robots may also have influence on decisions whether and how to use such robots. There might also be implications for how robots should be programmed "ethically".

The idea of morality as a human *construction* in a moral community is a useful assumption when investigating matters regarding robot morality. Moral responsibility can be expected to be a central notion in the moral community since the whole point of morality is to promote right actions and prevent wrong actions. Members of the moral community might be moral agents or moral receivers, i.e., agents whose well-being is morally relevant but who cannot be held morally responsible.

The paper is outlined as follows. First, the moral community is described in terms of its members and how members decide whether other members are agents (which are a part of the other minds problem). The solution to the other minds problem seems—in this community—to be functionalistic; that is, we look at other humans' behaviour and assume that their mental states are what caused their outward behaviour. Then the moral community is discussed focusing on the case for robots. I argue that the nature of morality—the way humans actually behave in moral matters—supports the idea that *the passing of a moral Turing Test* (MTT) is a necessary and sufficient criterion for being held morally responsible. Support for this idea also comes from the so called "*deceiving robot-aliens"-example*. In summary I will conclude that the "functional morality" of robots allows us to hold them responsible. We have no reason to be biased towards nonorganic potential agents.

THE MORAL COMMUNITY

In order to decide whether we can hold *robots* morally responsible we should begin by considering when we hold *humans* morally responsible. On what criteria do we—or do we not— hold humans morally responsible?

Whom do we consider morally responsible? Not children, at least not very young children, and not animals, for instance. The same goes for the severely mentally ill, which is why we often test the mental health of defendants who are tried in court. Consider, for instance, a child hitting another child, not realizing that the other child feels pain—or someone who is mentally ill, and believed that he was fighting trolls and not innocent humans. The reason for not holding such people morally responsible is that we doubt their ability to properly understand what they do and what the consequences might be, or their ability

The Issue of Moral Consideration in Robot Ethics

Article *in* ACM SIGCAS Computers and Society · September 2015

DOI: 10.1145/2874239.2874278

1 author:

Anne Gerdes
University of Southern Denmark
25 PUBLICATIONS **32** CITATIONS

The Issue of Moral Consideration in Robot Ethics

Anne Gerdes
Associate Professor
Department of Communication and Design
University of Southern Denmark
4565501323
Gerdes@sdu.dk

ABSTRACT
This paper discusses whether we should grant moral consideration to robots. Contemporary approaches in support of doing so centers around a relational appearance based approach, which takes departure in the fact that we already by now enter into ethical demanding relations with (even simplistic) robots *as if* they had a mind of their own. Hence, it is assumed that moral status can be viewed as socially constructed and negotiated *within* relations. However, I argue that a relational turn risks turning the *as if* into *if* at the cost of losing sight of what matters in human-human relations. Therefore, I stick to a human centered framework and introduce a moral philosophical perspective, primarily based on Kant's *Tugendlehre* and his conception of duties as well as the Formula of Humanity, which also holds a relational perspective. This enables me to discuss preliminary arguments for moral considerations of robots.

Categories and Subject Descriptors
K4 [**Computers and Society**]: Ethics.

General Terms
Design,Theory.

Keywords
Moral consideration, ethics of robotics, duties, as if.

1. INTRODUCTION
In a recent report on lethal autonomous robot systems, Heynes points to that personhood is what links moral agency to responsibility [11]. But is that necessarily the case, or is Heynes being species chauvinistic? The answer could well be a yes, since robots have started to come into our social lives and we interact with them in human-like ways, as if they had inner mental states. On this background, it seems that we have good reasons to dwell upon our concepts of moral agency and patiency. Especially since our interactions with, and reactions towards, robots also concerns our self-image. First, I discuss the possibilities of artificial moral agency and patiency and explore whether this counts in favour of anchoring the question of moral status in phenomenological observations of how we form relations with robots; the so called *relational turn*, favoured by Coeckelbergh [3] and Gunkel [9], who summarizes the idea as an alternative to standard explanations, which sets out to decide, who (or what) deserves moral standing on the basis of ascribing properties to the entity in question. Hence, according to Gunkel, the relational *"..alternative [..] approaches moral status not as an essential property of things but as something that is socially negotiated and constructed in face of others."* ([10]:13)

I sympathize with the relational turn, but still find that it is challenged by the fact that, over time, our human-human relations may be obscured by human-robot relations. Currently, it may seem reasonable to skip discussions about what a robot *really* is and instead focus on how it appears to us and how we engage with it by applying *as if* approaches. But in the long run, our experiences with robots may radically alter our *Lebenswelt*. Here, I'm in alignment with the ideas of Turkle [18], who fears that we may lose something of great importance if we turn to robots or even end up preferring robots over humans.

For that reason, I outline a Kantian moral argument in emphasizing his treatment of duties in the doctrine of virtues, *The Tugendlehre*, which is presented in the second part of *The Metaphysics of Morals* [13]. Related to Kant's analysis of duties, there is room for a relational perspective, which can be expressed via the Formula of Humanity. Moreover, I also make reference to virtue ethical reflections in general. Thereby, I am able to put forward preliminary arguments for granting degrees of moral consideration to robots without risking that we gradually lose sight of our folk intuition and lived experience with what it is to enter into social relations. As such, I prefer to stay within a human centered framework, even though I agree with the proponents of the relational turn that there are baffling problems inherent to this kind of mind-morality perspective. However, the mere fact that things are complicated and problems unsolved does not constitute a proper reason for rejecting a framework.

2. ROBOTS IN THE MORAL SPHERE
The role of robots in moral discourse has been widely debated both within science fiction, philosophy and science. Hence, The World Robot Declaration was issued in Japan in 2004 and within the last decade, humans have increasingly interacted with care bots, pet bots, robot toys and robots for various therapeutic purposes (see for instance [18], [6], [1]).

ETHICOMP, September 7-9, 2015, Leicester, United Kingdom.

One of the first to include robots in the moral sphere was Asimov, who issued his famous laws of robotics, which he used in science fiction novels to illustrate ethical dilemma situations in human robot interaction. From an engineering point of view, in *Moral Machines – Teaching Robots Right from Wrong*, Wallach and Allen [21] present the promises of machine morality from an engineering perspective by distinguishing between top-down, bottom-up and hybrid approaches to programming morality. Here, the first mentioned system suggests the implementation of formalizations of a given moral philosophical theory, whereas a bottom-up system requires neural network models, which gradually build up moral understanding by trial and error based performance optimization techniques. However, pure bottom-up systems are challenged by the lack of a guiding ethical theory, and as such there is no guarantee that a robot will develop a preferred kind of moral maturity. On the other hand, a hybrid model, which Wallach and Allen speak in favour of, combines these ideas from a virtue ethical outlook: Here, artificial moral agency might be obtained by integrating bottom-up learning scaffolded by top-down rules.

By the same token, from a philosophical angle, Verbeek [20] grasps the possibility of artificial moral agency by viewing technologies as mediating devices, which serve as morally active in shaping human understanding and action in the world. Consequently, even though technological artifacts do not hold human-like intentions, it can make sense to refer to distributed or hybrid intentionality and hence assign intentionality to technology in the sense that technological artifacts may play a directing role in our actions and experiences ([20]:57). Correspondingly, in moving beyond an anthropocentric understanding of agency, Floridi and Sanders [8] reject free will and mental states as necessary conditions for moral agency. On the contrary, they argue that moral agency may be assigned to intelligent artificial agents (AAs) to the extent that such AAs are interactive, i.e., able to react to stimuli by changing state, and capable of adaptive behavior as well as autonomous responses to the environment. What matters is whether an agent can perform good or evil actions, that is, whether its actions are morally qualifiable ([8]:371).

If we include robots in the moral sphere by assigning moral agency and responsibility to them, a next reasonable step would be to discuss if the time has come where we ought to discuss whether robots are worthy of moral consideration? Among others, Gunkel thinks the answer to that question might be a yes. In *The Machine Question – Critical Perspectives on AI; Robots, and Ethics*, Gunkel [9] argues that already by now the term "person" has been stretched out to include non-human agents, such as corporations. As such, we might benefit from including machines into the category of persons. If we do so, the question arises whether the kind of responsibilities we have towards robots would be on pair with the kind of responsibilities we have towards animals, corporations or other human beings?

A lot has been written about machine agency in trying to lay out how robots ought to treat humans. Typically interest centers on how we may protect ourselves from possible harm caused by robots. At the same time little has been said about machine patiency. ([9]:103]). Hence, according to Gunkel, a claim to moral consideration, or even rights, may arise based on our social interactions with robots. We design artificial companions with whom (or which) we do engage and bond. Our machines are no longer tools, but have instead gradually turned into social actors or social interactive objects. Consequently, it may be about time we begin to think about moral obligations towards robots, maybe even in the strong form of robot rights. The mere fact that Paro, the seal care robot, is not a consciousness being with inner mental states does not automatically justify that we should not grant moral consideration to Paro. Moreover, our ways of living with robots is not just about what we do with robots, but also concerns our self-perception – what do I become through the kind of relations I form with robots?

A contrast to the relational view can be found in the work of Sparrow [17]. He presents a so-called Turing Triage Test which allows him to illustrate that we would always chose a human life over a robot's life, regardless of how advanced the robot might be. The mere fact that we can never know what the robot is *really* feeling, and if it feels anything at all makes it implausible to talk about, for instance, 'punishing' a robot: "*Our awareness of the reality of the inner lives of other people is a function of [..] "an attitude towards a soul"".* ([17]:211). According to Sparrow, there exist an unbridgeable gap between reality and appearance ([17]:210).

On the other hand, Coeckelbergh, like Gunkel, suggests a relational turn and continues by arguing in favour of replacing "*..the question about how "real" or how "moral" non-human agents are by the question about the moral significance of appearance."(*[5]:181).

He displays problems with what he coins "a property approach to moral status assignment", which seems to rest on the assumption that we can settle issues about moral significance with reference to a set of properties (e.g., mental states, speech, consciousness, intentionality). In this manner, we can supposedly establish a firm ground for separating out entities worthy of moral standing. But, Coeckelbergh points to problems inherent in this line of argument. Especially, it appears to be impossible to establish which properties we exactly need in order to be able to assign moral status to an entity. Also, the whole endeavor is challenged by "the other minds problem" - i.e.; the fact that we can never know for sure anything about the inner lives of others. Instead, Coeckelbergh focuses on our perceptions of robots and the way this affects our interactions with such entities:

"My suggestion is that we can permit ourselves to remain agnostic about what 'really' goes on 'in' there, and focus on the 'outer', the interaction, and in particular on how this interaction is co-shaped and co-constituted by how AAs [artificial agents] appear to us, humans ([5]: 188)

Coeckelbergh's phenomenological conception reflects a relational perspective, which takes departure in the observation of our mutual dependency. This fundamental precondition – with which everyone is actually familiar – forms a central point in Coeckelberg's so-called relational ontology, which assumes that "*relations are prior to the relata"*([3]:45), and thereby view robots and humans as "relational entities". For that reason, Coeckelbergh emphasizes a social-relational approach to moral consideration ([4]:219). But, here, unlike Coeckelbergh, I shall be arguing that we need not lean against appearance in combination with a social relational ontology. Instead, I point to a Kantian outset, which emphasizes how we can have duties *to* others and *with regard to* non-humans. Before moving forward, I find it important to stress that this paper does nothing else than provide a tentative outline of my preliminary ideas. In that respect, and all though I have reservations towards their positions, I find the work of Coeckelbergh and Gunkel highly inspiring and thought provoking.

3. AS IF

Appearance is closely related to the notion of 'as if', which is also explicitly noted by Coeckelbergh in mentioning that we interact with e.g., humanoid robots or artificial companions *as if* they could be trusted, blamed or loved. Therefore, Coeckelbergh calls for a phenomenological starting point in the investigation of human-robot relations, which takes departure in the "*observed or imagined"* human-robot relations ([5]:184).

It makes good sense to turn to analogical reasoning or to introduce *as if* constructions when confronted with unfamiliar territory. This kind of idealization, or way of using representations as tools, has been given a thoroughly treatment in Vaihinger's influential book *The Philosophy of as if* [19] in which he illustrates how fictions, i.e. *as if*-models and constructions may inform science and philosophy.

Fictions are applied due to their utility, meaning that they are justifiable when proving useful in practice. But, they are not on pair with hypotheses, which can be proved or verified ([19]: xlii). Obviously, there are shades of pragmatism in Vaihinger's work on the philosophy of *as if*. But we are not dealing with the pragmatic conception, which implies that what is useful to believe is true, since here "useful to believe" may involve *both* that which is true or false. In opposition to this, the guiding principle in Vaihinger's philosophy is the observation that fictions are not just false but contradictory. Hence, fictions are errors, but fruitful errors. Yet, Vaihinger warns us that the use of fictions may also lead us astray, hence in legal practice women used to be treated *as if* they wore minor, which caused grave injustice ([19]:148).

However, fictions are widely used in everyday thinking as well as in science, philosophy, economics, legal practice and in the description of abstract objects ect.. For instance, Vaihinger mentions Adam Smith's *Wealth of Nations*, which apply the fiction that human nature is driven by rational egoism. This fiction forms the foundation of Smith's theory. Likewise, Also, Kant, in his treatment of rational agency, requires us to act *as if* we were free even though this is not the case in the real, phenomenal world. By the same token, the categorical imperative demands that you "*act as if the principle of your action were, through your will, to become a general law of nature"* ([19]:292). Hence, according to Kant, our *vernunftbegriffe* are fictions since they do not refer to objects in the world of experience [14]:KrV B799). Actually, in explaining the role of *as* if, Vaihinger points to the fact that the term "heuristic fictions" was coined by Kant:

"Kant introduces a new term for what [..] he subsequently called "heuristic fictions": he calls the ideas "regulative principles of pure reason": they are not "constitutive" principles of reason, i.e. they do not give us the possibility of objective knowledge either within or outside the domain of experience, but serve "merely as rules" for understanding by indicating the path to be pursued within the domain of experience. By providing imaginary points on which it may direct its course but which can never be reached because it is outside reality." ([19]:273)

Also, Coeckelbergh notes that we can never have access to reality, mental states or the minds of others'. But, as noted above, instead of a mind morality approach, he suggests an alternative route. Rather than discussing the moral significance of either human or robot, we must turn to the study of appearance and relations in situations involving moral considerations in human-robot interactions ([4]:215). Consequently, when people, now or in a near future, start to treat humanoid robots as if they were moral agents, we could benefit from letting these observations guide our investigations by focusing on how humans experience and form interactions with robots through *as if* approaches.

Nevertheless, according to Vaihinger, fictions are only justifiable, not probable hypotheses. As such, I doubt that we need to take a full relational turn and introduce a social relational ontology. To me, it seems that the relational *as if* approach is challenged by the fact that, over time, our human-human relations may be obscured by human-robot interactions. Currently, it might seem reasonable to skip discussions about what robots *really* are and instead focus on how they appear to us and how we engage with robots in social situations by applying *as if* approaches and ascribe human-like agency to them. But in the long run, our experiences with robots may radically alter our *Lebenswelt* and by then we will no longer be able to make use of *as if* approaches, because we have forgotten what human-like relations are, that is: we have become unable to 'measure' experiences up against the benchmark of human relations. Here, I am in alignment with the ideas of Turkle [18], who fears that we may let go of fundamental values, such as trust and friendship, if we turn to robots or even end up preferring robots over humans:

"At the robotic moment, we have to be concerned that the simplification and reduction of relationships is no longer something we complain about. It may become what we expect, even desire."([18]:295).

Likewise, if philosophers take departure in observed and imagined human-robot relations, they risk turning the *as if* into *if* ([19], [7]:9) and thereby lose sight of what originally constituted human-human relations.

4. A HUMAN CENTERED PERSPECTIVE

In *Robot Futures* [16], Nourbakhsh describes a future scenario in which some kids act with great cruelty towards a robot dog. The scenario reminiscences about children's abusive behavior towards animals, and the son in Nourbakhsh's story remarks that: *"These people…they're sick. Let's go home!"* ([16]:54). By the same token, Nourbaksh reports a more recent experience with an autonomous tour-guide robot, which people would get great fun from teasing while it was guiding guests visiting a museum. Nobody seemed to care when it said: "please step out of my way", it was not until the engineering team changed the phrase to also include the people being guided by the robot, that people's attitudes towards the robot were changed to the better - *even slow robots will be treated well by people when they are wrapped into a human social context* ([16]:58).

As discussed above, a justification of moral consideration to robots may rest upon the observation that once we start ascribing agency to robots, we may possible become ethical obliged towards them. Moreover, the way we treat robots will have an impact on our moral habitus. In order to take this into account, I choose to I introduce Kant's distinction between two kinds of duties, as duties *to* human beings and duties *with regard* to non-human beings and entities [13].

Consequently, in what follows, I shall be introducing a perspective, which of course, within a relational ontology, is viewed as flawed due to problems derived from this kind of anthropocentric line and its inherent "property approach to moral status ascription" [3]. Both Coeckelbergh and Gunkel argue that we need to move beyond the assumptions of mind morality philosophers. They in particular point to the vagueness of metaphysical concepts and the fact that there is no consensus on

what these concepts designate. Moreover, complications also arise from the fact that we do not have access to others' minds. Hence, the argument goes that we must rethink moral agency and patiency by turning to their alternative relational paradigm ([9], [3]).

But, in contrast to their approach, I think that one cannot reject the role of metaphysical concepts, such as consciousness, intentionality and freedom, with reference to the fact that complicated issues have not yet been settled. This would be like discharging logic on the basis of Gödel's incompleteness theorems.

Hence, In *Facing up the problem of consciousness* [2], Chalmers notes that consciousness is the outmost puzzling problem in the science of mind ([2]:200). He has coined the terms *the easy problem* and *the hard problem* of consciousness in referring to the fact that we already know about the part of consciousness dealing with e.g., our ability to categorize, discriminate, associate and recognize patterns. Additionally, over time, our knowledge about brain processes will gradually increase, and we will probably end up knowing all there is to know about the complexity of the brain. This is *the easy* problem. But, *the hard problem* of consciousness is the problem of experience, that is, to learn why all that processing accompanies my consciousness experience. As such, mental qualia escape reduction to biophysical matters, and in modern dualism, property dualism holds that the mind has two fundamentally different types of properties, bio-physical and qualia. According to Chalmers, despite interesting and advanced cognitive science and reductionist models "*the mystery of consciousness will not be removed.*" ([2]:221). As an alternative, Chalmers sets out to outline a nonreductive theory of consciousness, which I'll not go further into here, where I only wish to point to Chalmers' observation that : "*The hard problem is a hard problem, but there is no reason to believe that it will remain permanently unsolved"* ([2]:218).

By itself, the observation that the concepts of mind pose baffling problems is no argument for dismissing the project of mind philosophy. I argue in favour of re-instantiating the mind-morality perspective, which allows me to move on to a Kantian and virtue ethical perspective, in which there is room for arguments for moral consideration of robots as different from humans, as well as from other artifacts or tools.

Moreover, Kant's Formula of Humanity reflects a relational perspective in describing how we ought to treat others (persons) as ends in themselves, where by "ends" Kant means "*only the concept of an end that is also a duty, a concept that belongs exclusively to ethics.[..]."* ([13]: 6:389). As such, we can only have duties *to* human beings, since duties require being capable of obligation ([13]:192). Meanwhile, Kant's *Tugendlehre* [13] allows for a description of moral obligations *with regard to* other beings or entities. Actually, Kant gives similar reasons as above in emphasizing that a prevalent argument for having indirect duties *with regard to* non-human entities and animals rest upon our duties *to* ourselves:

"§17 [..] a propensity to wanton destruction of what is beautiful in inanimate nature [..] is opposed to a human being's duty to himself; for it weakens and uproots that feeling in hum, which, though not of itself moral, is still a disposition of sensibility that greatly promotes morality or at least prepares the way for it[..]. With regard to the animate but non-rational part of creation, violent and cruel treatment of animals is far more intimately opposed to a human being's duty to himself, and he has a duty to refrain from this; for it dulls this shared feelings of their suffering and so weakens and gradually uproots a natural predisposition that is very serviceable to morality in one's relations with other men. [..] – Even gratitude for the long service of a horse or dog belongs indirectly to a human being's duty with regard to these animals; considered as a direct duty, however, it is always only a duty of the human being to himself."([13]: 6:443)

Thus, a Kantian perspective, as formulated in his doctrine of virtues, enables us to introduce degrees of moral consideration along a continuum stretching from, e.g. simple artifacts, such as tools, over to, for instance, paintings and historical buildings. We have varying degrees of duties *with regard to* such entities: One could say, that I have a duty towards tools, such as for instance my garden kit, in the sense that I handle these objects with care, i.e.; I clean them after use, oil them when needed and so on. In that sense, the practice surrounding gardening includes taking proper care of one's tools, and if I fail to do so, I will either feel bad about myself and improve my behavior or continue acting carelessly. In that case, others might blame me for neglecting my duties as a gardener. Here, we are of course dealing with moral consideration in a minimal sense thereof. But, from a virtue ethical perspective [15], the way I succeed or fail in my role as a gardener is nevertheless important for my personal flourishing.

Likewise, but on a more serious scale: when confronted with acts of vandalism, for instance the destroying of historical buildings by Islamic State, we find that such acts are wrongful due to the lack of moral consideration to these architectural pearls.

We do not have duties *to* animals, but we have duties *with regard to* animals. This is so, primarily because animals deserve moral consideration because they can suffer and because the way we treat animals will influence our self-perception. Moreover, according to MacIntyre:

"*To acknowledge that there are [..] animal preconditions for human rationality requires us to think of the relationship of human beings to members of other intelligent species in terms of a scale or a spectrum rather than of a single line of division between 'them' and 'us'"* ([15]:55)

Again, the question arises: what do I, or we, as a moral community, become if we abuse animals? This indirect argument for moral consideration has been criticized by Coeckelbergh [4]:213) with reference to that it seems contra-intuitive to justify moral consideration by referring to our own well-being rather than to the well-being of the receiver of moral consideration. But, as illustrated above, actually both Coeckelbergh and Gunkel stresses the importance of a relational turn (social relational ontology) with reference to that living with robots will change our lives, hence we need to reflect upon what we become from interacting with robots. By the relational turn Coeckelbergh de-individualizes the concept of a person and holds that we have to be viewed as *relational entities whose identity depends on their relations with other entities* ([4]:215).

In addition Coeckelbergh problematizes the fact that virtue ethics faces the problem of application. Hence, we cannot establish, or delimit, what the virtues are, which ought to guide our lives, and we cannot point out precisely which entities we should grant moral consideration by exercising virtuous behavior towards them. This is a classic line of argument against virtue ethics, which has been countered by Hursthorse [12] in arguing that an ethical normative theory does not necessarily have to deliver the right answers as such, or, in the case of virtue ethics, provide a

complete catalogue of virtues. As such, a plausible normative ethical theory should not give us universal rules to guide our behavior. Instead, it should be sufficiently flexible to allow for different moral outcomes by taking into consideration relevant elements in a particular context. Consequently, when faced with dilemma situations in real life contexts, it might well be the case that two persons solve a dilemma differently. This is not a relativist standpoint, since it does not imply disagreement about the fact that there is a conflict of values, rather it takes into consideration that, in the given context, there might be more than one solution, which is in accordance with that, which is virtuous.

Thus, from a virtue ethical perspective, we develop to become what MacIntyre calls *independent practical reasoners* [15]:158) through our upbringing and through participation in moral communities, which stand as morally robust and sound practices because they are open to critical reflective examination by members from in and outside the given community.

Within this kind of human based social framework, it might still be possible to grant moral consideration to robots by introducing a continuum on a scale above artifacts - such as tools and things, which we handle - over to animals. Probably below living entities, like animals, we may place robots with which we do form *as if* social relations.

I too hold that living with robots will change our lives. But I doubt that we need to take the relational turn.

5. CONCLUDING REMARKS

Since, we already by now interact with humanoid robots, and even rather simplistic types of robots, as if they were moral agents; we ought to start deliberating about moral status. This observation might lend support to a relational turn, which allows for viewing robots and humans as relational entities, rather than subjects and objects, thereby assuming that morality is always already situated in the social sphere and phenomenologically rooted in mutual dependency between social actors – "*relations are prior to the things related*" ([3]:110). Moreover, we ought to pay attention to how human-robot interactions actually unfold, that is, focus on *appearance* or how we apply *as if* approaches when we enter into human-like relations with robots. Thus, if we follow suit with the relational turn, we might benefit from not having to struggle with the problems of property ascription and mind-morality. Even better: Coeckelbergh holds that he does not want to give up on folk intuition reflected in the idea that there is a special relation between humanity and morality ([5]:181).

Yet, in the long run, our experiences with robots may radically alter our *Lebenswelt*. Therefore, by taking the relational turn, I think we risk losing sight of something of great value to our humanity, perhaps without recognizing that this has been the case. Instead, I suggest staying within a human-centered framework. Here, I present a Kantian relational perspective, which distinguishes between others, *to* whom we have duties, and non-humans, such as robots, with *regard to* which we have duties.

Even though I place myself in (humble) opposition to the work of Coeckelbergh and Gunkel, I am deeply inspired by them. Compared to their thoroughly analyses in the field of ethics of robotics, my contribution represents nothing more than a preliminary note. For now, I have no fully fleshed out solution to offer regarding how to establish a continuum, which enables us to grant various degrees of moral consideration to non-humans. Nevertheless, when speaking about robots, I still find it worth being anthropocentric for the reasons given above, but also bearing in mind that morality is deeply linked with mortality.

6. ACKNOWLEDGMENTS

I am grateful to my dear colleague, Klaus Robering, for inspiring discussions about moral philosophy as well as for his suggestions, which helped me develop this paper.

7. REFERENCES

[1] Bartneck, C., Van der Hoek, M., Mubin, O., Al Mahmud, A. 2007. Daisy, Daisy, Give Me Your Answer Do! Switching off a Robot. *Proceedings of the 2nd ACM/IEEE International Conference on Human-Robot Interaction*. Washington DC. . DOI: 10.1145/1228716.1228746. 217-222.

[2] Chalmers, D.J. 1995. Facing up the Problem of Consciousness. *Journal of Consciousness Studies* (2): 3, 200-219.

[3] Coeckelbergh, M. 2012. *Growing moral relations: critique of moral status ascription*. Palgrave Macmillan, NY.

[4] Coeckelbergh, M. 2010. Robot rights? Towards a social-relational justification of moral consideration. *Ehtics Inf Technol*.12, 209-221.

[5] Coeckelbergh, M. 2009. Virtual moral agency, virtual moral responsibility: on the moral significance of the appearance, perception, and performace of artificial agents. *AI & Society*. 24, 181-189.

[6] Dautenhahn, K. 2007. Socially Intelligent Robots: Dimensions of Human-Robot Interaction. *Philosophical Transactions: Biological Sciences*, Vol. 362, No. 1480, (Apr. 29, 2007). 679-704.

[7] Fine, A. 1993. Fictionalism. *Midwest studies in philosophy,* XVIII.1-18.

[8] Floridi, L., Sanders, J. W. 2004. On the morality of artificial agents. *Minds and Machines*. 14(3), 349-379.

[9] Gunkel, D. J. 2012. *The Machine Question – Critical Perspectives on AI, Robots, and Ethics*. The MIT Press. MA.

[10] Gunkel, D. J. 2014. The Other Question: The Isssue of Robot Rights. *Proceedings of Robo-Philosophy 2014. Sociable Robots and the Future of Social Relations.* Frontiers in Artificial Intelligence and Applications. IOS Press

[11] Heynes, C. 2013. Report of the Special Rapporteur on extrajudicial summary or arbitrary executions on Lethal Autonomous Robot Systems. A/HCR/23/47 fhttp://www.ohchr.org/Documents/HRBodies/HRCouncil/RegularSession/Session23/A-HRC-23-47_en.pdf.

[12] Hursthouse, R. 1999. *On Virtue Ethics*. Oxford University Press. Oxford. NY

[13] Kant, I. 1991. *The Metaphysics of Morals*, transl. by M. J. Gregor. Cambridge University Press.

[14] Kant, I. 1785. Akademieausgabe, vol. IV *Grundlegung zur Metaphysik der Sitten*. http://www.korpora.org/Kant/aa04/Inhalt4.html

[15] MacIntyre, A. 1999. *Dependent rational animals: Why human beings need the virtues.* Carus Publ. Company. Chicago.

[16] Nourbakhsh, I. R. 2013. *Robot Futures*. MIT. Cambridge. MA.

[17] Sparrow, R. 2004. The Turing Triage Test. *Ethics and Information Technology*. 6, 203-213. DOI: 10.1007/s10676-004-6491-2.

[18] Turkle, S. 2011. *Alone Together – Why We Expect More From Technology and Less From Each Other*. Basic Books, NY.

[19] Vaihinger. H. 1924. *The Philosophy of as if*. Transl. by C. K. Ogden. London.

[20] Verbeek, P. P. 2011. *Moralizing Technology: Understanding and Designing the Morality of Things*. The University of Chicago Press.

[21] Wallach, W., Allen, C. 2009. *Moral Machines – Teaching Robots Right from Wrong*. New York: Oxford University Press.

.