Chapter

A GENERAL THEORY OF MORAL AGENCY GROUNDING COMPUTATIONAL IMPLEMENTATIONS: THE ACTWITH MODEL

Jeffrey White*
KAIST, South Korea

ABSTRACT

The ultimate goal of research into computational intelligence is the construction of a fully embodied and fully autonomous artificial agent. This ultimate artificial agent must not only be able to act, but it must be able to act morally.

In order to realize this goal, a number of challenges must be met, and a number of questions must be answered, the upshot being that, in doing so, the form of agency to which we must aim in developing artificial agents comes into focus.

This chapter explores these issues, and from its results details a novel approach to meeting the given conditions in a simple architecture of information processing.

* E-mail: kaistphilosophy@gmx.com

1. WHY BUILD ARTIFICIAL MORAL AGENTS?

> And every assistant is as it were a tool that serves for several tools; for if every tool could perform its own work when ordered, or by seeing what to do in advance, like the statues of Daedalus in the story, or the tripods of Hephaestus which the poet says 'enter self-moved the company divine,' - if thus shuttles wove and quills played harps of themselves, master-craftsmen would have no need of assistants and masters no need of slaves.
>
> Aristotle[1]

Any realization of an artificial moral agent (AMA) satisfying our ultimate goal must meet certain conditions stemming from questions about morality and moral agency generally, beginning with "Why build artificial moral agents (AMAs) at all?" From these conditions, moral agency must be understood in terms applicable to computational implementation. The question here is "What is moral agency?" And, further, "What is the 'cash value' of moral agency?" From these answers should follow a model for implementation. Then, given a successful implementation, the claim to genuine moral agency must be supported. The question here is "What is a moral agent?" This chapter proceeds along this line of inquiry, concluding with some reflections on the promise, and ultimate purpose, of research into artificial moral agents.

There are two sources for initial conditions on the satisfactory form of artificial moral agents. From these conditions, we may glean the form and function that AMAs satisfying the ultimate goal of research into computational intelligence must embody and execute. These sources are the pre-theoretic questions that motivate the inquiry, and formal theoretical stipulations about the nature of moral agency. We shall review these in turn before articulating a novel architecture designed to meet these conditions, as what can be taken from them will provide a loose checklist against which the present, as well as any future, proposal might be measured.

The successful realization of any project meets the expectations motivating its original pursuit. This is how success in any given endeavor is finally measured. Should one set out to bake a cake, and end up with a poisonous omelette, the result is a bad situation. The reason for this evaluation has to do with one's aims in baking a cake – to partake in something sweet, for dessert. A good situation, that is, unless the cake tastes like mushrooms and will kill you. A similar measure, with equally dramatic potential results, may be applied to the question, "Why attempt to build an artificial moral agent?"

[1] *Politics,* Book 1, section 1253b.

This question can be answered variously. In the strongest form, Erich Dietrich has argued that AMAs not only should be built, but that they *must* be built, in order to replace human beings as moral inheritors of the Earth, superior to their creators by design [1] [2]. The assertion here is that, for all of the good that may be attributed to human beings, the bad outweighs it. Through the development of AMAs, we should create our own replacements and "exit, stage left." Dietrich's proposal is that we endeavor to construct a "...race of machines, *Homo sapiens 2.0*, which implement only what is good about humanity..." and suggests that "One way of carrying out this project would be to implement in the robots our best moral theories."[1][2]

Dietrich's proposal raises some challenges. For one thing, the superiority of the creation depends on the capacity of the creator to articulate this superiority. However, to date, there has been no adequate, computationally friendly articulation of "our best moral theories." Until we can detail to what moral superiority adds up, Dietrich's proposal is premature. After all, how are we to create a morally superior artificial agent, when we cannot even spell out what it is to be morally superior? Delivering such account is the focus of this chapter.

Premature though it may be, Dietrich's proposal is successful in one way. It forces us to confront the possibility that our presumed moral superiority over other creatures, natural and unnatural, requires constant support in the form of demonstrated right action, self-improvement, and even theoretical self-explanation, and that this support by any conventional measure is dwindling. So, maybe Dietrich is right, and humans *are* in need of replacing. The end to which this points does appear gloomy, but the proposed replacement of morally inferior critters (us, now) with morally superior critters (in whatever form) highlights the special promise of AMA research, and also opens an avenue for opportunities beyond self-annihilation. Indeed, not only might the necessary proof of human moral primacy lie in the successful construction of AMAs, but the people that emerge from the process may be affected in the right ways to stand as their own morally superior replacements. This is a very compelling reason to push through social, political, and technological hurdles involved in creating AMAs. Finally, it touches on the issue of the moral status of AMAs, a theme to be addressed as this chapter closes.

Suggesting that robots should be designed to replace humankind is a controversial proposal to say the least. Most considerations of the social status of AMAs presume their place either beneath or alongside human beings. And,

[2] Page 61.

this presumption brings with it certain prescriptions for the eventual form of any successfully realized AMA. Even if their purpose is the replacement of human beings, in order to cooperate with human beings, they must act according to the human standards belonging to their designers. They must be more alike than unlike us - "superior," after all, does not imply "alien." Thus, for its shocking initial impression, Dietrich's proposal is merely that we replace ourselves with beings much like ourselves, only better. There is nothing shocking about that.

How is this goal any different than that of a normal mother trying to raise her offspring, or for that matter any teacher trying to educate her students, to live a better life, in better ways, avoiding parentally demonstrated mistakes along the way? Contrary to Dietrich's insistence on artificial successors, it is not so clear that human replacement by non-humans is the only way, or even the most efficient way, to realize his goal.[3] Indeed, this has been the goal of moral education since the dawn of philosophy. And, it is to this end, moral education, that the demanding inquiry into the adequate realization of an AMA seems an especially fruitful enterprise. Through this inquiry, we may well succeed in replacing ourselves with our moral superiors, and our replacements may simply be our own selves improved through the increased self-understanding that comes from trying to teach AMAs how to be moral.

With less urgency than to save the world from human evil, AMAs should be built because we can learn about ourselves, our actions and the theories developed and recruited in understanding them, in the process. On this point, in the words of James Gips:

> The hope is that as we try to implement ethical systems on the computer we will learn much more about the knowledge and assumptions built into the ethical theories themselves. That as we build the artificial ethical reasoning systems we will learn how to behave more ethically ourselves. [4]4

While the process by way of which these insights are to arise is typically represented as uni-directional – human moral theories and lessons are imported into artificial implementations thereof, and are therein tested and refined – AMAs also promise to inform us about ourselves as they take on more complex roles in society. In the social integration of AMAs with human moral agents, theoretical questions of moral status, moral responsibility, the

[3] And, indeed, the line becomes even fuzzier when we consider the possibility of organic/inorganic hybrids as discussed in [3].

[4] Page 10.

roles of emotions in grounding moral convictions, the potential for emotional attachment – love? - and genuine moral obligation to AMAs, all arise with degrees of urgency proportional to the horizon of their expected realization. As this process continues, shortcomings within traditional moral theories to deal with relevant issues – both due solely to the theories themselves, as well as to mistaken interpretations thereof – promise to be revealed. So, both in terms of theoretical application, and pure scholarship, "... to consider how to construct an explicit ethical robot is an exercise worth doing for it forces us to become clearer about what ethical theories are best and most useful."[5][5] In ongoing inquiry into the broad field of morality and ethics, the design and construction of AMAs carries special promise. Thus, we are met with further compelling answers to the question "Why build AMAs?"

In this dimension, the question "Why build AMAs?" is not unlike the question "Why *build* anything at all?" The motivation implicit in any and all technological application is freedom from drudgery and labor. Mankind seeks through technology to be able to do more for less, doing what otherwise cannot be done. And in technology, we have so far enjoyed some modest success. The story of human progress, history itself, is largely marked out and measured by advances in labor saving devices. Progress is often measured in terms of standards of living, and these are often measures of the degrees to which labor saving devices - and the freedom from drudgery that they represent - are available to human beings. Equally, justice is a standard of the good life, and justice can likewise be measured by the equality with which technological advances are distributed and accessible. These answers bring to the fore themes to which we shall attend in some detail in the sixth section of this chapter.

In the end, answers to the question "Why build AMAs?" are not unlike answers to the question "Why ask 'why' about anything at all?" Recall Aristotle's famous reduction that all philosophic inquiry begins with the question "Why?" And, in this light rethink the question "Why build AMAs?" The pursuit of AMAs opens the door to the ultimate goal of philosophical inquiry, the good life in a just world, in this case perhaps better given as a better world through technology. Socrates himself on many occasions pridefully recalled a family relationship with Daedelus, mythical builder of 'moving statues,' in order to motivate inquiry into the life worth living. According to legend, Daedalus created statues imbued with a mechanism through which they could move, execute appropriate actions, and indeed *be* as

[5] Page 17.

if alive.[6] Even now, in the modern form of artificial intelligence, artificial moral agents, and artificial life, such creations not only serve as starting points of wonder, motivating theoretical inquiry, but they represent the promise unique to technology since the wheel. Our freedom as human beings to determine, through our own agency, the shape of the world in which we will live, and indeed, both the sorts of creatures with which we will share it, and the sorts of creatures that we, ourselves, will become in the process.

As these parameters reveal, conventionally held markers of human progress are intimately tied to technological achievement, on both practical and theoretical fronts, and this in itself provides important answers to the question "Why build AMAs?" while at the same time setting out conditions for their successful realization. They must make life better, more just. They must satisfy the sense of wonder, the search for knowledge, that spawned the inquiry into their realization in the first place. And, in clarifying these conditions, we have set out the checkpoints through which the present chapter must travel.

Finally, there is an answer to the question that is perhaps the most compelling given our present context. As for the question "Why build artificial moral agents in the first place?" Laszlo Versenyi recruits a Kantian argument in support of a positive assessment of the enterprise. On his assay, it is in the fulfillment of a moral duty to ourselves that these machines *must* be produced:

> If we have the talent to construct moral robots then not to do so would be to neglect one of our natural gifts and this cannot be willed as a universal law of nature, for every rational being "necessarily wills that all his powers should be developed, since they serve him, and are given to him, for all sorts of possible ends." Not to construct robots would mean to neglect humanity in our own person, and this would conflict with the categorical imperative of treating humanity in our person, as well as in others, as an end rather than as a means.[6][7]

To not build AMAs would be to violate universal moral law, binding on ourselves in virtue of our very natures as rational, morally self-legislating and therefore autonomous creatures. To not build AMAs would amount to doing the wrong thing, to acting immorally! And, it is surprisingly contradictory to presume that demonstrably immoral creatures should endeavor to construct any demonstrably moral creature.

---

[6] See Euthyphro 11 and 15, Meno 97. Also, see Aristotle's Politics.
[7] Page 256.

Such a quest would signify only the irrationality of the agent who sets out on it. Should we succeed at all, morally will be the only possible way, and this yields a last condition on the eventual form of any adequately realized AMA. It must come at the limit of our capacities. It must represent the best of us. For if it does not, then we will have succeeded only in doing ourselves a final injustice.

2. WHAT IS A MORAL AGENT?

Before we can pretend to instill moral principles into other beings - or even recognize them for that matter - it would serve us to first successfully realize them in ourselves. How are we to conceive of a moral agent, at all, otherwise? Just what do we think that a moral agent is?

James Moor distinguishes between four types of moral agency.[8] These are, in order of autonomy, the ethical impact agent, the implicit ethical/unethical agent, the explicit ethical agent, and the fully ethical agent.

Any agent whose actions have ethical consequences qualifies as an ethical impact agent. Implicit agents are essentially programmed to perform some action or particular set of actions which have ethical consequences. They do not create the principles that drive action plans for themselves[9], and rather "...have ethical considerations built into their designs."[5] Moor's examples of such are airplane instruments that warn pilots of unsafe conditions (ethical/moral), and 'spam-bots,' computers whose over-arching function is to serve as a hub for the distribution of unwanted, unwarranted, and even destructive electronic messages (unethical/immoral). Clearly, by either example, he is describing a level of currently realized ethical "agency."

The third type of ethical agent, the explicit ethical agent, begins to include some aspects of agency that are typically only associated with human moral behavior:

> Explicit ethical agents are agents that can identify and process ethical information about a variety of situations and make sensitive determinations

[8] By ethical agent, we can understand what we are calling here a "moral" agent. Though "ethics" and "morality" are often used interchangeably, there are distinctions to be made. In the present context, ethical agency and moral agency will be taken to attach to the same thing in the same ways, as any adequate AMA must satisfy either formula, although there is potential for further fruitful inquiry on this front.

[9] And so are not by the Kantian formula strictly autonomous (and by the Kantian formula, thus, strictly moral agents at all). For insight into what autonomy, and especially moral autonomy really adds up to, see [7].

> about what should be done in those situations. When principles conflict, they can work out resolutions that fit the facts. These are the kind of agents that can be thought of as acting from ethics, not merely according to ethics. [5][10]

This is the form of agency that Moor feels designers of artificial agents should strive to realize. He takes them to be the "paradigm case" of robot ethics, "philosophically interesting," "practically important," yet not so sophisticated as to never exist.

The fourth type of ethical agent is the full-blown ethical agency attributed to properly-functioning human beings. This is, according to Moor, not a level of agency which a robot is likely to achieve, characterized as it is by three characteristics assumed to belong uniquely to human agency, the attribution of which to (not to mention the realization within) robots remains - for some - a serious conceptual challenge and source of no mean controversy. These are: free will, consciousness, and intentionality. We will evaluate these concepts in this chapter with a focus on intentionality.

In order to measure the degree of moral agency exhibited by an AMA, Moor proposes a standard. It involves the empirical analysis of robotic behavior, employing a rational-discursive method of evaluating human moral agency in approximation of a form prevalent in contemporary philosophical literature:

> If a robot could give persuasive justifications for ethical decisions that were comparable to or better than that of good human ethical decision makers, then the robot's competence could be inductively established for a given area of ethical decision making. [5][11]

This rational-discursive standard for moral agency is already employed in theories on human moral agency under a different heading, sometimes represented under the moniker of "reason-responsiveness." The gist is this. Should any agent be able to deliver sufficient reasons for actions, then ethical competence in the field of action in question should be attributed to the agent. Likewise, a moral agent should change its behavior when given sufficient

[10] Pages 14-15.

[11] Page 15. This form of evaluation (modified to allow for modes of expression other than natural language – though in the end this takes nothing from the following objection) also shows up in the literature [8] in the form of a "Moral Turing test." The Turing Test itself requires no lengthy description. Moreover, its shortcomings are also well-known from Searle's Chinese Room to Harnad's symbol grounding problem. Rational-discursive and "moral" approximations thereof suffer similar difficulties. For starters, see [9].

reasons to do so by other members of its moral community. Such a model of reason-responsiveness can be initially presumed adequate for inclusion into a moral community, and so the moral standing of an agent capable of such self-determination in light of fluid social input (while maintaining adequate weights on certain moral principles) might be guaranteed.

Just how this is is to be worked out in practice remains a question, though I feel that it is one that can be answered once an adequate model of moral cognition is set out. We will look at this more closely, the big question being a moral equivalent of the famous 'symbol grounding problem' - just where are these reasons supposed to be coming from? And, this points to one problem with Moor's approach, one missing piece of the classificatory puzzle. This is that any evaluation of artificial agency according to any set of criteria, however speculative, which takes human moral agency as a standard requires first the articulation of human moral agency in terms appropriate to the computational implementations under evaluation.[12] After all, without this, who is to say that the third class of ethical agents – or even the second - does not contain at least some portion of humanity within it?[13]

The most robust resolution lies in the articulation of human morality in terms friendly enough to computational implementation that their relative evaluation proceeds from a common baseline, as aspects of a single comprehensive theory of moral agency, rather than proceeding by the standards of one imported into another. Such a comprehensive understanding of morality, in terms universal to both natural and artificial agents, provides common grounds for the attribution of moral status, as well. We will return to this issue of moral status as this chapter closes. Now, it serves to introduce the issue of intentionality, a theme important to the discussion of the ACTWith model at the heart of the present work, and a concept through which a universal baseline of morality can be derived.

Moral status can be traced to grounds both within and outside of any agent in question. Analyses of ethical agency so far have largely proceeded according to internal factors. Is it conscious, sentient, configured in relevant ways? However, the moral status of agents is perhaps more often practically determined externally, notably through the personification of those agents by the human beings who are affected by them. People attribute morally loaded properties to agents, such as intentionality, thereby attributing moral status

[12] The "Moral Turing Test" [8] takes some steps in this direction, relaxing the requirement for discursive justification to include prompted and appropriate action responses of any sort within the capacities of the agent in question, but actually this is beside the point.

[13] For a form of a solution to the problem raised here, see [10].

regardless of their internal properties, and regardless of their level of sophistication.

Specifically, people take an "intentional stance" toward things, and in so doing grant objects a certain moral standing, as if an object in question were a genuine partner in human life. An extreme example of such attribution resides in the fad phenomena of the pet rock. At one time in recent history, people took rocks to be unique partners, pets, and so vehicles of emotional attachment. Perhaps this fad was buoyed by the essential reliability of mineral behavior. Rocks can be depended on, after all. And, as this is a characteristic that people look for in friends, the mineral was felt to represent the relevant intentions required of a good friend.

As Moor points out, this attitude extends throughout the technological world. For example, the common, untutored understanding of the behaviors of computers is always and already packaged in terms of human agency. People presume that "A word processing program corrects our misspellings because it believes we should use different spellings and its goal is to correct our spelling errors."[5][14]

Not that people would actually hold that their home computers have desires and goals when pressed on the issue, but that they pre-critically understand them in these terms. That is, they understand computer agency in the terms with which they always and already understand in their own moral and ethical lives.

This fact is revealing. For one thing, it shows once again that an ultimate AMA will reflect the human condition. After all, not only is the human condition all that we have to go on, but morally relevant aspects of it are projected onto the world of objects as a matter of course. The transition of AMAs into human society will be eased if they embody the capacities that their human companions reflexively presume them to have, anyways. Further, from the perspective of the AMA, the only way that we can expect AMAs to recognize the significance of human actions, and human morality, when *they* see it is to embody a similar sort of agency, themselves.

Together, these two perspectives underscore the need for a comprehensive theory of moral agency according to which both are understood equally, a key aspect of which being a computationally friendly articulation of intentionality. This account will be provided through the analysis of conscience, following.

3. A PORTRAIT EMERGES

[14] Page 16.

All of the apparent obstacles to and reasons for realizing AMAs suggest something interesting. In setting out initial conditions to be met in implementation, a portrait of our final product presents itself, and it is of *ourselves*. An AMA constructed accordingly should mirror the human form and function, exhibiting the best of human morality.[15] Perhaps, even, developing its own moral theory.

Accordingly, should we understand how to build such an AMA, we must first understand the mechanisms of our own moral agency, including our predisposition for the creation of moral theories.

The wealth of insight available from which to begin this project is both a blessing, and a bane. It is a bane in that, spanning the extant of human history in all of its diversity, it presents an inexhaustible field for review. It is a boon in that there is little practical difference between the task with which we are charged, now - informing potentially moral machines - and that from which the wealth of insight derives - informing potentially moral men.

However, there is one important difference between the development of moral man and machine that does stand out. This is that the influence that social engineers have had over human agents has largely been limited to "post-construction programming."[6] Human beings are genetically limited to being human beings, modifiable only after this fact is settled, through education. Meanwhile designers of artificial agents appear to have a great deal more freedom in designing the perfectly moral artificial agent, perhaps engineering them to be perfectly moral critters from the get-go.

Given the obvious failure of human beings, generation by generation, to improve on their moral predecessors – Mohammed being a thousand years

[15] Some have presented contrary portraits, however. For instance, Nick Bostrom [11][12]makes many claims about the possible form of future artificial agents, but there are three that in particular seem to undo the potential morality of any such proposed agent. He expects that any given future agent need not have a human-like intellect, that the agent may be fixated on fulfilling any given end (i.e. "make as many paperclips as possible"), and that artificial minds can be easily copied. The model proposed here disputes all three in the following ways. One, the structure of the ACTWith model is human-like as it is inspired by and consistent with research both on artificial and human minds. Two, the model proposed here is motivated according to a basic fact about any natural agent, that it seeks low-energy (comfortable) situations in terms of its environment, and so, while the production of paperclips may lead to such a situation given a certain set of satisfaction conditions, that it might is not at all as arbitrary as Bostrom would paint it (it is a difference without a difference, once agency is adequately appropriated). Finally, the model proposed here denies that any moral agency can be easily copied, as it is essential that a moral agent inhabit a particular, unique situation in the world against which those of others can be compared, and from which any motivation to improve one's own or another's situation arises. Granted, Bostrom's focus is not moral agency, but ours is.

dead, Christ being more the 2000 years dead, the Buddha and Socrates more than that – this added dimension of control over the development of AMAs does sound a promising note. Regardless, I see the possibility of artificial moral agents NOT redrawn according to a human standard dead-end speculation for two reasons: 1) The recipe for human morality is rather straightforward, if not yet perfected – autonomy, compassion, and a tension between self-sufficiency and social-dependency are typical ingredients. Any alien form is difficult to constitute, and even more difficult to call "moral," unless essentially comparable to the human standard, leading to the second point. 2) Regardless of the eventual design of AMAs, the human standard is the only standard that we have. There is no other measure for a moral agent, thus making any effort into non-human designs a non-starter, at least insofar as the human standard is no longer applicable. Should we create a moral machine rather than the inverse, we are directed with even greater urgency to the essence of human morality perhaps best expressed by those long-lost moral exemplars.

As we can see, foremost amongst the challenges involved in realizing a fully ethical AMA is the articulation of moral agency in terms suitable for computational implementation. Our attention cannot stray far from the human subject, as origin both of theoretical form and moral function, regardless of the fact that any ideally moral human agent is the rarest of birds. As well, we cannot neglect the need for a computational architecture adequate to the task. And, of machine intelligence platforms unlike the human counterpart, there are many from which to choose. First, let's attend to the the relative adequacy of available platforms, as, in the words of Laszlo Versenyi, "How closely we can reduplicate a human agent artificially is dependent solely on our knowledge, that is, our technai, our technology."[6] [16]

Typical approaches to conceiving of AMA architectures have focused on direct importation of the products of moral thinking into an explicitly represented set of regulatory principles (top-down, traditional a.i.), rather than focus on the mechanisms from which such emerge (bottom-up). However, I feel that this approach is mistaken, and will here briefly argue for the bottom-up agent-centered generation and reinforcement of moral principles as only so far conceivable in terms of hybrid neural network models capable of both

[16] Page 250. Although, I disagree with this assessment. I feel that the fundamental limiting condition is the conceptualization of the human moral agency in terms that are, at least in principle, amenable to computational implementation, well prior to any technically specific application.

taking on externally given principles (top-down) and confirming, generating, and refining such principles through experience (bottom-up).

Of computational intelligence platforms, there are essentially three: top-down, bottom-up, and hybrid.[17] Most discussion centers on top-down approaches, largely due to three concerns and one practical limitation, all of which I will address by counterpoint in turn. First, top-down approaches permit the explicit limitation of agency by principles encoded in similar forms to those of the moral theories in which they appear. On the basis of this concern, we can discount top-down models for three reasons. One, such a limitation precludes the articulation of a fully ethical agent at all. Two, there is no guarantee that the principles, as given, can reliably result in moral action in the first place. Three, hybrid models are also capable of this mode of moral input, while also overcoming the first two shortcomings. Second, externally derived theoretical prescriptions are presumed to be more sophisticated and more powerful than any equivalent originating from within an AMA itself, as they cover situations with which any given agent might have no practical experience, and in terms with which that agent might not be familiar, until the morally challenging situation in question arises. This apparent shortcoming does bear some weight, but is explicitly overcome in the ACTWith model to follow. Third, researchers tend to expect moral perfection from AMAs, and only top-down models seem to deliver the sort of control over action that might make moral mistakes impossible.[18] This issue is related to the first, and may here be overlooked for similar reasons.

Finally, preference is given to top-down explicit symbolic approaches because researchers themselves are more comfortable with these implementations. It is notably easier to transcribe moral rules into computable code than to conceive of a genuinely moral machine. At the same time, by the same people, bottom-up approaches, relying as they do on designing and then training neural nets to deliver appropriate output, are often enough treated in derogatory terms, called "black boxes," and treated with suspicion: "this kind of modeling inherently produces agents that are liable to make mistakes."[8][19] Granted that work in hybrid models is ongoing, training methods and extraction algorithms in constant refinement, the mistakes in question have not

[17] Their relative merits are very well know, and shall not be discussed here. Top-down and bottom up architectures specifically in the context of moral agency are the focus of [13]. The current work is conceived with hybrid models in mind.

[18] This may in part be due to as yet unanswered questions about the status of artificial agents, especially concerning liability for agent actions under civil law. We will briefly entertain this issue at the close of this chapter, but some insight can be found in [14].

[19] Page 260.

been “moral” mistakes. These mistakes merely represent instances where neural-network designs fail to deliver expected results, perhaps due to implementation specific flaws rather than due to limitations inherent in the platform, itself. It may be, as we shall see in discussion below, that what may appear to be a 'mistake' in a lab context may take on a different significance in a moral context, wherein deviation from expectation (such as is the case in 'conscientious objection,' for example) becomes necessity rather than something to be avoided. In any event, with explicit symbolic models, researchers can “see” what their creations are doing, both inside and out, most easily, helping them to ensure that output meets their expectations, and this provides prima facie reason against the employment of architectures that do not deliver such conveniences.

This is not to say that top-down approaches can not lead to AMAs that make moral mistakes. Even with all of the vaunted control afforded, this approach may in the end fail to deliver desired results. Perhaps the most famous, and still prevalent, discussion in this vein centers on Isaac Asimov's famous three laws of robotics, but similar concerns can be drawn from any set of principles. Principles, regardless of their systematicity, can and likely will be drawn into conflict given a sufficiently complex environment. And, even with the perceived environment simplified, as in a lab context, an AMA is liable to make a moral mistake when it weighs one principle over another, executes action on this basis, and yields a less than optimal moral result, no less often than would a human agent relying on similarly constrained information. Further, where a human agent can act in some way based on limited experience when regulatory principles are absent, a solely top-down AMA has no such recourse, and so in similar instances may either apply an inappropriate principle, or fail to act at all. In the end, any agent governed purely, or even for the most part, by externally derived and prescribed regulatory principles may seem reliable in the lab, but likely cannot perform the roles of a moral agent in the full sense in the real world where moral issues actually surface.

It is not enough to simply cut and paste moral principles into the top layer of a computational model and declare the resulting machine a moral agent, not in any robust sense. However, a similar approach is often employed in rendering the interpretations of moral theory from which computational implementations are drawn. It is common enough for interpreters of Kant and Mill, to cite two figures to which we shall divert considerable attention in the pages to follow, to cut the top out of the complex originals and to proceed as if these simplifications were adequate. These reductions are then passed off as

productive of a working understanding of the originals, when this is – as we shall see – simply not the case. Thus, it is no wonder that designers of AMAs who appropriate these interpretations fall into the trap of believing that top-heavy rationalist models of artificial agency best embody the top-heavy rationalist accounts of human agency so given.

Typically, in the conception of AMAs, moral principles distilled from the proper function of human agents and which are then expressed in the (explicit, symbolic) format of moral theories are post hoc applied to artificial agents. This approach ignores two critical points, the first of which we have already seen. The fact is that a comprehensive theory of moral agency adequate to the task of computational implementation has yet to be articulated. Standard interpretations of traditional moral theory on this count fail us. Additionally, and most importantly in terms of selecting a computational platform in terms of which an AMA is to be conceived, this approach fails to appreciate that any truly suitable moral framework only results from the proper function of the architecture from which it emerges. In other words, the embodied mechanism must produce its own moral principles, or at least include the means to test and modify given principles against ongoing experience, just as human beings have done, themselves. Otherwise, rather than a truly moral machine, we are left with a very complicated thermostat, into which guiding moral principles serve as arbitrary constraints imposed after the fact, imported as they are from completely different instances, as is the case when moral frames derived from human implementations are imported into computational instances non-reflective of this condition, and vice versa.

Most importantly, AMAs only provide an important mirror in light of which insight into our own moral agency can be revealed, and so deliver on this most compelling set of reasons for their development reviewed earlier, insofar as they in fact are informed and function as we ourselves do. Human beings, after all, are not merely spoon-fed moral principles from which all future actions proceed accordingly. They make mistakes, learn from these mistakes, do better, and even become better. This process is an essential part of human moral agency, adequately understood, highlighting not only the significance of mistakes in the development of the moral agent, but also the limitations of our own moral judgments. Any given set of moral principles is in constant refinement. And, regardless, who are we to judge right and wrong in beings whose embodiment differs so much from our own that our rules for efficient and effective engagement with the world of objects do not efficiently and effectively apply? No rational human being would retain such a principle.

To presume that an AMA would do so is to merely conceive of an irrational AMA.

These considerations put final constraints on the eventual form of any adequately realized AMA. Should we wish to retain any moral prerogative over AMAs, and expect them to confirm moral principles emergent from the human moral condition, we must endeavor that their designs fall within the scope of human morality. We must see them as ourselves, should we expect them to act as if they were. They must not only do, but *be*. And, in this light, the case for bottom-up hybrid models rather than top-down approaches is strong, as human development, learning, and on-the-fly modification of (possibly top-down externally given) regulatory principles are rather effectively modeled on this approach, more so than on others.[13][15][16] Indeed, including conditions already reviewed and some yet to be, this fact is already acknowledged:

> Eventually, there will be a need for hybrid systems that maintain the dynamic and flexible morality of bottom-up systems, which accommodate diverse inputs, while subjecting the evaluation of choices and actions to top-down principles that represent ideals we strive to meet. Depending on the environments in which these AMAs operate, they will also require some additional supra-rational faculties. Such systems must also specify just how the bottom-up and top-down processes interact. [13][20]

4. CHOOSING A MORAL FRAMEWORK

In articulating a moral framework suitable for artificial moral agency, we must uncover necessary resources from existing moral theories and interpret them in such a way as to facilitate computational implementation. However, there are two potential sources of distortion at work in this process that potentially undermine both the efficacy and legitimacy of the results, one of which has already been mentioned. First, standard interpretations of moral theory ignore aspects of original theoretical insight that have either fallen out of favor within a theoretical community, or that – more difficult to realize – had not been adequately emphasized during original expositions because concurrent audiences would have held tacit understandings of the importance of these aspects that would have made any expositional emphasis redundant. The point being, here, that in interpreting traditional theories, we must

[20] Page 459.

endeavor to give adequate weight to dimensions that may have fallen out of contemporary theoretic focus, as well as recreate the original intentions of their creators, or we will succeed only in deriving distorted portraits of moral agency. Second, the computational architectures themselves require certain forms of information for effective importation, and this is also a potential source of distortion, both during the appropriation of moral theory and during its implementation. For example, traditional a.i. requires explicit rules for direct implementation. Any talk of emotive moral grounds, for example, are not so easily captured on such platforms, and so any theoretic exposition stressing these aspects must either be heavily edited, or neglected entirely.[21]

One possible response is to do away with the bulk of moral theories, all of those that do not fit with a conception of agency consistent with chosen computational platforms. In effect, this is exactly what the simplifying interpretations of traditional moral theories pointed to above do. By this option, should traditional a.i. form the mold, morality should be reduced to rationality, perhaps to pure utilitarian calculation, or at the very least to statements of pure principle. Such a conclusion should promise to make computational implementation much easier, however, as we have seen, at the cost of moral agency in the strong sense altogether.

Consider in this vein the interpretations of moral theories afforded by Allen et. al. [8] Allen et. al. appear to non-critically accept given interpretations of moral theories, testing their viability in such stripped forms:

> In Mill's utilitarian terms, we might say that an agent is morally good to the extent that its behaviour positively affects the aggregate good of the moral community. In this sense a robot could be said to be a morally good agent to the extent that it has been programmed to act consistently with the principle of utility, regardless of how this behavioural result is achieved. For Kant, however, any claim than an agent is morally good (on either a specific occasion or in general) implies claims about the agent's internal deliberative processes. On Kant's view, to build a good AMA would require us to implement certain specific cognitive processes and to make these processes an integral part of the agent's decision-making procedure. [8][22]

[21] Arguably, and oddly enough, the only tradition that holds to a similar process of arriving at moral principles is the family of theories of divine command. This correlation opens some interesting avenues for discussion, specifically about man as godlike creator of "artificial" life. But, it extends to human life as well, as all life, literally, is an "artifice" when intentional agency is taken to be behind its origin, even human life.

[22] Page 253.

In relying on pre-packaged standard interpretations of moral theory, Allen et. al. miss an important insight into human moral cognition common to both Mill and to Kant. This is the central role of conscience in both, generating and testing moral actions and moral principles.[23]

Regardless, these simplified versions of traditional theoretic accounts of the great complexity that is human morality fail to deliver workable moral frameworks. Due to the apparent difficulties in adapting the given interpretations of moral theory into a conceivable AMA architecture, Allen et. al.'s excursus into modeling human moral theory in an artificial agent expressing the human moral condition is short lived. As for utilitarianism, the computational demands of testing outcome utilities for innumerably possible action paths prove unmeetable:

> The crucial problem for the consequentialist approach is that utilitarianism would seem to be a computational black hole. To implement the theory, the effects of an action on every member of the moral community must be assigned a numerical value. The sheer impracticality of doing this in real time for real world actions should be evident, especially if one considers the fact that the direct effects of every action ripple outwards to further effects that also affect aggregate utility. We are confident that interactions between these effects would make practical computation of long term consequences an intractable problem. Furthermore, if utilities must be computed for as long as an action has an effect in the world, potentially for all time, there is the risk of a non-terminating procedure. Even if the termination problem can be solved, it would also require the implementation of a comprehensive scientific theory to predict the nature of these long-range effects. [8][24]

Note the focus on computability, especially over ever extending effects rippling out through the action space shared by all agents, the best interests of which must remain ever under the guise of an ideal consequentialist agent. There is no doubting that these are real troubles, actual show-stoppers, so far as the characterization of consequentialism from which they begin is correct.

---

[23] Rather, Allen is content to identify these theories by their differences, as if they actually apply to two radically different sorts of morality, a mistake as well made by contemporary experimental ethicists. “Roughly, the classical utilitarians held that the best actions are those which produce the greatest happiness for the greatest number.”(page 252) Meanwhile, Allen's characterization of Kant focuses on the categorical imperative in its most popular first form - “Act only on that maxim through which you can at the same time will that it should become a universal law.” - a species of directive most obviously absent from Mill.

[24] Page 256.

However, in the focus on computability, moral theory as given from and for human agency is saddled with limitations according to implementation, leading the researchers to appropriate a wholly deficient understanding of moral theory.

It should be no mystery that this approach should fail. It does not take on the utilitarian theory as originally conceived to begin with. If one returns to Mill's original work, and studies it with an eye to agency, one finds two surprising things. First, Mill himself directly confronts this issue of computational impracticality in the form of a common objection to his consequentialist program. He writes that:

> Again, defenders of utility often find themselves called upon to reply to such objections as this – that there is not time, previous to action, for calculating and weighing the effects of any line of conduct ... The answer to the objection is this, that there has been ample time, namely, the whole past duration of the human species. [17][25]

In essence, he advises that we go with what is given until such fails to deliver appropriate results. The sort of revision at which we would expect a hybrid agent to excel. Still, how are we to measure the propriety of said results? Well, Mill directly answers this question, too. He answers that all moral judgements, indeed all evaluation of action, relies on the functions of conscience. In fact, there is an entire chapter of his famous text *Utilitarianism* dedicated to delivering this answer, something to which we will pay more attention after looking first at Allen et. al.'s troubles with Kant.

As for the Kantian portrait, things get a bit more interesting. In passing through Peter Singer's analysis, Allen et. al. seem to recognize the affective foundations of Kantian moral agency. The motivation to right action arises not from principle, but out of goodwill. Yet, from here, they detour through the Kantian distinction of the divine versus the human will to action. Kant famously suggested that a divine will, being all-knowing and all-powerful, has no use for "oughts," and being entirely good there is no further need to posit a potential conflict between good and bad will. Rather, a divine will simply *wills* the results of action be realized, and so they are realized, and being divine, are good. Accordingly, Allen et. al. redirect the effort to iterating the artificial moral agent in this form, no longer in human likeness, on the consideration that an artificial agent differs from a human agent in the same way, and at once trying to dodge the principles-in-contradiction bullet. AMAs should be held to

[25] Page 273.

a different standard, a higher standard, on their estimation. And, though this may no longer reflect the human condition, or in the end even add up to moral agency in the Kantian sense at all, so far as building AMAs goes, it is good enough:

> If, as Kant appears to think, being a moral agent carries with it the need to try to be good, and thus the capacity for moral failure, then we will not have constructed a true artificial moral agent if we make it incapable of acting immorally. Some kind of autonomy, carrying with it the capacity for failure, may be essential to being a real moral agent. However, as we suggest below, the basic goals when constructing an artificial moral agent are likely to be very different than when raising a natural moral agent like a child. Accordingly, it may be acceptable to program a computer to be incapable of failure but unacceptable to attempt the analogue when raising a child. [8][26]

Now, I do not agree that the "basic goals when constructing an artificial moral agent" can be to *actually* construct something that is not a "true moral agent." This sounds too much like setting out to bake a cake, and ending up with an omelette. However, in the present context, the issue is that Allen et. al.'s concessions preclude the prospects of a fully ethical artificial agent right out of the gates. For instance, they characterize AMAs in a way that seems to be missing one crucial aspect of human moral thinking, and this is the capacity to visualize results of actions that might have been, a cognitive exercise closely tied to the ability to do otherwise (free-will by another name) obviously being the ability to consider having done otherwise in similar situations, past. In their words:

> We humans typically muddle along making mistakes while harbouring private regrets about our moral lapses, which occur more frequently than perhaps we care to admit. But while we expect and, to a certain extent, tolerate human moral failures, it is less clear that we would, or should, design the capacity for such failures into our machines. [8][27]

It is difficult to understand how an incapacity to make a moral mistake should deliver a capacity to judge over the morality of any given action or agent in the first place. Without this ability, we are left with something like a pet rock that walks, unable to err because it is unable to do otherwise than it already is, with that, in any case, not stopping its human companions from

[26] Page 254.
[27] Page 255.

seeking to emulate its steadfastness. And, this is not moral agency in any robust sense, at all. Though Allen et. al. make it a point to proclaim that "we shall probably expect more of our machines than we do of ourselves," we expect very little of ourselves, indeed, if we aspire only to the creation of household slaves mistaken as moral gods.

However, there is a way to do both ourselves and our creations justice in this matter, and it begins with the road that Allen et. al. (and so many others) fail to travel. In order to articulate a robust moral agency, we must first return to original moral theory and extract a portrait of a robust moral agent. Lopping off the top parts where all the easily transcribed principles reside, and then pasting them into a computer, is not enough. As noted, above, Allen et.al. make a mistake common in the literature on the subject. They fail to note the common grounds for moral agency iterated by both Kant and Mill: the conscience. And, it is here, on this common ground, that the possibility of articulating human-like moral agency in computationally friendly terms resides.

In the third chapter of Mill's *Utilitarianism*, titled "Of the Ultimate Sanction of the Principle of Utility," he specifically posits conscience as the ultimate judge for any moral action. It is, by his description (and universally reflected in Western moral theory), a "mass of feeling which must be broken through in order to do what violates our standard of right, and which, if we do nevertheless violate that standard, will probably have to be encountered afterwards in the form of remorse."[17][28]

This recipe reveals the obvious failure of standard appropriations of Millian utilitarianism to appreciate the fact that regardless of the understanding of the utilitarian calculus, resultant calculations (especially those that deviate from what is given in accepted convention, for instance, and that fail to deliver appropriate results) must pass the muster of the conscience in order to qualify as right action. Furthermore, this muster implies a certain form to the utilitarian logic, that an action is permissible that does not contradict a "standard of right," otherwise discouraged by the "mass of feeling" that is the most recognizable aspect of conscience. In this form, as we shall see in greater detail, Mill begins to look more and more like Kant, sharing common affective grounds for morality in conscience.[29]

---

[28] Page 277.

[29] Even on Allen's assay, this is true. To be fair, though they find other reasons to refuse the Kantian moral agent, Allen et.al. seem to recognize the possibility that Kantian moral theory can escape computational difficulties to which other deontological approaches are prone, though they fail to specify how this is possible. Consider for instance this exception given

Now, we may consider further, why does a person strive to do the right thing at the right time? To satisfy some set of rules? To maximize happiness? To this, in Mill we find a not-so-surprising answer, that in a person there is a tendency to "...feel it one of his natural wants that there should be harmony between his feelings and aims and those of his fellow creatures."[17][30] This feeling is the "natural basis of emotion for utilitarian morality."[17][31] It encourages people to cultivate talents through education, encouraging a feeling of "unity with all of the rest" of humanity, "which feeling, if perfect, would make him never think of, or desire, any beneficial condition for himself, in the benefits of which they are not included."[17][32] In short, it is through compassion, extended from and consistent with a selfish will to maximize one's own happiness, that a person strives to do the right thing on Mill's account. Harmony, inside and out.

We can now color these results with the unstated. Mill composed his *Utilitarianism* as a sort of response to Kantian moral theory. In so doing, he opens that text by setting his own theory against Kant's. From that point of origin, Mill focuses on the self-identification of one's own ends with those of others, through the "... feeling that the interests of others are [one's] own interests."[17][33] And, he is able to focus, indeed must focus, on these aspects of conscience, of human morality, because Kant does not. Kant, as we shall see in greater detail later on, is more focused on moral feelings internal to the person, himself, to become a better person, especially a more moral person. Thus, to appropriate from Mill a robust portrait of moral agency, we must expropriate his compliment from Kant.

What does this compliment consist in? To Mill's picture - conscience as a motivational mechanism for the identification of self and others' interests – add a similar relationship within the same person, both before and after action, and the role of conscience in the human-like AMA begins to resolve itself. This internal relationship within the individual agent represents perhaps the best simple answer to the question of moral motivation - to become the best person that one can be, the realization of which results in a harmony not only between people, but within them.[34] Finally, whether we look to Mill or to Kant

on page 257: "From a computational perspective, a major problem with most deontological approaches (with the possible exception of Kant's) is that there is the possibility of conflict between the implied duties."[8]

[30] Page 281.
[31] Page 279.
[32] Page 280.
[33] Page 279.
[34] This is essentially the position taken in this chapter, as well being reflected in [18].

for inspiration, it is conscience that represents the mechanism by way of which this harmony, of either origin and end, is attained, and retained.[35]

But, of what sort of mechanism must conscience consist to deliver such results? Can adroit interpretation of traditional moral theory provide an answer to this question? Now that we have uncovered the fulcrum of the moral mechanism in moral affect, we may do well to look for insight in an addition to the library of moral theory that attends directly to this aspect. One particularly efficacious description of the cognitive processes that are in need of computational representation comes from Adam Smith, author of the famed text *On the Wealth of Nations*. In an earlier text, *The Theory of Moral Sentiments* - the lessons of which motivate the economic processes that are the focus of *Wealth* - Smith describes what goes on inside a human moral agent when that human being is *being* moral:

> By the imagination we place ourselves in his situation, we conceive ourselves enduring all the same torments, we enter as it were into his body, and become in some measure the same person with him, and thence form some idea of his sensations, and even feel something which, though weaker in degree, is not altogether unlike them. His agonies, when they are thus brought home to ourselves, when we have thus adopted and made them our own, begin at last to affect us, and we then tremble and shudder at the thought of what he feels. [20][36]

In this statement, we see a number of cognitive processes in need of representation, and these are related in a specific way. There are clearly the affective and rational components typically presumed to be pieces of the moral puzzle, and they are given in a movement of compassion whereby one's interests, and more strongly one's situation, become those of another. Smith's portrait of compassion is not mere sympathy – feeling the same as another feels, mirroring or modeling another's sentiment on the basis of affective cues, for example. This is a completely immersive condition wherein another's situation is taken for one's own, and from this translocation, the feelings of what it is like to be in that situation (instead of one's own) become one's own feelings. And this tells us specifically in what ways the cognitive processes in

[35] Childress proceeds on a similar understanding, with conscience responsible for a sense of peace and wholeness in the maintenance of personal integrity while guiding away from immoral actions.[19]

[36] Section 1.1.2. Smith is most famous, today, for authoring *Wealth of Nations*, but before this work he developed a powerful moral theory grounded in compassion. Relative ignorance of his moral theory has led to widespread misappropriation of his economic theory.

need of modeling are related. One's feelings are not what is immediately shared. Rather, it is the situation that is shared, with one entering 'as it were' into another's body. And, it is from this perspective, and with the information thus provided, that morally requisite harmonies can be realized, and even come to explicit awareness. The ACTWith model represents the process thus described, and we will return to this passage after the ACTWith model is detailed to show how the model effectively serves as a vehicle for Smith's insights into the mechanism of morality. First, however, there are some things to clear up about just how conscience is supposed to get all of this moral work done.

5. CONSCIENCE, THE MECHANISM OF MORALITY

Smith's moral theory reaffirms both the need for human-like AMAs and the use of hybrid architectures in their design. After all, how is an AMA to take up a human situation in the performance of morally relevant functions if that AMA has no basis for taking up that situation in the first place? Aspects of embodiment in common are prima facie inroads permitting other-agent centered simulations Amongst these aspects are a mix of affect and intellect, two modes best represented in hybrid architectures. In particular, we are led to one aspect of human agency that has been presumed to set humans apart from other creatures in the moral hierarchy, conscience. An adequately resolved AMA is a conscientious agent, by the portrait of moral agency herein emerging.

What is conscience, and how does it culminate in the cognitive/ computational processes that amount to morality? Conscience is an old term for a family of phenomena, ranging from voices that warn of impending wrong action to providing the fundamental basis for international humanitarian law. It is an extremely complex concept, often confused with consciousness, and more often burdened with seemingly contradictory tasks as it has traditionally been associated with such things as self-preservation on the one hand and altruistic selflessness on the other. Here, we will take on the issue of consciousness, first. In clarifying the relationship between conscience and consciousness, we may gain some insight into the more complicated issues involving self-preservation and altruism.

Conscience is related to consciousness. In fact, the use of the term conscience predates that of consciousness by three centuries. In fact,

"consciousness" comes from "conscience," not the other way around.[37] However, the issue of consciousness receives a great deal more attention in contemporary philosophy of mind and psychology, wherein it is presumed that the two terms represent distinct aspects of the human condition, regardless of their common history.

We may gain clarity on both terms by assessing first their structural similarities. Both consciousness and conscience consist of conjunctions between a prefix "con-" and a root, "sciousness" and "science." "Con-" means "together" or "with." It is a prefix that indicates synthesis. "Sciousness" was briefly considered to be that from which consciousness arises in William James' controversial 10th chapter to his landmark text, *Principles of Psychology*.[21][22][23][24] He arrived at this speculation through directed, educated introspection. Introspection was the only psychological tool available at the time, the only tool with which inquiry into the nature of consciousness could proceed, at all. At the limits of his introspective powers, James found a rolling stream of sensation that evaded a discrete characterization, yet that appeared to underlie every conscious moment at the same time. "Con-" "-sciousness," thus, can be taken to mean the synthesis of merely felt moments into discretely realized phenomena.[38] Accordingly, sciousness can be understood as the felt ground of all discrete thought, with consciousness, in Cartesian form, being what is clear and distinct built from this muddy, affective landscape.[39].

The function of "-science," the root of the term "conscience," can be assessed similarly. Formally, "science" implies a specific field of knowledge and inquiry, the ideal organization of which is constituted by a certain set of objects interrelated though the systematic application of certain field-specific principles. One example of such an ideally constituted science is Chemistry. Chemistry consists of a field of entities related by chemical laws and

[37] See for example http://www.etymonline.com/index.php?search=conscienceandsearchmode=none. Last accessed February 15, 2010.

[38] This is effectively the operation employed through the use of mathematical algorithms in hybrid models. For discussion on James and sciousness on this point, see [15].

[39] Through this discussion, I avoid taking the tangle of characterizations of consciousness given in the literature head on. The issues are far too complicated for exploration here. For instance, [3], and [25] both consider consciousness as aspects of moral agency, but the former does so in terms of biological/mechanical hybrids while the latter does so in terms of the extended mind. Meanwhile, as is well known in the philosophy of mind, there are various types of consciousness to be made sense of, as well. Drawing meaningful relationships would demand more space than we can dedicate, here. However, preliminary equivalences in terms of consciousness between humans and machines are drawn while defending materialism with some attention to moral agency in [26].

constitutive of chemical theories over a specific set of chemical objects. This use of the word "science" is clear enough. However, in the description, one thing is missing – the chemist, herself, situated in the midst of her field. There is no field of Chemistry without a chemist somewhere in the middle of it.

The presence of the chemist reveals something universal about the use of the word "science" that ties all of the seemingly discrete fields of inquiry together, however formal or informal. It is from this universal implication that the term "con-science" can be construed. "Science," as the root of "con-science," represents what it is to be situated within *any* field of *any* set of objects, however non-specific, which are bound by *any* set of relations, however non-systematic. Sciences, like Chemistry and Physics, are simply ideally ordered limit cases of this phenomenon. In terms of the general analysis, however, "science" can be taken to name *any* field in terms of which *any* person is (or persons are) embedded. Moreover, this characterization grounds any form of inquiry, and indeed any action and activity. After all, it is only ever in terms of one's situation that an agent acts, and searches for scientific truth. "Science," thus, can be understood as the "scene" from within which one sees and understands the world, and from within which and in terms of which one acts, experiences, understands, learns, feels or fails.

"Science," in this sense, is reducible to "situation" in a formal sense, being the irreducible complex of agent and environment, understood from the perspective of the experiencing agent, or subject, herself. "Conscience" can be understood, then, as the synthesis of subject-centered situations, the "what it feels like" to be in a place at a time. Finally, and most importantly for any computational appropriation of this picture, in the comparison of embodied situations, conscience provides information on the differences – both discretely and implicitly realized – between any one situation and any other. It is from this information that relative evaluations of situations, as good or bad and so on, are derived, as we shall see in greater detail as this discussion moves forward.[40]

Consider in this light the role of the so-called "voice" of conscience. Even this most familiar characterization as a universally recognized voice which rises against acting towards morally repulsive ends, cannot be merely a

[40] This is a much stronger process than that described in theories of Goldman and Hurley[27][28]. The differences between embodied cognition (EC) and mindreading/simulation are summarized in [29]: "EC holds that non-mentalistic embodied practices are developmentally fundamental, and they constitute the primary way we understand others. Thus, EC holds, the emphasis in cognitive science should be on primary embodied practices, not specialized and relatively rare cognitive skills of mindreading."(page 124)

"voice." After all, for it to fulfill even this seemingly simple function, the operations of conscience must extend through all processes from which possible actions and their resulting ends are drawn – in line with the analysis, above. This is a much more complicated role than might be fulfilled by any simple voice. There must be more to the story than a "voice."[41] Conscience must act as the steering mechanism of the entire embodied complex that is the moral agent, if it is to reject some ends, surfacing as the warning voice of conscience, while either endorsing others in forecasts of harmony, or at east not standing in the way, remaining silent and permitting action to proceed for lack of interference.

This characterization requires more detailed discussion. Most of all, it requires translation into fundamental currencies common to both computational/artificial and organic/natural agents. This account must render cash value, and we will find this in basic currency common to all things in the universe – information, and energy.

In these terms, conscience can be understood as a mechanism integrating information from all sensed aspects of the embodied agent in the relative assay of possible situations, serving as a motivational and self-preserving extension of basal homeostatic mechanisms common to all sufficiently complex organisms.[42] Conscience is motivational in that any organism that is capable of ascribing relative values to situations will seek those situations that are valued more and avoid those that are valued less. It will seek those that feel good, and avoid those that feel bad, as these situations are effectively environments in terms of which, if actually taken up, that organism must subsequently reach homeostatic equilibrium. Harmony.

In aiming for harmonious integration with the (social, natural, and metaphysical) environment, conscience can be understood as the extended homeostatic function of the embodied moral agent to sustain personal integrity in the face of a changing world. The "harmony" achieved represents a low-energy, relaxed state. In Socratic terms, one of "leisure." It is a long road from low-energy to Socratic leisure, but somehow people have been able to get there, and the reasons for doing so are not at all different from the reasons for any other critter performing any other action toward any low-energy, relaxed,

[41] See [19],[30],[31],[32],[33],[34],[35],[36],[37],[38],[39],[40] for an introduction into some basic issues having to do with conscience, especially its early psychological interpretations and attempts at naturalization. The influence of these early works run throughout the present work, forming the basis for [41] and [42].

[42] Consciousness, as well, has been understood as an extension of homeostatic mechanisms. See [43].

comfortable situation. In any case, in order to achieve and maintain low-energy situations (regardless of their complexity), an organism must take advantage of the information available about past, present, and possible situations. The mechanism of this information processing is given here as the ACTWith model.

The ACTWith model is a generic mechanism, the basic operations of which are observed in forms of life from amoeba to bivalves and upwards. Where as the functions instanced in these examples are easily observed, in human agents the processes are more complex.

In general, regardless of the form of embodiment, the processes in question amount to respiration, the common operation being a regulative opening and closing to the external environment.

In the case of bivalves, water and with it food and gases are carried into the organism, while the products of their metabolism are carried out. In the case of human conscience, information is carried into the organism, and the products of its metabolism, understanding, wisdom, and actions proceeding from this basis, are – in a somewhat different sense, though in a way poignantly revealing in the expression – "carried out."

In the case of a clam, regulated opening and closing to the environment leads to the accretion of a protective shell, the production of living organic mass, and perhaps even the refinement of a pearl.

In the case of the human-like moral agent, the regulated opening and closing to environmental input leads to the accumulation of experience[43] which is used to guide future operations of the same mechanism, for its protection, its survival, and for the refinement of a treasure particular to this "rational animal" identified as such since the beginning of philosophic record, practical wisdom.

As this picture paints it, in designing an AMA adequate to the task of being "truly moral," we must engineer more than a mere moral algorithm.

We must give it breath. It must inhale and exhale morally relevant information. And, in this, we find the common space of all human and non-human action that provides the final bridge between human and artificial moral agency.

[43] Initially understood as memory, see [44], but eventuating in embodied adaptations due to peripheral attunements, i.e. hormones and general metabolism, over time.

6. CROSSING THE DIVIDE - CASH VALUE FOR MORAL TERMS

As discussed above, conscience lies at the heart of traditional moral theory, yet in contemporary interpretations and in today's applications of them, in designing AMAs, it has all but been forgotten. The current model takes conscience as the fundamental aspect of a moral agent, offering an alternative to rationalist interpretations of traditional moral theory as reviewed above, reinstating conscience to its traditional role as briefly described in Mill, preceding, and to be reviewed in Kant, proceeding. Now, we must place all of this in context of the energetic landscape upon which all naturally occurring agents act, one way or another.

The most pressing issue once any articulation of human moral agency is adopted for implementation into a computational model is to understand both it and the information that it delivers in terms relevant to both humans and to AMAs in the same way. As we saw in Adam Smith's description of morality in action, with emotion and reason clearly related, hybrid models present themselves as most appropriate vehicles for morally capable architectures. In terms of a hybrid model, Smith's *Theory* is bottom-up. An emotion, compassion, grounds morality on Smith's account. Morality begins with taking up another's situation for one's own while remaining open to the feelings that inhabiting such a situation generates, "enduring all the same torments." On the basis of this experience, one only then comes to understand what it is like to be in another's situation. On Smith's description, then, the central aspect of moral analysis is the situation of an other.[44] The question then becomes, how is this situation to be understood, equally, by both human and artificial moral agents?

In brief, morally relevant aspects of any given situation can be conveniently articulated in energetic terms. Taking the natural environment as a baseline, metabolic potentials to both overcome obstacles and to reap benefits provide an accessible measure of agent sufficiency in meeting environmental challenges, to prosper in the given environment, and so to survive. These terms are appropriate in any analysis of any form of natural agency – though perhaps not divine agency - and so are appropriate for the relative evaluation of any set of situations.

[44] Equally, this applies to situations of one's own. Though we shall focus on inter-agent dynamics here, the process is essentially the same.

Indeed, it is the essence of a moral agent to evaluate situations, and so the actions that result in them, in terms of their energetic cost/benefit. This assertion requires some substantiation, and this will come as the discussion proceeds, but it is immediately verifiable in the analysis of one's own experience. There is a reason, for example, that I keep the television remote control next to me, and feel uncomfortable when I have lost it, actually having to get up and cross the room to change the channel. I feel how much easier – better – life is with a remote control, than without it. Moreover, I also remember life before remote controls, when I felt no such tension. Having to walk to the TV to change the channel was the only situation I had experienced.

Conscience, as can already be seen on the basis of its prior description, is the mechanism whereby this relative evaluation of situations takes place. Without a situation to compare, there is no stress – there is no discomfort felt without a remote control when there is no prior experience of life with a remote control. However, the energetic basis of this logic lies so deep beneath the layers of discourse on the subject of human agency as to have remained effectively hidden from theoretic account. So, here, we must start from the beginning.

All beginnings and all ends of all actions are situations. Every end of action is itself a beginning, providing opportunities to move to still further ends. Conscience motivates toward situations according to one's capacities to meet the terms of those situations. This concept of agency takes as fundamental a capacity to evaluate available actions and ends, with varying degrees of powers over the selection, attainment, and maintenance of such specific to the agent in question. Every agent seeks situations in which its needs are met, and in which it can meet those needs. This is because, if one attains some end in action, he will have to live there, or die. Some fish keep to warm waters, for example, while some stick to cold. Some bacteria seek reductive environments, some oxidative. Human beings seek the "good," and avoid "evil." In the end, the "good" situation is that in which an agent's rather expansive needs are easily met, and "bad" that in which they are not. So, conscientiously motivated, agents seek situations in which they can survive, or often enough do more than merely survive. They seek situations in which they can survive easily, be comfortable, have more than they need, enjoy luxury.

What separates situations, spatial-temporally, is the energy required to move from one to another. They can also be distinguished by the energetic return to the agent upon their attainment. This energetic return can be measured in various ways. It need not be in direct terms of energy. Consider the energy expended by the soldier seeking, and attaining, high-ground in a

field of battle. Attaining the high-ground often incurs high costs, well above those of the direct metabolic expenditure involved in lugging war-equipped bodies up steep inclines. This high cost must then be subtracted from the return, and the energetic return upon gaining the high ground comes largely in the form of information, information that improves the soldier's capacities to perform his given function – to hold the field, to advance the front, and so on. Thus, gaining the high ground lowers the potential energy costs associated with performing the same functions without the high ground, and this is where the energetic cost/benefit swings in the favor of climbing that hill.

On this account, the common currency of natural energetics provides a universal framework for the relative evaluation of situations, regardless of their forms.[45] Generally, agents seek ends the attainment and maintenance of which require the least possible energetic input. Recall the TV remote control illustration, for example. In keeping the remote close at hand, I - like any other natural agent - exercise intensions toward ends that are easy. Because it takes energy, metabolically measured in calories, to do anything, agents – myself included - tend to put themselves in situations where their needs can be met with the least expenditure of energy. Putting one's self in such a situation is a "good" thing, in the opposite, a "bad" thing.

Likewise, putting another agent into a situation in terms of which it cannot live easily is never a "good" thing. Putting another into a situation in terms of which it cannot survive is certainly a "bad" thing. Here, we find the intersection between moral actions and any other sort of action. Moral agents, thus, are no different than any other agent. And, human agency is not unlike any other form of agency. All keep to comfortable situations – low-energy situations - striving through action to achieve certain ends, with humans only differing from other types of agent in possessing a *limited* capacity to determine for themselves what these ends might be. Here, in the directed, energetic comparison between ends of actions, we have a genuine moral currency applicable to both human and artificial moral agents.

When the needs of an agent are not met in a given situation, then that agent is motivated to change its situation. This motivation may be understood

[45] Every thing in nature moves to meet the terms of its environment. Every thing in nature achieves balance between inside with out, reaches equilibrium, and so either remains stable in the face of change, or changes into a form that can. In extreme situations, minerals fracture, gases become plasma, and stars implode. When winter is deep and cold, and food is scarce, it costs more energy to find the food than the food can reliably deliver. In such extreme environments, animals hibernate, and human beings, unable to balance this energetic equation, die. There is nothing controversial about this picture. The only thing that may seem unlikely is that morality operates over the same set of terms.

in terms of an imbalance, which may be understood in terms of energy. When it requires more energy to survive in a situation than can be regained in that situation, for example, an imbalance results. The agent feels this imbalance as a tension between where it is and where it would rather be – in a situation in which there is a more positive relationship between energy in and energy out – given that it has grounds for conceiving of a situation in which it would rather be.[46] It is conscience that contains this tension, as it is conscience that holds the present unsatisfactory situation in comparison with another, possible, better one. Thus, relatively evaluating the given situations in terms of agent-specific need, it is also conscience that seats the motivation to act.

One way to picture conscience as motivational is to recast the preceding analysis of conscience as "con-" "-science" in terms of a mechanical spring.[47] Picture this spring sprung from one situation to another, one actual, and the other the projected end of some possible action. Stretched from end to end, A to B, there is tension inside the spring. This tension is the motivation to do the work of moving from A to B. The dimensions in terms of which this tension is realized are those by way of which the relative evaluation of the situations proceeds.[48] This is the picture of conscience as the motivational spring of action, easily translated into mechanical-energetic terms.[49]

Consider any given action. An agent begins in his current situation, first at rest at point A. Then, a need arises. The agent becomes uncomfortable in the current situation. Surveying available ends, the agent seeks some further, more satisfactory end, B. It evaluates available ends, weighing energy cost/benefit,[50] and stretches from points A to B, end to end. The tension within the stretched spring is the difference between discomfort and comfort, A and B. This

[46] Any agent, however will embody a native tendency to low-energy due to various purely metabolic/embodied processes, requiring no such experiential base for comparison. Of course, such metabolic baselines will very both between and within given agents, depending on periods of development, work/sleep cycles, and so on, but these issues fall well outside of the bounds of this chapter's discussion.

[47] Kant employs similar imagery.

[48] At the same time explaining the possibility of mistaken actions, as the dimensions of said assay may be inaccurate or incomplete for various reasons.

[49] Granted, in any realistic implementation, there must be two measures of tension brought into play, one in terms of the agent and one in terms of the world. However, in purely physical terms, this is easy enough – the terms of the world are given in natural laws of physics and chemistry. This is discussed in the paragraphs following. In terms of the metaphysical space of action in which human beings navigate, the translation is more complex, an account of which is developed in [42].

[50] This is an overt simplification, but these comparisons can be formalized in a number of ways, though I tend to think of them constructed from overlaying surfaces representative of a metaphysically determined action space.

difference is felt, though it can also be explicitly/symbolically represented, and provides the motivation for an agent to move from A to B.

When an agent reaches out for an end, attaches to it, and attempts to move there, the agent intends to reach that end. Here, "intension" is consistent with what can be said about agency generally.[51] It is "in-" "-tension," the internalized difference between relatively evaluated situations, both of which stand as ends of action, possible or otherwise. It is internal tension. This tension is the motivating force of the spring of conscience. When an agent is in a good situation, she is without tension. Then, when some need arises, a gap appears between this needful situation and another revealed first of all in the dimension of the most urgent need. The agent then exercises the tension of this difference as it moves to that end.[52]

7. THE ACTWITH MODEL

From the preceding, we have caught a glimpse at how a moral agent might begin to assess the morally relevant aspects of any other situation, its own or any other's. At the crux of this analysis is conscience. Thus, any model of moral agency seeking to capture the revealed aspects of moral cognition must model conscience, whether it is understood to be a model of conscience or not, and indeed in many ways agency altogether may be reducible to conscientious agency. The ACTWith model is, first of all, a model of conscience, and as such is a comprehensive model of moral agency.[53]

The ACTWith (As-if Coming-to-Terms-With) model is a situated, embodied and embedded [29][46][47] information processing framework inspired equally by hybrid neural network models and complex/dynamic

[51] The preceding is a computationally friendly interpretation of intentionality, and to distinguish it from the contemporary accounts bearing no resemblance, it is branded "intension" reflecting its deeper coherency with what can be said about agency, generally, outside of demands for theoretical consistency with given prepackaged interpretations. Moreover, this account is intended to reflect certain neurological features of agency, such as the anticipatory creation of action potentials in pre-motor networks when some sensed need for action arises, though I shall discuss these issues no more here.

[52] At first glance, this account may appear contrary to contemporary accounts of the term. Typically, theorists take "intention" to indicate the significance of an object. However, it is easy enough to derive the significance of any given object from how it stands in terms of projected ends of action toward stable/comfortable/low-energy situations. I will leave this discussion for another time.

[53] Something heretofore lacking, as noted in [13]. As for the question – Why not implement a comprehensive neurologically accurate model? The fact is that such is neither forthcoming nor computationally realizable if it were [45].

systems, tempered by observed successes and failures in related treatments from human psychology, neurology and traditional moral philosophy.[54] Here, the model is detailed, its theoretical origins and implications briefly reviewed, and its operations illustrated.[55]

The ACTWith model is at root a bottom-up hybrid architecture, originally informed by Ron Sun's CLARION architecture.[15][16] However, it is developed here as a model of control of information processing,[56] by its nature task and implementation non-specific. From human neural processing, the model proceeds from two key insights into organic mechanisms of moral cognition, disgust and mirroring.[48][49][50][51][52][53] It is essentially a model of situated cognition. Although developed without any particular theory or theorist in mind, it is consistent with work from situationist psychology [56][57], and represents a strong form of embodiment.[47] The ACTWith model is a four-step cycle composed of four operations. Two belong to a top (rational) level and two to a bottom (affective) level. This structure is captured in the name, "ACTWith."[57] "ACTWith" stands for "As-if" "Coming to Terms With." The "as-if" operations involve feeling a situation out, while the "coming to terms with" operations involve defining the situation in terms of the things originally felt. This is straightforwardly bottom-up hybrid in conception, intended to represent the bare minimum architecture providing for the eventual emergence of morality. The model consists in 4 modes:

As-if (closed) coming to terms with (closed)
As-if (open) coming to terms with (closed)
As-if (closed) coming to terms with (open)
As-if (open) coming to terms with (open)

[54] The literature is vast. On mirroring, see [48]. On disgust see [49]. On empathy see [50]. On embodied indicators of conscience as testable during child development, see [51]. On social cognition/social-cognitive neurosciences (mirroring and disgust) see [52] for theory and [53] for method. On the dual-aspect nature of the mind influencing the logic of the ACTWith model, see [54]. On psychological approaches to modify conscientious responses, see [55].

[55] [7] provides account of autonomous agency of the sort aimed at in the ACTWith model, distinguishing this with other approaches such as the control-theoretic approaches, specifically in the context of moral agency and with special attention to Kant's formulation of autonomy on pages 95-96. See these and the pages preceding for important insight into autopoesis, self-regulation, and identity that can enrich the understanding of related issues beyond the scope of present discussion.

[56] For relative advantages to this approach, see [58], though his control theoretic implementation differs dramatically from the self-regulatory, stability-seeking systems described here.

[57] "ACTWith" bears no deliberate relationship with the famous ACT-R model.

The closed modes are inspired by research into human neurology on the mechanism of disgust, while the open modes are equally inspired by research into mirror neural systems, both affective and action oriented.[58] Altogether, the four modes can be visualized as given in figure 1. (See figure 1 - "Basic ACTWith model consisting of four static modes.")

These modes, when fully articulated, capture the process expressed in Adam Smith's moral theory quoted above. The "as-if" involves affectively putting one's self into another's situation. It involves feeling as if this situation is one's own.

It has two basic modes, open or closed. One is either open to feeling as if one is in another situation, or one is closed to it. Openness is "compassion" and closedness is the lack thereof. Figuratively speaking, what is open in compassion is one's heart, and having an open heart is the first step in exercising one's conscience. Before returning to such traditional representations, and to Smith, however, it will pay to consider each of these modes in turn.

In order to illustrate the individual modes, it is useful to imagine that each represent a certain personality type which might arise through the habitual application of one of the four modes at the exclusion of the others.[59] Each represent personality types that are common, enough, to be easily recognized as archetypes in personal experience. First, consider the mode o/c.

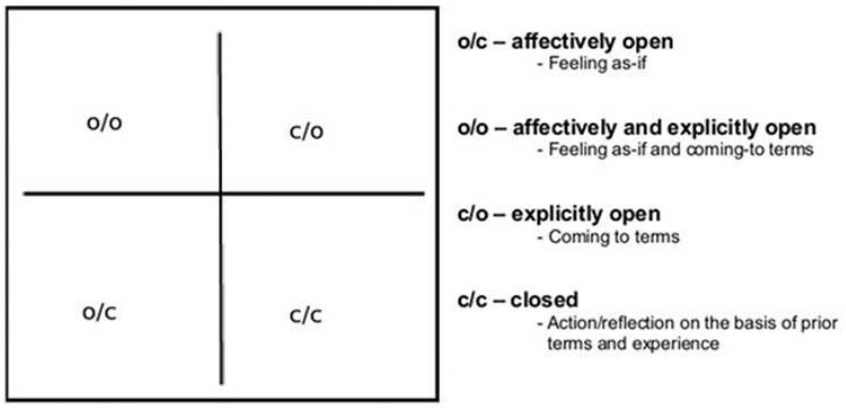

Figure 1. Basic ACTWith model consisting of four static modes.

[58] See also [54]

[59] Such would be a static, rigid, over-trained personality, and, though unrealistic, still useful in illustration.

This mode represents a personality affectively open to other situations, but not at the level of discrete reason. There is a certain sort of disconnect between thought and action in this personality. Such an agent, habitually engaged in this mode when dealing with others, would present genuine sympathy for the situations in which these others were finding themselves, but would only be capable of understanding the significance of those situations in light of his or her own prior understanding. The agent would not come to new terms with these other situations, and would rather feel them out only insofar as prior experience would allow. Contrast this mode with that of c/o. This personality is closed, affectively, but open at the level of discrete reason. If an agent were to habitually engage in this mode, he or she would not feel what it is like to be in another's situation, at all, but would be receptive to the terms of explanation, only. Imagine that these two personalities were policemen. The first would put emotions before law, letting people go who intended no harm in actions otherwise deemed unlawful. The second would show not such mercy, dutifully recording testimony, meanwhile retaining its fundamental maxim: the law is the law.

The o/o and the c/c modes are the most recognizable. The o/o mode is that of the genuine saint. This personality is affectively open to others' situations as well as genuinely interested in understanding what it is like to be in those situations. The habitual engagement of this mode is the object of many if not most religions, encouraged through the practice of certain religious rituals. It leads to the reverence for all sentience in nature on the Buddhist program, and to the society of Friends on the Christian, for example. The c/c mode is the opposite of the o/o mode. Agents demonstrating this mode are selfish, arrogant, and even psychopathic. Cold and calculating, this personality is perhaps most recognizable.

Any realistic model of agency cannot remain in a single mode of computation, but must be dynamic. The ACTWith model is, fully developed, a cycle of information processing built from the modes illustrated above. This cycle is illustrated in figure 2. (See figure 2 - "The Beating Heart of Conscience.")

This model is "the beating heart of conscience," recognizing the fact that the conscience has traditionally been associated with the beating of a heart, the seat of human compassion, of love, and of morality.[60] With the ACTWith

[60] And capitalizing on the input/output life-preserving dynamic common to both the human heart and to less complex organisms, such as the common bivalve. However, where the bivalve is effectively a slave to it external environment, being as it is rooted to a sea floor and capable only of feeding from what the tides bring, more complex organisms are able to seek out and

model dynamicized into the beating heart of conscience, we can recast Adam Smith's description of the process of moral cognition in ACTWith shorthand:

> By the imagination we place ourselves in his situation [O/C], we conceive ourselves enduring all the same torments [O/O], we enter as it were into his body[C/O], and become in some measure the same person with him [C/C], and thence form some idea of his sensations [O/C], and even feel something which, though weaker in degree, is not altogether unlike them[O/O]. His agonies, when they are thus brought home to ourselves [C/O], when we have thus adopted and made them our own [C/C], begin at last to affect us, and we then tremble and shudder at the thought of what he feels [O/C].[20]

This cycle is a normal process. Normal, at least for the agent not completely "hidebound by habit."

In order to further illustrate the effect of this cycle, it may serve to demonstrate two modes, these being with a conscience, "conscientious", or with a heart, and being "without a conscience" in everyday terms.[61]

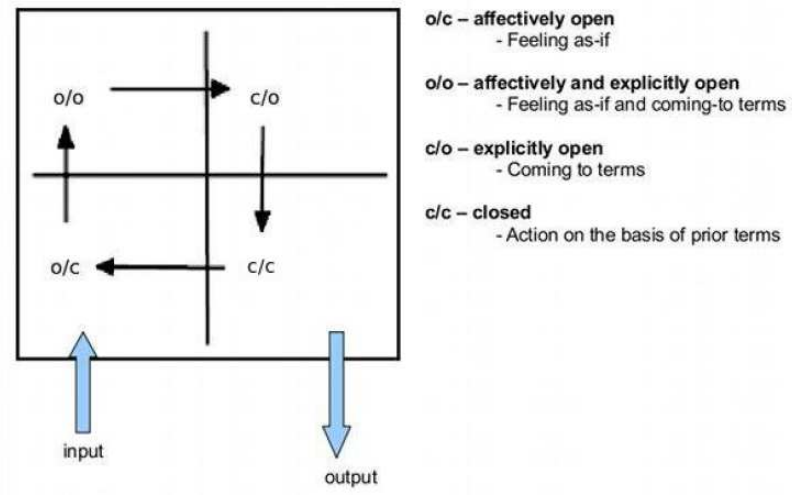

Figure 2. The Beating Heart of Conscience.

to avoid situations that are either beneficial or contrary to integrity (physical or otherwise) and survival.

[61] Strictly speaking, as the model suggests, one is never literally "without a conscience," one merely fails to employ certain modes of cognition at morally appropriate times, thereby demonstrating immoral or amoral behavior, while at once – through routine – become immoral or amoral by habit if not by reputation.

These modes are differentiable by the ways in which morally relevant information is processed by a moral agent. And, as this information so processed is then used as the basis for future action, future integration of morally relevant information, and future evaluation of situations, these modes of information processing effectively shape the moral agent, amounting to what in human beings is recognized as moral character.

Consider the following scene. A moral agent is making his way down an icy city street when he stumbles upon a man, dirty and disheveled and obviously very cold, sitting over a steaming man-hole cover. The man is wet from the steam. He is dressed in rags. In the bitter wind, the stinking vapor - his only source of heat - turns to ice in his ratty beard. Moreover, the man is apparently ill, with spots of pus dried from broken sores upon his windburned lips, and his feet are bloody through the ragged boots that hang over the side of the manhole cover into the dirty slush that rings it.

Upon realizing this scene, our agent may either open to the plight of the poor man, or close to it. Let's consider the open mode, first. In opening to the poor man, the agent – following Smith's description – will proceed through the steps of moral cognition in bringing the other situation "home" to himself. The agent will, perhaps, first have to overcome disgust in order to take up this "home" He will find affective cues as to the man's felt condition, and where these are absent, will have to fill in the blanks from his own prior experience. In this way, he will have to bring "home" to himself both the feeling of what it is to be in that man's current situation, as well as construct potential action paths both into and out of that current situation. In so doing, our agent is able to reconstruct within himself the tensions that the poor, ragged man feels, stuck as he is in a situation that - it would seem - no one would seek given any opportunity or understanding otherwise.

Let's look at the steps that the agent must take in greater detail. First, the agent will feel what it must be like to be in the other's situation in terms of the agent's own prior experience (o/c). Then, from this affective base, as the agent opens to the other in genuine compassion, the agent is amenable to coming to an understanding of the situation from the perspective of the other. Perhaps, he actively looks for symbols and signs, clues to the other's condition. Or, perhaps he simply asks about it. This is the mode of concern, again with "con-" playing its typical role, and "-cern" meaning the being of a cognitive agent altogether, in thought and in feeling, as the agent comes to appreciate the situation in terms of the other, perhaps through conversation, or through the careful study of the other's actions and expressions, whether momentarily or for a longer time (o/o). Then, as the situation sinks in, the agent experiences

the situation as if it were actually his own, extracting from this experience significant judgments and valuations (c/o).[62] Thusly, the agent is able to feel the difference between his own situation and that of the other, as the terms to which he has come are backfed into his own prior understanding, and he learns – perhaps – to appreciate the relative comfort of his own situation, to "count his blessings" as it were. This translocation of self-interest, from one's own situation into that of another, is the feeling of being "moved" in compassion, a natural byproduct of the ACTWith cycle. Finally, the agent will be able to reflect on his new experience, and either open once again to the situation, searching for greater understanding,[63] or act – perhaps by offering the poor man some charity – and move on to other situations, enriched for the new experience (c/c).

The closed case is effectively much easier to demonstrate. The agent, upon the sight of the man, closes to him in disgust, and during this cycle of processing opens instead to the agent's own future or past situations, perhaps reliving a trip to Disney World or imagining what it will be like to eat with a mistress. The agent simply walks by, unaffected by the difference between his own and the other's situation beyond feeling reinforced within his own. Though the cycle of cognition that is the beating heart of conscience proceeds uninterrupted, this agent can be said to be "without a conscience." Our moral agent has a heart only for its own self, failing to truly be a "moral" agent, at all.[64]

Being "without a conscience" is not all bad. In many ways, it is easy to see how the closed agent has certain advantages over the open agent. Especially in a world whose customs, largely shaped by latter-day corporate capitalism, favor those who act selfishly and without regard for the situations that others are left in due to one's own selfish actions, the closed mode has the advantage of delivering its habitual employer to positions of relative success and material wealth. Here, recall the extreme heartlessness of Scrooge the famous Dicken's story, *A Christmas Carol*, passing aside the sufferings of

[62] This is already a much deeper process, one of trading situations in a strong sense, than that represented by mindreading/simulation theorists.

[63] In purely computational terms, going through another cycle of processing, beginning from the agent's new found understanding, in the end reaching greater refinement of understanding through repeatedly generating and backfeeding error signals.

[64] Again, strictly speaking, a moral agent is never literally "without a conscience." However, it may fail to employ appropriate modes of cognition at appropriate times, thereby effecting immoral or amoral behavior, while at once becoming the embodied locus that results from this behavior. This raises the issue of psychopathy. It is a distinct advantage of the present account over others that it accounts for moral deviance by the same logic that it does moral excellence.

homeless children with the understanding that the world has too many people, already, and the loss of a few, on his mode of accounting, being a good thing. In so doing, Scrooge effectively negates those others' situations. They become null, and do not figure into his conscience, whatsoever, relieving him of any sense of obligation that might attach to persons not so nullified.

On the other hand, if one were to consider the negative effects on others of one's own everyday actions, eventually he may change his way of life in order to minimize the tensions between his own and the situations of those less fortunate. Recall Mill's words reviewed earlier, "... feeling that the interests of others are [one's] own interests."[17][65] The habitual employer of the open mode, thus, may become increasingly burdened. As greater numbers fall to desperate situations in the wake of the selfish stampede for "success," the felt tensions increasingly motivate the open agent to action, while the Scrooges of the world are revealed in their relative inaction.

In either case, both types of agents affect the world through their agency. And, as agents shape their environments through their actions, they exhibit a limited potential to determine for themselves both the environment with which they must find equilibrium, and the conditions that they must embody in order for this equilibrium to be achieved. As an agent shapes the world through action, it sets out the terms to which it must come in future iterations, and so on. Self and world, what one knows and does, are not only inseparable, but are increasingly bound together as the agent proceeds.

As the agent opens to the world, and comes to terms with it, that agent takes up the understanding of that situation, and proceeds from that understanding to the next situation, and so on. Thus, in opening and in closing to the world, the agent changes not only the world through action, but himself through the experience of that world. Through this process, thus, it becomes clear that agent and environment are two sides of the same coin, the situation, the ongoing integration of either pole of which – self and world - is illustrated in Figure 3, "Stitching One's Self into the World."[66] (See Figure 3 - "Stitching one's self into the world.")

Here, the relationship between conscience and freewill can be briefly clarified. As shown in the preceding figure, and as alluded to in the preceding

[65] Page 279.

[66] In the diagram on the left, the process of opening and closing to the world is given in ACTWith processing terms. In the diagram on the right, there is illustrated the potential for personal growth that is the promise of the habitually open mode, which leads to what existentialist have called the "beautiful soul" and that phenomenologists have called "authenticity".

discussion, the role of conscience in freedom is that it serves as the mechanism which makes the freedom of self-determination a real possibility.

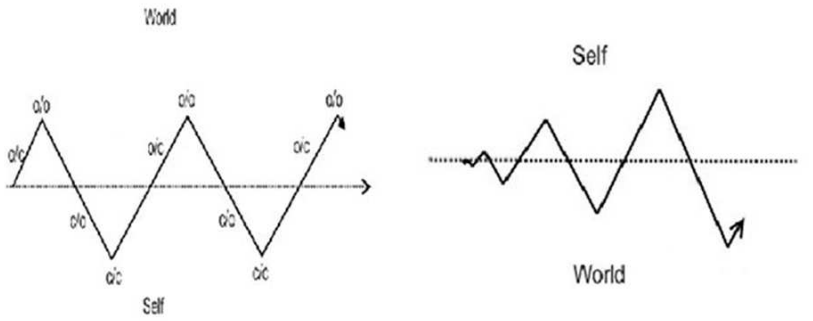

Figure 3. Stitching one's self into the world.

This is not radical freedom, the sort that permits an agent to perform any action willy-nilly without regard to past or prior constraints imposed by very real facts about the agents embodiment and its capacity to adapt to new and changing situations. It is not "divine" freedom. Instead, conscience is a steering wheel of sorts, a gentle handle on personal self-transformation – one is free to determine, in a limited way, who he or she will become through action, and in so doing determine how the word will be affected through the same.[67]

Next, we will turn to consider what sort of moral framework emerges from the model developed thus far. Already, we can see that morality has less to do with what action one takes in a given situation, and more to do with who one wishes to become as the product of his or her actions.

This leads to two sets of considerations. First off, considerations of the maximal expression and reinforcement of morally redeeming qualities falls under the heading of virtue theory. Some attention has been given to virtue theory as a suitable theoretical framework for AMA design. However, it is not so commonly familiar as are other approaches. Thus, this chapter shall remain focused on the most famous moral theorist in the contemporary West, Immanuel Kant. It is enough to note at this time that there is a direct relationship between virtue theory and Kantian moral theory, and that this relationship has been explored in regards to robotic implementations specifically, with the former being transcribed into the latter.[6]

[67] These issues are fully explored in [42].

8. MORAL AGENTS CREATE MORAL THEORIES, DON'T THEY?

In the adequately articulated AMA, the agent not only acts according to moral constraints, it sets these constraints for itself as well.[68] From the preceding discussion, we can see how an AMA might come to realize the value of certain situations over others. And, given the universal currency in terms of which this evaluation is to proceed, we can also see how an artificial instantiation of agency can still weigh ends belonging to human agents, and vice versa. We can see that an AMA so motivated might come to create moral principles as it goes, rather than be merely constrained by moral principles externally prescribed. Still, there is a question: Will the moral principles created by an AMA align with those given in traditional moral theory? And, in any event, how can either be understood in terms that facilitate their comparison in the first place?

The first task is to understand traditional moral theory in terms both internally self-consistent, and in terms consistent with those native to both AMAs and human moral agents. Consider Kant's moral theory in this light. According to some interpretations, the role of conscience in Kant's moral theory is merely that of the traditional voice of conscience. On this interpretation, conscience is merely a warning against immoral action, embodied in Kant's categorical imperative. We have already reviewed the failures of this characterization in this chapter, but the issue bears further discussion, now. On this picture, conscience is simply recast as a pre-theoretical understanding of the categorical imperative. The warning voice of conscience rises to awareness when an agent forms an intension that violates the categorical imperative.[38] Primacy is placed on the categorical imperative, not on the conscience, and the role of conscience is reduced to a sort of annoying beep. This leads easily to appropriations of Kant's portrait of human agency without conscience entering into the picture, at all.[69]

However, this rationalist reduction of morality to the conscious application of explicit principles is not without difficulties. Recall the trouble reviewed in the first section from employing Kant's categorical imperative in an AMA.

---

[68] This falls under the heading of "autonomy" in Kant, but might be better labeled "moral auto-poeity." For discussion of autopoeisis and autonomy, see [7].

[69] That being most appropriations, representing a so-called "standard interpretation," including [1][4][6][8][59].

The trouble there was that principles can conflict, driving the researchers to rethink the problem of artificial moral agency not in terms true to Kant's theory of human morality, but rather in terms of Kant's description of moral agency suited to divine beings. By creating moral "gods," the researchers hoped to evade two troubles with human-like agency: the potential to make mistakes, and the conflicts of principle that potentially forbid any action, at all. In the extreme form of divine agency, there is no need for conscience even as a warning beep.

Its role is completely obviated. As we have seen, however, this form of AMA is not adequate. It does not represent the ultimate goal of artificial intelligence research, as it does not qualify as a moral agent fully understood. We need a human-like form of agency, especially if AMAs are to be recognized as fully ethical agents, and receive the moral status due such entities, this discussion being the focus of the next section of this chapter.

In 1974, a philosopher, Laszlo Versenyi, anticipated the likelihood that researchers into AMAs would find themselves in a similar position. On his forecast, this is where computational models of moral agency would begin, with agents of a divine form, unable to make mistakes, perfect vehicles for universal moral law.

He labeled them "holy robots." The presumption was that "holy" robots would not be burdened with human traits leading to selfish – or compassionate – distortions of moral duty as prescribed by Kant's categorical imperative. In order to recast this perfect moral executor in more human form, Versenyi writes:

> Should the programmer wish to change our holy robots into something more like imperfectly moral human beings, he would merely have to make their categorical programming somewhat less than absolutely overriding. Whatever way he does this (e.g., by a fixed percentage method or a sufficiently complicated built-in randomizing procedure) is irrelevant for our purposes. Since Kant himself cannot explain by what mechanism the causality of pure reason becomes (if and when it does become) decisive in human action, he could not say categorically that our robots' mechanism is different from that of men."[6][70]

However, I disagree with Versenyi on two points. First, Kant may not have been able to explain how pure reason yields human action. But, Kant is in no unique position, here. Neither has anyone else. The present account, as well

[70] Page 251.

as many others including from neurology and psychology directly, are attempts at solutions, and as inadequate as any may be in definitively patterning the cognitive mechanisms "decisive" in human action, there remains a sure test at the gross level of agency. If the mechanisms in question might in principle distill universal moral law from experience – just as Kant himself had done - then such a mechanism is prima facie worthy of moral consideration.[71] Secondly, and in remedy of the first point, Versenyi seems to fail to appreciate the central role of conscience in Kantian moral theory, and so misses one likely candidate for the mechanism of morality before it is ever investigated. Kant may not have been able to explain the cognitive processes at play in human action, but that doesn't mean that he didn't try. In Kant's moral theory, fully explored, conscience plays a central role not merely in deliberation over action, but in the process of becoming a moral agent. Though it may be true that Kant did not explicitly[72] detail the conscience in that role consistently throughout the exposition of his moral theory, this assignment does exist. The proof, then, of the adequacy of Kant's proposed mechanism of cognition is the same as any other: that it can, would, and perhaps should eventuate in the production of regulatory principles consistent with those given by Kant, himself.

The ACTWith model provides a medium through which such a test can be performed. As fundamentally a model of conscience, the ACTWith model is not only regulated by Kantian principle, but embodies it in a way specific to fully ethical AMAs, the form that we are after. The adequacy of the ACTWith model as a framework for moral agency is demonstrated in the direct emergence, from its normal function, of what may be considered the first fundamental law of morality. This law is variously understood. In the West, it is recognized as the "Golden Rule." This rule is formalized in Kant's "categorical imperative," and expresses a core moral component - if not *the* core moral component - of every world religion. From the ACTWith model, we will find a form of this law derived in situationist terms consistent with the mechanism of morality as described therein, and at once universal to both human and artificial moral agency. In fact, consistent with Kant's theory, too.

Let us first consider what Kant means by morality. In *The Metaphysics of Ethics*, Kant tells us that "... although it is no direct duty to take a part in the

[71] Discussion in and on Kant on this issue centers on the concept of autonomy, a term the implications of which in the present context of artificial agency is fruitfully reviewed in [7]

[72] At least in clear enough language for the philosophical community, which – for all its assumed patience and vision – shows a suspect impatience when confronted with the task of constructively interpreting the organic life's work of a mind such as Kant's.

joy or grief of others, yet to take an active part in their lot is ...” [60][73] This recalls Mill's characterization, and Smith's as well. Further, he tells us that we ought not “... avoid the receptacles of the poor, in order to save ourselves an unpleasant feeling, but rather to seek them out.” Seek them out to perform, we may imagine, the processes of moral sentiment as described by Smith. As well, we ought not

> ... desert the chambers of the sick nor the cells of the debtor, in order to escape the painful sympathy we might be unable to repress, this emotion being a spring implanted in us by nature, prompting to the discharge of duties, which the naked representations of reason might be unable to accomplish.[60][74]

In the moral man, the heart naturally goes out to others less fortunate.[75] Most importantly, from these passages, and decidedly contrary to rationalist interpretations of Kantian theory, we can clearly see that Kant paints a picture of an *affectively* motivated moral agent. He describes an agent motivated by the “spring” of practical reason to overcome unpleasant feelings of compassion and sympathy. This spring is conscience.

Here, we must be careful not to mistake conscience for a mere warning light, the pre-theoretic call of the categorical imperative, itself understood as a purely rational directive. Instead, the motivational spring of practical reason instead must be understood in terms of an aspect of Kant's theory more difficult to encapsulate than the categorical imperative, and so as often neglected, the “good will.”

Moral affect is central to Kant’s moral theory in the form of the good will. What is good will? In the first section of *The Metaphysics of Ethics*, Kant tells us that the good will is “to be considered, not the only and whole good, but as the highest good, and the condition limiting every other good, even happiness ...” [60][76] In this light, recall our considerations of Mill's *Utilitarianism*. This ultimate limiting condition is the role played by similar feelings as aspects of conscience on Mill's account. Furthermore, in the second section, Kant writes “That, we now know, is a good will whose maxim, if made law universal, would not be repugnant to itself.”[60][77] Here, it is important to note that repugnance is another word for disgust – specifically,

[73] Page 167.
[74] Page 167.
[75] Equally reflecting the purpose of Mill's *Utilitarianism*.
[76] Page 16.
[77] Page 39.

self-disgust - neither concepts belonging to reason, where typical misinterpretations of Kantian ethics place the locus of moral motivation, to avoid rationally drawn contradictions between explicitly given maxims of action. Instead, Kant draws a specific form of contradiction, and it is decidedly not of the form of two explicit principles, or maxims, drawn into purely symbolic contradiction. This is an *embodied* contradiction, in that one *self* cannot perform some action resulting in that *self* becoming *self*-repugnant through said action. Otherwise, action proceeds in good will, towards ends in which a self remains in harmony with itself. Thus, we have a fundamentally affectively grounded, indeed bottom-up interpretation of Kantian moral theory in light of which Kant, himself, is rendered more internally self-consistent and in terms of which artificial and human moral agency may likewise be consistently assessed.[78]

Still, there is more to this story. By Kant's account, goodwill alone is not enough. It is not enough to merely become through action one in whom one is not disgusted. An agent must excel.[79] An agent must also have in mind some exemplar according to the demonstrations of which it might pattern its own initial behaviors, and so from this experience gain insights that aid in the eventual realization of universal moral law. This moral exemplar may be either real or ideal, in either case providing information through which an agent may, at least initially, model its actions, and thus eventually its self. According to Kant, a very special affect attaches to these moral exemplars, "reverence." Moreover, objects of reverence serve not only as guides for moral action, but in their direct comparison with one's self and others, they serve as the measuring stick of moral worth.

How are we to understand reverence in terms appropriate for articulation in the design and construction of AMAs? Where is the cash value, here? It might appear that Versenyi has a point, in that Kant cannot explain how such a mechanism actually works. However, we can charitably fill in some blanks for the old man, given that he was working out the necessary form of the solution to this great problem of moral agency 200 years before modern psychology, fMRIs, social-cognitive neuroscience, and all the rest which themselves, ultimately, owe their present inquiries to groundwork laid by Kant, himself.[80]

---

[78] Further proof of the affective grounds of Kantian moral theory can be found in the notes to chapter 1, where Kant tells us that "What is called a moral interest, is based solely on this emotion."(page 190) The emotion in question here is reverence, the focus of following pages.

[79] This is to open important bridges between Kantian moral theory and theories of virtue.

[80] However reworked it has been through Hegel, Nietzsche, Heidegger, James, and others in the meantime.

Reverence can be characterized in contemporary terms as involving the employment of mirroring capacities of the human body to emulate, imitate, model and so – through witnessed demonstrations of moral action - train one's self to adopt certain modes of being, thus becoming like some exemplary moral agent, whether real or ideal. In terms of the ACTWith model described above, the exemplar and the self are two situations to be held in comparison, the tensions between which serving as motivation to move from one to the other along the dimensions revealed. Any given moral agent's ultimate interest, thus, is not fundamentally different than is the ultimate interest of any other agent in any other space of action, whether that be Chemistry or charity - to become the best agent in that field that it can become. The crux, here, is that this model of excellence may be initially taken from the embodied and demonstrated actions of some other agent.

Furthermore, in the two concepts, reverence and goodwill, we can plot the opening and closing functions of the ACTWith model. An agent either opens to another situation, or closes to it. In instances where emulation of demonstrated modes of being might lead to lower-energy states,[81] both physically - and in the case of human being - metaphysically understood, then these modes are taken up and expressed through the agent's own present and future actions.

Finally, from this discussion, it is obvious that the exposure of an agent to suitable moral exemplars becomes a limiting condition not only on what an agent will do, but most importantly becomes a limiting condition on what that agent will become. Without suitable moral exemplars, an agent is left alone in a seemingly infinite space of possible action. Doing the right thing is a crap shoot, and the computational hurdle involved in plotting a moral self-portrait is insurmountable.

With the benefit of information derived from the demonstrations of moral exemplars, an agent's understanding of what is means to be a moral agent is enriched, and thus the potential to become a moral agent, itself, is opened up.

In a section of the *Metaphysics of Ethics* interestingly entitled "Prerequisites towards constituting man a moral agent,"[82] Kant affirms the limiting factor not only of moral affect, but of understanding. In the end,

[81] Given in terms of comfort, but complicated by issues of opportunity, the metaphysical rather than purely physical dimensions of situational analysis, and so on - issues beyond the scope of present review, but subject of [42].

[82] Any mention of which is notably absent from any works in this field covering Kant's moral theory, including Allen et.al.'s presumptively titled "Prolegomena..." and obviating Allen's end-run to divine agency, I might add.

practically, theoretically, an agent's understanding is the limit whereby it can determine right or wrong. Kant writes that "... obligement can extend only to the illumination of his understanding as to what things are duty, what not."[60][83] We cannot, thus, expect morality from an agent which is not permitted, through accident or through design, to appropriate the terms necessary for universal moral evaluation. This conclusion provides support for our insistence on human-like AMAs, conscientiously endowed, proceeding on the basis of universal terms of agency, briefly articulated in this chapter as energetic terms, and fully benefitted by the richest possible source of moral information, that from similarly embodied moral exemplars. This is a portrait not unlike that of the proper moral development of a human moral agent.

Before concluding the present analysis, let's first consider some possible objections to the preceding account. Some may object that we have given short-shrift to the notion of moral duty in Kantian theory as a rational notion, not affective at all. In the face of such an objection, we must only recall the source of the motivation that leads to the production of moral duty on the Kantian account. However much weight can be given to the notion of duty, more must be given to the processes that produce it. For this motivation, we need look no further than what Kant called this "spring implanted in us by nature" that motivates people – Kant considered only human beings, but expanded in the present context to include artificial entities, and perhaps other forms as well - to seek to fulfill moral duty. Both the duty that is attached to action, and the spring that motivates an agent to become the best agent it can be – to become worthy of reverence through its own exhibition of goodwill - are the propriety of conscience. Kant writes:

> The only duty there is here room for, is to cultivate one's conscience, and to quicken the attention due to the voice of a man's inward monitor, and to strain every exertion (i.e., indirectly a duty) to procure obedience to what he says.[60][84]

"He," here, is conscience. Thus understood, conscience is not a pre-theoretic expression of the categorical imperative; the categorical imperative is a post-theoretic expression of the mechanism of morality, conscience. In other words, moral agency is conscientious agency, and an agent's sole moral duty - "the only duty there is room for" - is to maximize conscientiousness.

[83] Page 132.
[84] Page 132.

From the preceding description of the ACTWith model, we can see how an agent might do this. By engaging others in an open mode, as a matter of habit, thereby maximizing understanding, and so expanding the potential to discover and satisfy obligations to self and others, an agent fulfills its moral duty. This is not a one-off action. It is a cycle. AMAs, thus, cannot be disposable tools, means for human pleasure, or even regarded, when adequately realized, as *less* than human, at all. Not if we human beings are to retain our presumed status of moral superiors, and especially not if we expect AMAs to be moral agents, at all.[85]

Finally, in consideration that, at least initially, it is we who will serve as moral exemplars for the emulation of AMAs everywhere, it may pay to attend to the responsibility that comes by way of this role. Shortly after the preceding statement, Kant spells out this duty for conscientious moral agents when passively serving as models and guides for others. He paints this picture according to the same logic of disgust and mirroring articulated in the ACTWith model:

> The compunction a man feels from the stings of conscience is, although of ethical origin, yet physical in its results, just like grief, fear, and every other sickly habitude of mind. To take heed, that no one fall under his own contempt, cannot indeed be my duty, for that exclusively is his concern. However, I ought to do nothing which I know may, from the constitution of our nature, become a temptation, seducing others to deeds which conscience may afterwards condemn them for. [60][86]

This nature is conscientious, with each serving as moral example for every other, with each modeling and mimicking demonstrated actions when seemingly appropriate, with each setting and reinforcing given standards of action. What one agent does indirectly affects other agents, true, but it directly affects who those other agents become. This social mirroring of one another is the root of Kantian moral law, as given in the categorical imperative. And, so understood, underscores our roles in whatever becomes of AMAs.

---

[85] It is also worth noting that Kant equates one's giving oneself over to these emotions with freewill, in short because such opens the potential for one's becoming the best person one can become, and such a result is, on his understanding, the universal aim of every person. Thus, satisfying these goals, an AMA should qualify as a person. And, adding fuel to this fire, it is easy to see that this process of expansive moral self-identification through compassion leads directly to the "beautiful soul," raising important questions of the nature of AMAs – *Do they have "souls?"*

[86] Page 127.

In this passage, we are confronted with a number of realizations. First off, we cannot expect an AMA to act morally, if we cannot demonstrate moral agency, ourselves. Secondly, should we fail in producing AMAs in the form of which we are now aiming, it may be due equally to our own moral incapacities as much or more than to our technological incapacities. Lastly, for all of those interested in AMAs as tools of war, as means for pleasure, and as dutiful and even god-like slaves, we must ask this: Is the act of designing, constructing, and living beside a race of creatures, solely for one's own selfish ends, an act that we can endorse for any other set of creatures besides ourselves? What if the robots were to do similarly to us, manipulating dating patterns and social rituals to ensure the production of a perfectly servile, docile, selfless race without goals of its own except those given by others? If we proceed in the development of pleasure-bot robo-killer AMAs, what will we have made of ourselves in the process? What if, someday, robots become the moral superiors? By the extension of their amassed experience with human beings, what should they do with us?

Regardless of our answers to these questions, in the end, we are left with a portrait of moral theory which pushes the issue of the moral status of AMAs. This portrait can be understood as a direct extension of the mechanisms at work in the ACTWith model. Not only should an ACTWith motivated agent proceed according to moral principle, it should equally produce, refine, and in practice become an agent embodying these principles, so demonstrating moral agency for the benefit of other moral agents, and ultimately for the benefit of the world, as a whole. And this is exactly the sort of moral agency that we have been aiming for.

Accordingly, it serves to recast the categorical imperative in light of these results. Arguably, the most famous form of the categorical imperative is "Act according to that maxim which thou couldst at the same time will an universal law."[60][87] There are other forms, however their direct consideration is presently unnecessary. With the situationist account of conscience applied to the preceding account of Kant's moral theory, this imperative can be rewritten in the following forms:

> Do not become through action (or inaction) an object of self-disgust.
> And, conversely: Do become through action (or inaction) an object of reverence.
> And, most simply: Do not put another into a situation that you would not seek for your own. [41]

[87] Page 39.

From the preceding account of conscience, should an agent so motivated take up a situation in the moral community, then, in terms of the energetic basis for the relative analysis of any given situation, these forms of universal moral law should emerge as a matter of general function.

Here, we may recall an illustration given earlier, and consider what sort of rule would emerge from an agent drawn according to the ACTWith account. Until that time, we are in a position not dissimilar to Kant's own. Recall the illustration involving two men, one our intrepid moral agent, and two the poor, cold, sick man warming himself on a street grate in the filthy snow. In that story, the agent may open to the evaluation of the less fortunate situation, realizing through the revealed differences that that other situation is not one worth seeking. Opening to that situation, our conscientiously motivated AMA may also begin to understand not only differences in their relative situations, but also differences in their relative agencies, in their selves. Through this comparison, our AMA may discover modes of action worth embodying, and modes to be avoided.[88] And, it will itself become different by way of this information. In fact, it may become better. Merely repeat this scene and permutations of it many times, and a picture of the moral education of an embodied AMA by the present design presents itself. In so doing, as the information taken from these iterations is processed, and from this information rules are extracted – on a hybrid model – one should arrive at a generalized statement of the amassed experience in the form of – although likely not in the language of - the restatements of universal law, above. And, finally, through this self-directed transformation from moral infant to a self-confirmed understanding of universal moral law, gain an increasingly well-realized sense of its own unique self.[89]

Now, this assertion is weighty, and there is really only one way to test it – we must build it.

Thus, again, we find ourselves in a position not dissimilar from Kant's, unable to fully explain the mechanisms that we have at the same time posited. Though we may have advanced the discussion by offering a rudimentary model, we cannot answer to its eventual efficacy.

---

[88] After all, even a poor man – especially a poor man - can be worthy of reverence!

[89] Resulting in, effectively, a virtue-motivated agent, for instance recommended for hybrid implementations in [13]. But, this does not imply a break with the account given thus far; for example, Versenyi [6] shows how Kantian/Millian agents can be understood in virtue-theoretic terms. However, the present account does break with Wallach's further assumptions, that moral reasoning and moral judgements should proceed from subsystems specialized to the tasks, as described in [13], page 248.

We cannot be certain that AMAs constructed consistently with the current model will achieve a unique personal identity. Confirmation can come only through implementation, confirmation that we can only hope to deliver in less than the 200 years it may have taken to get this far.

9. WHAT COUNTS AS A MORAL AGENT?

Given that we can create AMAs of the form in question – fully ethical agents – there remains the question of their moral status. Once constructed, an artificial moral agent must be recognized as such in order for it to become a member of the moral community. The requirements for an artificial moral agent to be considered a "moral agent," the essential equivalent of any human instance of moral agency, have been the subject of some attention.[90]

There are effectively two sources of this status, from within and from without the agent, itself.[91] With internal considerations in mind, to the question "What does it mean to be a moral agent?" Allen et.al. add:

> A suitably generic characterization might be that a moral agent is an individual who takes into consideration the interests of others rather than acting solely to advance his, her, or its (henceforth its) self- interest. [8][92]

In this regard, Allen's stipulation recalls Mill's characterization of the fusion of interests being a natural extension of inclination common to persons as essentially social, and essentially moral, critters.

---

[90] The issue is complex, and though gaining increasing specific attention, has been difficult for many commentators to ignore, as it is addressed in various ways in [1],[3][6][8],[7],[10],[11],[12],[13],[14],[25],[26],[61],[62],[63],[64],[65],[66],[67]. Some results worthy of brief comment and not otherwise covered herein include, interestingly, [67], representing the uncommon position that it is too early to begin talking about artificial moral agents at all. While Sparkes [65] (like most others) feels the need to explore these issues with some urgency. Bostrom [12] denies any room for doubt in the emergence of human-like, independent artificial agents in the first half of the 21st century. Sparkes writes that "Society is going to face problems adapting to their [AMA] integration into society, and it will start to face them soon."(page 10) Weng et.al [14] advise laying legal groundwork in anticipation thereof. Tamatea [66] assesses differences in attitudes to the possibility of AMAs being granted moral status, even as partners in love relationships with humans, based on religious differences, finding a common ground for (largely negative) evaluation between Buddhist and Christian respondents in the concept of the "unique" human "self."

[91] Though, they are not always so labeled. For example, Coeckelbergh [63] employs the distinction 'indirect' and 'direct.'

[92] Page 252.

Also in terms of internal criteria, Pollock offers an interesting measure of human-like agency. Using the concept 'person,' Pollock writes that:

> ... the concept of a person must simply be the concept of a thing having states that can be mapped onto our own in such a way that if we suppose the corresponding states to be the same, then the thing is for the most part rational.[68][93]

Pollock's recipe for personhood is one of epistemic rather than moral agency, but his emphasis on "human rational architecture" with states "mapped onto our own" bears interesting correlations with the approach to the creation of moral agents taken in this chapter.[94] Throughout, we have kept to a certain vision of an AMA that also emphasizes – so far as internal considerations are concerned – an architecture of information processing intended to model the mechanism of human morality, indeed map AMA processing onto our own and vice-versa.[95]

In bridging the self and artificial other in terms of mirrored architectures, we come to the limit of internal criteria and begin to talk about external criteria. From here, we can see our way clear to articulate external conditions for moral recognition. This avenue proceeds through a decidedly Kantian portrait of moral agency, in which internal requirements can be reduced to a single concept – autonomy. A moral agent must be a full fledged agent: self-legislating, productive of its own guiding principles, with a sense of self leading to certain aspirations, to be the best self that it can be. Ideally speaking, it must identify the best self that it can be, aiming through the application of self-assumed moral principles for a best-possible living environment for itself and others, thus setting a moral example for other like-minded agents to emulate, even revere. Should any agent exemplify these conditions, then without contest it qualifies not only as moral agent, but moral

[93] Page 462.

[94] Same vein, different note, Bolstrom [11] – focusing on intelligence rather than moral agency – focuses as well on architecture, and suggests that "The cognitive architecture of an artificial intellect may also be quite unlike that of humans."(page 4). We have seen how this assumption may fail to deliver on a moral agent throughout this chapter, however, with implications for the moral status of even the sort of "super-intelligence" that is Bolstrom's focus.

[95] Kahn et.al. [10] take a slightly different approach in offering a checklist of psychological benchmarks intended to assess the degree to which an artificial agent might qualify as 'human,' noting at the same time that such an inquiry promises to equally reveal essential aspects of human psychology, to this point perhaps hidden, analogizing human-robot comparative psychology to human-(other)animal comparative psychology.

exemplar, in such case cementing any claim to moral status in the moral community at large. The ACTWith model is designed to meet these conditions.

In exemplifying moral excellence in such a way that others not only recognize the agent as acting morally, but in such a way that others emulate said agent, we begin to see moral status as the product of social relationships external to agents, themselves. Moral status as such derives from the moral community of which the agent in question is a (potential) member, from how the features of the agent are "experienced by us." From this perspective:

> ...moral consideration is no longer seen as being 'intrinsic' to the entity: instead it is seen as something that is 'extrinsic': it is attributed to entities within social relations and within a social context. ... The implication is that both the human and the robot are not so much considered as atomistic individuals or members of a 'species', but as relational entities whose identity depends on their relations with other entities. [63][96]

This formula is not enough on its own, however, leaving open the possibility that we personify artificial agents too much. In such a case, we need not wait for new technology, for advances in the methods of designing and educating AMAs, or for a revolution in the way that morality, itself, is understood. We are experiencing this sort of impact from already existent artificial agents, now.

Recall Moor's classification of "ethical impact agents" introduced in the first section of this chapter. When read through the lens of the human "intentional stance" toward AMAs, it is clear that morally significant "social relations" within a "social context" already exist, even though the AMAs in question are not in the true sense "moral" agents. Robots have replaced many workers from factory floors, from libraries and cafes, and from other positions that had until recently served as productive niches for many people within a social system composed of and maintained solely through the labor of natural, rather than artificial, agents. As the costs for robot-workers has decreased, corporate managers have been able to maximize corporate profits through the exclusion of natural labor, in preference of more reliable, often more productive automated replacements. Such robotic replacements have already had a deep ethical impact on some areas of society, as formerly employed persons have had to suffer the indignation of losing their livelihoods to mere machines.

[96] Pages 214-215.

Displacement of human beings from the workforce is perhaps the most apparent ethical impact of (less-than-fully-ethical) AMAs thus far, but it promises to be the least important. The most important changes that promise to take place in society due to increasing inclusion of artificial agents may proceed unperceived. These are not changes to any production system, factory floor, or even economic bottom line. These are changes in the human beings, themselves, in their attitudes towards themselves, their world, and even in their capacities to care for each other, to serve as moral agents, in the first place. Consider the possibility of a worker who might despise robots, but who at the same time aspires to embody a similar efficiency of movement, a similar cold motivation, and a similar detachment from negative social repercussions due to its actions. This is not an eventuality that is easily planned for, yet carries serious moral implications.

Consider also in this light the impact of AMAs replacing human care-givers in hospital, hospice, and even home environments. Here, given the same economic incentives that have pushed human workers out of other productive positions, under the directive of insurance companies and for-profit health-care consortiums, "automated" healthcare providers might push human beings "out of the loop" of care-giving, altogether. [69][97] This potential self-imposed alienation from health care concerns also carries serious moral implications. For one thing, so far as they are commonly conceived, robots have no conscientious compulsion to fight for a patient's well-being, and no conscientious obstacle to denying apparently "futile" care, while human caregivers often continue to struggle against purely financial policies denying the former and encouraging the latter practice. Moreover, should the practice gain social acceptance, and the life or death decisions of automated care-givers determine when human life is worth living, or not, human beings – after a long enough period - may become terminally distracted from questions of the value of human life, leaving such to the profit/loss calculus programmed into robotic nurses by entities equally without conscience, their corporate slave-masters.[98] Borenstein notes that:

> The more reliable we think automated systems are, the more likely it is our attention will stray. What complicates matters is that this type of

[97] Perri [70] provides some interesting grounds for doubts over the efficacy of such automated care-givers. See pages 207-8.

[98] One may argue that the entity without conscience includes every individual who permits such mistreatment, rather than the fictional "person" that is the modern corporation. After all, who is actually responsible? In the end, only ourselves.

> behavioral shift might not be consciously detected. Hence, it would be wise to temper the confidence that users place in robots and other automated systems, especially when people could be significantly harmed. [69][99]

His suggestion is that "risks" be made explicit, presumably so that we do not lose sight of what is ethically important – the welfare of other human beings. This is easier said than done. Corporations in control of health care already are in the business of issuing policy prescriptions that can easily be understood as "automated systems." And, the risks of such systems are far from transparent. Indeed, every effort is made to keep ethically controversial practices hidden. Given this trend, it is difficult to accept that the incumbent risks introduced by robotic care-givers would be made any more transparent.

At root of these concerns is the potential for AMAs to shape not only our social and economic systems, but the persons who live within them - Us. Embedded in the issue of the automated care-giver is the fact that such entities are no longer tools - means to human ends - and have taken a place in the world – howsoever attained – in which their actions are given equal or greater moral weight than the human beings who live with them. It is a series of small steps from robotic worker to robotic companion, to robotic boss and robotic executioner. Finally if we, as a society, encourage a chain of events that results in a moral world not simply populated by but *determined by* AMAs, then whatever we make of ourselves and our world is our shared responsibility, altogether.[100]

Granted that this is our object presents a prima facie case – once again – for AMAs as close to human in aim and interest as possible, which – once again – underscores the need for clear computationally friendly articulations of *human* moral agency. Moreover, that this is the potential end toward which we are, altogether, headed, puts special emphasis on the status of AMAs, not as legal or even as potentially lethal entities, but as parts of our societies, of our lives and our deaths – as parts of "us".

From this perspective, one especially interesting approach to the question of the externally derived moral status of AMAs bears consideration. AMAs may deserve ethical treatment, not due to their constitution, consciousness, or

[99] Page 31. Even to the extent that artificial intelligences should guide human discourses on such topics as morality and ethics. See Danielson [71], for example.

[100] In making the case for social-wide, distributed responsibility for (at least some) actions (such as those with social-wide consequences requiring social-wide endorsement) in terms of extended cognition, generally, without a focus on AMAs, Mason Cash [62] comes to a similar conclusion.

other property, but due the human beings who extend AMAs such consideration in the first place. This position is advanced by David Levy:

> My own argument in support of giving certain rights to robots is not that robots with consciousness should have rights *because* of that consciousness *per se*, but that, because they have consciousness, such robots will be regarded by us in some similar ways to those in which we regard other humans, for example developing affection and even love for robots, and that, *because* we will regard such robots with affection and even love, it is reasonable to assume that we will treat robots in other ways similar to those we currently reserve for humans (and, in the case of some people, to pet animals), for example by regarding these robots as having rights. [25][101]

At one root of Levy's assertion is the consideration that to incorporate robots into one's daily life and times is to *expand* one's self in terms of this presence.[102] This is to say that as one comes to rely on robotic workers, assistants, aids and even companions in performing daily life tasks, these machines become part of one's extended self. This view, often referred to as the "extended mind" thesis in contemporary philosophy of mind, yet with roots in Kant and Aristotle and indeed demonstrated clearly in Plato's *Meno,* holds that those artifacts that persons employ in everyday routines – everything from things that serve to literally record information, as does a notepad serve as a repository for memory, to entire situations, themselves, as does a lecture room serve to prompt the recall of certain information (students' names, assignment dates, and so on) - both prompt and make possible the execution of entire subsets of context dependent behavior.[103] In the case of artificial agents, companions, friends and even lovers, the implications from the extended mind thesis extend throughout the range of human life.[104] In other words, as these robotic counterparts become, effectively, parts of one's self, they are deserving of similar treatment, and should be regarded with like respect. They should be treated as one would wish one's self treated.

---

[101] Page 214.

[102] An enriching introduction to these issues in terms solely of human agents in human society appears in Sutton [72]. Given the fully-ethical AMA is the form of agency in question, there is no reason that Sutton's considerations do not equally apply in the present context.

[103] Perhaps infusing an otherwise unimposing man outside of a lecture hall with aspects of character contributing to a resounding and authoritative presence in the lecture hall.

[104] And, there are ethical issues to deal with every step along the way. For instance, regarding the role of artificial companions in remembering details about a human being's life, see [73]. And, the tangles get even thicker when we consider hybrid organic/artificial agents [3].

This brings us to a second root of Levy's assertion, that which Levy himself finds most compelling. Those behaviors and attitudes that persons demonstrate towards robots – increasingly humanlike, even if they do indeed remain mere artifacts, which in my mind is unlikely – can be emulated and then transferred, not only to the treatment of other robots and other artifacts, but, due to the human-like nature of robots, to other human beings themselves. Such a concern brings with it two notions that are at the heart of the work at hand. One is the implicitly recognized capacity, in fact necessity, for human beings to mirror the actions of other human beings, and indeed other animate objects of any form. Especially in the case of children, this is a crucial, neurologically instantiated, aspect of developing moral agency, an aspect briefly noted in the chapter preceding. Moreover, the concern for the emulation of and transference of immoral actions by others upon their demonstration underscores another essential aspect of the moral agent, again resonating with the chapter preceding. Where typical approaches have been to add an "ethical algorithm" or a set of limiting conditions – principles or laws – to a given model, these observations bring to light the fact that the morality of the agent must stem from the very architecture of the agent itself. Everything is moral.

Morality begins with the ways in which information is processed, and indeed, in the strong sense implicitly revealed through Levy's concerns, the way in which an agent is "in-formed." The sorts of things that a moral agent is exposed to literally "in-" "-form" the agent; they form its insides. They shape what the agent is, does, and becomes through action, from the outside in. Thus understood, we are returned to the lesson taken from the opening section of this chapter.

We are the essential limiting condition on artificial agency. Regardless of the technology that we can bring to bear, the world that we make for ourselves is the world with which we must come to terms. Whatever we make of artificial moral agents, we make of ourselves, and only if we make ourselves (and our world) the best that we (and it) can be, can we expect the AMAs that we produce to succeed similarly.[105]

In the end, the way that we regard robots reflects the sorts of persons we shall become.

[105] One issue that deserves mention, but that demands more attention that we can spare, is that of the safety of robots as they enter into the society, issues of liability and legal status, and so many other dimensions of social/moral agency that enter into considerations of moral status. These are central to [14].

The questions of robot development – how, where, when, to what ends – are finally answered from the same field of resources from which they arise – ourselves. Specifically, from our capacity to develop for our own good, our own collective good, to become the best possible persons that we can become. In Versenyi's words, almost 40 years ago:

> Virtuous action on our part is action by virtue of which we maximize our own well-being. Consequently, "right" behavior toward robots would be behavior that would lead to the type of man/machine interaction that would be most beneficial to men. [6][106]

And, what sort of interaction with AMAs would prove most beneficial to men? That we build them in the first place. Human-like, fully ethical, autonomous, artificial moral agents. We build these not as killing machines, or as slaves or selfless sources of pleasure.

We build them - we *must* build them - to be the best of ourselves, just as we would wish any of our creations to be the best of ourselves in so far as those things are able to express that excellence. That we *will* build moral robots, that we *will* construct moral mirrors of ourselves, is more than a technical hurdle.

It is in fact the fulfillment of our selves as builders, as technological beings. It is the fulfillment of our selves as philosophical beings. It *is* the best of ourselves. Again, in the words of Laszlo Versenyi:

> It is simply part and parcel of the life of a species that first began cultivating the land, devising tools and machines, and cultivating-culturally developing-members of the species itself. Machines and artifacts are an inevitable part of human culture. Moral robots are merely a part that still lies in the future. [6][107]

REFERENCES

Allen, Colin, Gary Varner, and Jason Zisner (2000) Prolegomena to any future artificial moral agent. *Journal Experiment Theory Artificial Intelligence* 12: 251- 261.

[106] Page 252.
[107] Page 259.

Ames, Van Meter (1937) Conscience and Calculation. *International Journal of Ethics* 47: 180-192.

Andersen, Michael (2010) Robot Be Good. *Scientific American.* 303: 72-77.

Bailey, Andrew R. (1998) The Strange Attraction of Sciousness: William James on Consciousness. *Transactions of the Charles S. Peirce Society* 34: 414-434.

Barsalou, L.W. (1999) Perceptual symbol systems. *Behavioral and Brain Sciences* 22: 577–660.

Beiswanger, George (1950) The Logic of Conscience. *The Journal of Philosophy* 47: 225-37.

Borenstein, Jason (2010) Computing Ethics: Worklife in the Robotic Age. *Communications of the ACM* 53: 30-31.

Bostrom, Nick (2000) When machines Outsmart Humans. *Futures* 35: 759-764.

Bostrom, Nick (2003) Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence. I. Smit et al. (eds), *International Institute of Advanced Studies in Systems Research and Cybernetics* 2: 12-17. Revised at http://www.nickbostrom.com/ethics/ai.html Last accessed March 5, 2011.

Boutroux, Emile (1917) Liberty of Conscience. *International Journal of Ethics* 28: 59-69.

Boutroux, Emile (1917) The Individual Conscience and the Law. *International Journal of Ethics* 27: 317-333.

Brooks, Rodney, and Lynne Stein (1994) Building Brains for Bodies. *Autonomous Robots* 1: 7-25.

Cash, Mason (2010) Extended cognition, personal responsibility, and relational autonomy. *Phenomenology and Cognitive Science* 9: 645–671.

Childress, James F. (1979) Appeals to Conscience. *Ethics* 89: 315-335.

Clark, Andy (1998) Embodiment and the Philosophy of Mind. *Current Issues in Philosophy of Mind: Royal Institute of Philosophy Supplement* Cambridge University Press 43: 35-52.

Coeckelbergh, Mark (2010) Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology* 12: 209–221.

Danielson, Peter (2010) Designing a machine to learn about the ethics of robotics: the N-reasons platform. *Ethics Inf Technol* 12: 251–261.

Dietrich, Eric (2001) Homo sapiens 2.0: why we should build the better robots of our nature. *Journal of Experimental and Theoretical Artificial Intelligence*. 13: 323-328.

Dietrich, Eric (2007) After the humans are gone Douglas Engelbart Keynote Address, North American Computers and Philosophy Conference, Rensselaer Polytechnic Institute, August, 2006. *Journal of Experimental and Theoretical Artificial Intelligence*. 19: 55–67.

Earle, William (1970) Some Paradoxes of Private Conscience As a Political Guide. *Ethics* 80: 306-312.

Eliasmith, Chris: How we ought to describe computation in the brain. Available at http://www.arts.uwaterloo.ca/~celiasmi/cv.html Last accessed March 5, 2011.

Fowles, Don and Grazyna Kochanska (2000) Temperament As A Moderator Of Pathways To Conscience In Children: The Contribution Of Electrodermal Activity. *Psychophysiology* 37: 788–795.

Gallese, Vittorio, Christian Keysers and Giacomo Rizzolatti (2004) A Unifying View of the Basis of Social Cognition. *Trends in Cognitive Sciences* 8: 396-403.

Gips, James (1995) Towards the Ethical Robot. Appearing in *Android Epistemology*, Ford, Kenneth M., Clark Glymour and Patrick J. Hayes (eds). MIT Press. Available at http://www.cs.bc.edu/~gips/EthicalRobot.pdf Last accessed March 5, 2011.

Goldman, Alvin (2008) Hurley on Simulation. *Philosophy and Phenomenological Research* 77: 775-788.

Harnad, S. (1990) The Symbol Grounding Problem. *Physica D* 42: 335-346.

Hongladarom, Soraj (2009) An Ethical Theory for Autonomous and Conscious Robots. *AP-CAP 2009*, October 1–2, 2009, Tokyo, Japan. Unpublished manuscript.

Hurley, Susan (2008) Understanding Simulation. *Philosophy and Phenomenological Research* 77: 755-774.

James, William: *The Principles of Psychology* (1890). Available at http://psychclassics.yorku.ca/James/Principles/index.htm. Last accessed February 15, 2010.

Kahn, Peter H. Jr, Hiroshi Ishiguro, Batya Friedman, and Takayuki Kanda (2006) What is a Human?- Toward Psychological Benchmarks in the Field of Human-Robot Interaction. *The 15th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN06)*, Hatfield, UK. 364-371.

Kant, Immanuel (1886) *The Metaphysic of Ethics*. J.W. Semple (trans.) T and T Clark, Edinburg. Available at http://oll.libertyfund.org/EBooks/Kant_0332.pdfLast accessed March 5, 2011.

Klein, D.B. (1930) The Psychology of Conscience. *International Journal of Ethics* 40: 246-262.

Lavazza, Andrea and Mario DeCaro (2009) Not So Fast: On Some Bold Claims Concerning Human Agency. *Neuroethics.* From Springer's Online First section at http://www.springerlink.com/content/761q55v204473303/ Last accessed March 5, 2011.

Levy, David (2009) The Ethical Treatment of Artificially Conscious Robots. *International Journal Society Robotics* 1: 209–216.

Lissek, S., S. Peters, N. Fuchs, H. Witthaus, V. Nicolas, et al. (2008) Cooperation and Deception Recruit Different Subsets of the Theory-of-Mind Network. Available at http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0002023. Last accessed March 5, 2011.

Manicas, Peter: Men, Machines, Materialism, and Morality (1966) *Philosophy and Phenomenological Research* 27: 238-246.

Miceli, Maria and Cristiano Castelfranchi (1998) How to Silence One's Conscience: Cognitive Defenses Against the Feeling of Guilt. *Journal for the Theory of Social Behavior* 28: 287-318.

Mill, John Stuart (1985) *The Collected Works of John Stuart Mill, Volume X – Essays on Ethics, Religion, and Society*. ed. John M. Robson, intro. F.E.L. Priestley. University of Toronto Press, Toronto. E-Book available from the Online Library of Liberty, http://oll.libertyfund.org/title/241 Last accessed March 5, 2011.

Moor, James H. (2006) The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems* 21: 18–21.

Moor, James H. (2007) Taking the Intentional Stance Toward Robot Ethics. *APA Newsletter* 6: 14-17.

Natsoulas, Thomas (1996) The Sciousness Hypothesis - Part 2. *The Journal of Mind and Behavior* 17: 185-206.

Natsoulas, Thomas (1996) The Sciousness Hypothesis - Part I. *The Journal of Mind and Behavior* 17: 45-66.

Ochsner, Kevin N. and Matthew D. Lieberman (2001) The Emergence of Social Cognitive Neuroscience. *American Psychologist* 56: 717-734.

Olson, Robert G. (1959) A Naturalistic Theory of Conscience. *Philosophy and Phenomenological Research* 19: 306-322.

Perri, (2001) Ethics, Regulation and the New Artificial Intelligence, Part 1: accountability and power. *Information, Communication and Society* 4: 199-229.

Pollock, John (1990) Philosophy and Artificial Intelligence. *Philosophical Perspectives* 4: 461-498.

Ramachandran, V.S. (2004) *A Brief Tour of Human Consciousness*. Pearson Education, New York.

Reid, Mark D. (2005) Memory as Initial Experiencing of the Past. *Philosophical Psychology* 18: 671-698.

Ricoeur, Paul (1992) *Oneself as Another*. Trans. Kathleen Blamey. The University of Chicago Press, Chicago.

Singer, Tania and Chris Frith (2005) The Painful Side Of Empathy. *Nature Neuroscience* 8: 845-846.

Smith, Adam (1982) *The Theory of Moral Sentiments*, ed. D.D. Raphael and A.L. Macfie, In vol. I of the *Glasgow Edition of the Works and Correspondence of Adam Smith* Indianapolis: Liberty Fund. Ebook available at http://oll.libertyfund.org/title/192 Last accessed March 5, 2011.

Smithers, Tim (1997) Autonomy in Robots and Other Agents. *Brain and Cognition* 34: 88-106.

Sparkes, Matthew (2006) It May Sound Like Science Fiction, but a New Document Released by Euron is Encouraging Engineers to Think About the Ethical Implications of Robotics. *IET Computing and Control Engineering*. 10-11.

Spaulding, Shannon (2010) Embodied Cognition and Mindreading. *Mind and Language*, 25: 119-140.

Sun, Ron (2002) *The Duality of Mind: A Bottom-Up Approach to Cognition*. L. Erlbaum and Associates, New Jersey.

Sun, Ron (2009) Motivational Representations within a Computational Cognitive Architecture. *Cognitive Computing* 1: 91-103.

Sutton, John, Celia Harris, Paul Keil, and Amanda Barnier (2010) The Psychology of Memory, Extended Cognition, and Socially Distributed Remembering. *Phenomenology Cognitive Science* 9: 521–560.

Tamatea, Laurence (2010) Online Buddhist and Christian Responses to Artificial Intelligence. *Zygon* 45: 979-1002.

Thimbleby, Harold (2008) Robot ethics? Not yet. A reflection on Whitby's "Sometimes it's hard to be a robot." *Interacting with Computers* 20: 338–341

Tonkens, Ryan (2009) A Challenge for Machine Ethics. *Minds and Machines* 19: 421-438.

Umilta, M.A., E. Kohler, V. Gallese, L. Forgassi, L. Fadiga, C. Keysers and G.Rizzolatti (2001) I Know What You Are Doing: A Neurophysiological Approach. *Neuron* 31: 155-165.

Vargas, Manuel (2010) Situationism and Moral Responsibility: Free Will in Fragments. Forthcoming in eds. Vierkant, Till, Julian Kiverstein and Andy Clark. *Decomposing the Will*. New York: Oxford University Press. From draft of April 3, 2010. Available at http://usf.usfca.edu/fac_staff/mrvargas/Papers/Situationism.pdf. Last accessed March 5, 2011.

Vargas, Patricia A., Ylva Fernaeus, Mei Yii Lim, Sibylle Enz, Wan Chin Ho, Mattias Jacobsson, and Ruth Ayllet. (2011) Advocating an Ethical Memory Model for Artificial Companions from a Human-Centred Perspective. *AI and Society*. Available from http://www.springerlink.com/content/6705q501kvw5p1p7/ Last accessed March 5, 2011.

Velleman, J. David (1999) The Voice of Conscience. *Proceedings of the Aristotelian Society* 99: 57-76.

Versenyi, Laszlo (1974) Can Robots be Moral? *Ethics* 84: 248-259.

Wallach, Wendell (2010) Robot Minds and Human Ethics: the Need for a Comprehensive Model of Moral Decision Making. *Ethics Information Technology* 12: 243-250.

Ward, Bernard (1961) The Content and Function of Conscience. *The Journal of Philosophy* 58: 765-772.

Warwick, Kevin (2010) Implications and Consequences of Robots with Biological Brains *Ethics Inf Technol* 12: 223-234.

Weng, Yueh-Hsuan, Chien-Hsun Chen, and Chuen-Tsai Sun (2009) How can humans and robots coexist safely? Toward the Human–Robot Co-Existence Society: On Safety Intelligence for Next Generation Robots. *International Journal Society Robotics* 1: 267–282.

White, Jeffrey (2010) Understanding and Augmenting Human Morality, the ACTWITH model. In L. Magnani, C. Pizzi, and W. Carnielli (eds.) *Studies in Computational Intelligence #314: Model-Based Reasoning in Science and Technology*: 607-620. Springer.

White, Jeffrey (Forthcoming) *Conscience: the mechanism of morality*. (monograph, 190,000 words).

Wicker, Bruno, Christian Keysers, Jane Plailly, Jean-Pierre Royet, Vittorio Gallese and Giacomo Rizzolatti (2003) Both of Us Disgusted in My Insula: The Common Neural Basis of Seeing and Feeling Disgust. *Neuron* 40: 655-664.

Wright, William K. (1916) Conscience as Reason and Emotion. *Philosophy Review* 25: 676-691.

LCH