Understanding and Augmenting Human Morality: An Introduction to the ACTWith Model of Conscience

Jeffrey White

Abstract. Recent developments, both in the cognitive sciences and in world events, bring special emphasis to the study of morality. The cognitive sciences, spanning neurology, psychology, and computational intelligence, offer substantial advances in understanding the origins and purposes of morality. Meanwhile, world events urge the timely synthesis of these insights with traditional accounts that can be easily assimilated and practically employed to augment moral judgment, both to solve current problems and to direct future action. The object of the following paper is to present such a synthesis in the form of a model of moral cognition, the ACTWith model of conscience. The purpose of the model is twofold. One, the ACTWith model is intended to shed light on personal moral dispositions, and to provide a tool for actual human moral agents in the refinement of their moral lives. As such, it relies on the power of personal introspection, bolstered by the careful study of moral exemplars available to all persons in all cultures in the form of literary or religious figures, if not in the form of contemporary peers and especially leadership. Two, the ACTWith model is intended as a minimum architecture for fully functional artificial morality. As such, it is essentially amodal, implementation non-specific and is developed in the form of an information processing control system. There are given as few hard points in this system as necessary for moral function, and these are themselves taken from review of actual human cognitive processes, thereby intentionally capturing as closely as possible what is expected of moral action and reaction by human beings. Only in satisfying these untutored intuitions should an artificial agent ever be properly regarded as moral, at least in the general population of existing moral agents. Thus, the ACTWith model is intended as a guide both for individual moral development and for the development of artificial moral agents as future technology permits.

Jeffrey White
KAIST, South Korea
e-mail: jbenjaminwhite@mail.com

L. Magnani et al. (Eds.): Model-Based Reasoning in Science & Technology, SCI 314, pp. 607–621.
springerlink.com

1

The ultimate goal of A.I., generally, is the construction of a fully embodied and fully autonomous artificial agent. This task poses special challenges, of course, especially the reconciliation of neural research with traditional thinking on intelligence and autonomy [7]. In the development of autonomous moral agents, some authors have contended that the starting point is in the selection of a suitable moral framework for implementation into moral machines [22][1]. However, I disagree with this tact. Though the necessary and sufficient physical mechanisms cannot yet be articulated, either for artificial or natural moral agents (humans), the approach that the following work takes is to first specify the necessary architecture, and then to see what moral framework arises from the proper function of that architecture[2].

The architecture at issue is the ACTWith model of moral cognition, or the ACTWith model of conscience. The scope of the present work forbids exhaustive review of pertinent research from diverse fields all touching on the issues of conscience in practice and theory, artificial morality, and neurological mechanisms at work in moral cognition. However, a brief review is necessary in order to indicate important points of reference. The ACTWith model is at root a bottom-up hybrid architecture, originally informed by Ron Sun's CLARION architecture.[21] However, as it is developed here, it is intentionally task and implementation non-specific, being essentially a model of control of information processing[3]. The model builds from two key insights into moral cognition from neurology, disgust and mirroring[4]. It is essentially a model of situated cognition, and although developed independently, it is consistent with work from situationist psychology[3], and represents a strong form of embodiment[8].

The scope of this paper forbids an exhaustive inquiry into the nature of conscience[5]. However, in this section, I will provide some disambiguating

[1] This seems to mirror the method in which moral theory is often pursued, as well.

[2] That neurology, especially, has ~~not~~ already delivered the final word on human morality is a common misconception amongst many. Though one might presume the case closed on moral theory, that we must only wait for the neurologist to tell us what the brain tells us is right and wrong, this is an overly hasty position. Even if the neurosciences level some incontrovertible facts, there remains the issue of interpretation of these facts and the integration of such into existing practices. For discussion, see [9, 15].

[3] For relative advantages to this approach, see [10].

[4] See [11, 23, 27] for initial discussion. The view put forward here is not to be confused with that of popular "mindreading" theorists. I have trouble with this program for reasons too detailed to develop here, but one issue involves the disputes within the body of researchers themselves over what mindreading actually amounts to. See [12] and [13] for examples.

[5] I take on this task in my current book manuscript Conscience: the mechanism of morality, forthcoming with publication expected 2010. See [1, 4, 5, 6, 26, 14, 18, 24, 25, 28] for an introduction to some basic issues in conscience, especially concerning its naturalization and psychological interpretation.

remarks, first in regards to conscience itself, and later in regards to conscience and consciousness.

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. Even the seemingly simple and most familiar characterization as a warning voice carries deep implications that demand some specification. For instance, conscience as that universally recognized voice which rises against acting towards morally repulsive ends cannot be merely a simple voice[6]. After all, for it to fulfill even this seemingly simple function, the operations of conscience must extend through all levels of end selection. In order to reject some ends while endorsing others, conscience must act as the steering mechanism of the entire embodied complex that is the moral agent. And that is a very complex concept, indeed. In simple terms and for purposes of introduction, conscience can initially be understood as naming the extended homeostatic function of body to sustain personal integrity in the face of a changing environment, presented in the basic ACTWith model as a generic mechanism which regulates the opening and closing to environmental input, a process which leads to the accumulation of experience[7] which is used to guide future operations of the same mechanism.

Conscience is historically, and linguistically, related to consciousness[8]. In fact, the term conscience precedes that of consciousness by some 300 years, and it is from conscience that the term consciousness originally derives[9]. However, the historical use of these terms is beside the point, now, as consciousness receives a great deal more attention than does conscience, and either clearly represent two very distinct aspects of the human condition, however less than clear their namesakes remain.

We may gain clarity on both terms by exploiting their structural similarities. Both consciousness and conscience consist of conjunctions between a prefix “con-” and a root, “sciousness” and “science”, either of which carry individual connotation. “Con-” means “together”, or “with”. It is a prefix that indicates synthesis. “Sciousness” was proposed by William James in the

[6] And to say that it is raises further questions about the nature of verbal language and the origins of symbols, themselves.

[7] Initially understood as memory, see [20], but eventuating in embodied adaptations due to peripheral attunements, i.e. hormones and general metabolism, over time.

[8] Consciousness, as well, has been understood as an extension of homeostatic mechanisms. See [19]

[9] See for example http://www.etymonline.com/index.php?search=conscience\&searchmode=none. Last accessed February 15, 2010.

10^{th} chapter of his landmark text, Principles of Psychology, to be a foundation for consciousness. He employed introspection, the only psychological tool available to him at the time, to inquire into the nature of consciousness and found a rolling stream of sensation that receded from his introspective projections just outside his conscious reach. “Con-sciousness”, thus, can be taken to mean the synthesis of merely felt moments into discretely realized phenomena[10]. Accordingly, sciousness can be understood as the felt ground of all discrete thought, consisting of clear and distinct ideas in the classical Cartesian sense of self-awareness [2, 16, 17].

“Science”, the relative root of the term “conscience” conveys a strikingly different sense, at least on initial inspection. Typically, “science” implies a specific field of knowledge and inquiry, constituted by certain systematic principles of relation between a specific and select body of objects. Examples such as Chemistry, consisting of chemists working in the field of entities related by chemical laws and constitutive of chemical theories over a specific set of chemical objects, make this use of the word “science” clear enough.

Yet, there is something universal about the use of the word “science” that ties all of the seemingly discrete fields of inquiry together, and it is from this universal implication that the term “con-science” should be construed. This universal nature is that “science” as the root of “con-science” represents what it is to be in *any* field of *any* set of objects, however non-specific, which are bound by any principles however non-systematic. In effect, “science” can be taken to name the field in which each each person is individually (and persons are collectively) embedded, and in terms of which he or she seeks successful action and even truth. It 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, further understands (learns), or fails. “Science”, in this sense, is reducible to “situation” in a very strong sense, being the irreducible complex of agent and environment, understood from the perspective of the experiencing agent, or subject. “Con-science” can be understood, then, as the synthesis of embodied situations, the “what it feels like” to be in a place at a time, and in such processing produces information on the differences – both merely felt and otherwise cognized – between the relative value of one situation with any other.

In this way, conscience, understood fully as an embodied mechanism, serves as a motivational and self-preserving extension of basal homeostatic mechanisms common to all sufficiently complex organisms[11]. An organism that is able to evaluate the relative values of situations will seek those situations that feel good, and avoid those that feel bad, as these situations are effectively environments in terms of which that organism must subsequently reach

[10] 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 [21].

[11] Again, these issues are more adequately developed in *Conscience: the mechanism of morality.*

homeostatic equilibrium. Conscience, thus, and morality by further extension, operate according to this logic, but present themselves in recognizable forms only in organisms of necessary complexity, such as human beings.

The scope of the present paper does not permit a thorough explication of the relationship between conscience and consciousness, or of the place of conscience as part of an organism's homeostatic mechanisms. But, the preceding brief account does specify the guiding role of conscience in the motivation of any autonomous moral agent, artificial or otherwise, and opens the window to develop in simpler terms a generic mechanism from which we might conceive a moral framework emerging, that being a framework from which moral action proceeds and in terms of which moral judgment can be based. In the next section, I will detail the basic ACTWIth model in a more easily appropriated form derived from hybrid neural net models.

2

The ACTWIth model is a four-step cycle, with two belonging to a top (rational) level and two to a bottom (affective) level, with each step a related mode of information processing. This structure is captured in the name, "ACTWIth"[12]. "ACTWith" stands for "As-if" "Coming to Terms With". "As-if" involves feeling a situation out, while "Coming to Terms With" involves defining the situation in terms of the things originally felt. 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)

These modes are intended to represent the bare minimum for the eventual emergence of morality. The closed modes are derived from the mechanism of disgust, while the open modes are from mirroring mechanisms, both affective and action oriented. While the systems in which these mechanisms must operate are not here specified, at the level of implementation, inspiration might be drawn from primate or from human brains, as some researchers are doing in non-moral realms presently, or they can be taken purely from computational intelligences, which are also common in the study of human learning, cooperation, and motivation. In any event, it is not the purpose of the present work to detail potential applications.

Altogether, the four modes can be visualized as follows (see Figure 1).

[12] ACTWith, either in name or function, bears no deliberate relationship with the famous ACT-R model.

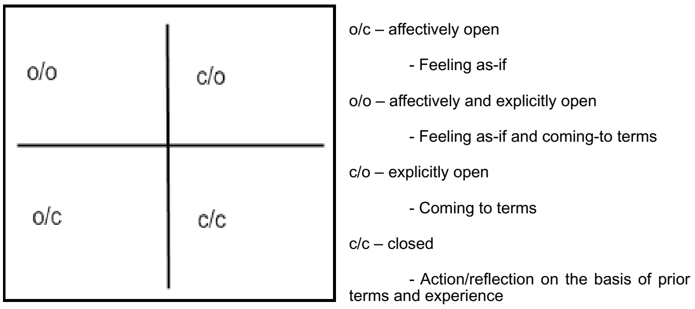
o/o
c/o
o/c
c/c
o/c – affectively open
- Feeling as-if
o/o – affectively and explicitly open
- Feeling as-if and coming-to terms
c/o – explicitly open
- Coming to terms
c/c – closed
- Action/reflection on the basis of prior terms and experience

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

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. For instance, consider the mode o/c. This personality is open to other situations affectively, but not at the level of discrete reason. If a person were to habitually engage in this mode when dealing with others, he or she 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. Contrast this mode with that of c/o. This personality is closed, affectively, but open at the level of discrete reason. If a person were to habitually engage in this mode, he or she would not be able to feel what it is like to be in another's situation, but would be interested in having an explanation for why that person is in that situation, how he or she plans to get out of it, and etcetera. The first may seem warm, but "flaky", while the second may seem cold, and calculating. Ether represent personality types that are common, enough, to be easily recognized as archetypes.

The o/o and the c/c modes are the most interesting, and the most recognizable. The o/o mode, when habitually employed, represents the genuine saint. This personality is both affectively open to another's situation as well as genuinely interested in understanding what it is like to be in that situation at least insofar as that other understands it. In practice, this sort of person is exceptionally rare, while the habits that lead to its realization are the object of many if not most religions. Buddhist practitioners (of some strains) stand out as exemplifying this mode as habitually employed. Meanwhile, the c/c mode is the opposite of the o/o mode. Persons habitually employing this mode are selfish, arrogant sorts who come off both as cold and calculating. This personality is perhaps most recognizable, as it represents a being who is both unable to feel what it is like to be in another's situation, as well as being disinterested in understanding why he or she is in that situation and how or why he or she would plan to leave it. This is the mode of the psychopath.

Different personality types can be rendered more finely by recognizing that these modes may be habitually employed only in certain types of situations. As “another situation” equally means one’s own or another’s situation, one may be completely open to one’s own different situations (o/o) while being indifferent to those of other persons (c/c) or interested in them solely insofar as understanding those situations fortifies his or her own understanding of his or her own place in the world (c/o). Over the long course of personal development, it is easy enough to see how the habitual employment of one or another of these four modes of information processing can lead to a wide diversity of personality types.

To articulate these four modes in static terms, in terms of habitual employment at the exclusion of one another, is useful for illustrative purposes. However, any realistic model of agency must be dynamic. The ACTWith model is, fully developed, a cycle of information processing. It can be represented thusly (Figure 2):

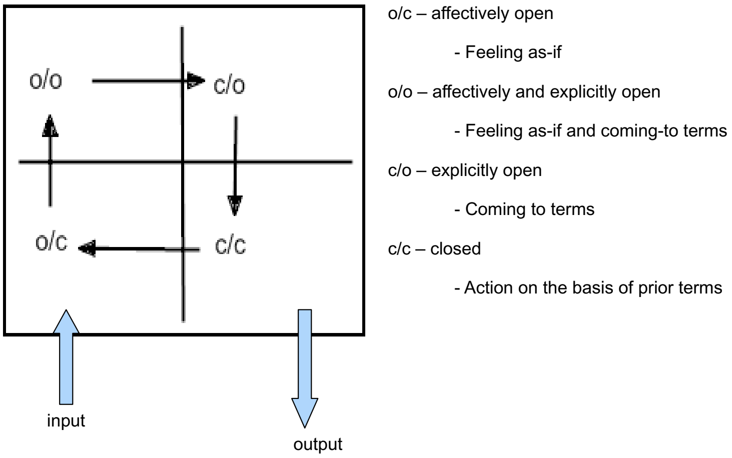

Fig. 2 The Beating ~~Heat~~ Heart of Conscience.

This model is “the beating heart of conscience”, recognizing the fact that the conscience has traditionally been associated with the beating of a heart, 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 to avoid situations that are either beneficial or contrary to integrity (physical or otherwise) and survival.

In order to 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”[13]. Consider the following scene. A cold and lonely 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, dressed in rags, and in the bitter wind, the stinking vapor - his only source of heat - turns to ice in his ratty beard. At first glance, the man is 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.

At the instant that the agent comes onto the scene, he has a chance to either open to the plight of the poor man, or to close to it. Let’s consider the open mode, first. In opening to the poor man, the agent will perhaps at first have to overcome disgust in order to mirror both the feeling that is expressed by the man as well as mirror potential action paths, as opening means feeling as if the agent were that other man, however momentarily[14]. First, the agent will appreciate the situation from the agent’s own prior experience (o/c). Then, 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. 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 in the words of Adam Smith “makes himself at home” in the situation, (c/o). 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. The feeling of being literally moved in compassion for another is ~~he~~ the product of this process. Finally, the agent will be able to reflect on his new experience, and either open once again to the situation, searching for greater understanding[15], or act – perhaps by offering the poor man some charity – and move on to other situations, enriched for the new experience (c/c).

[13] 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 an immoral or amoral person by habit if not by reputation.

[14] 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 an immoral or amoral person by habit if not by reputation.

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

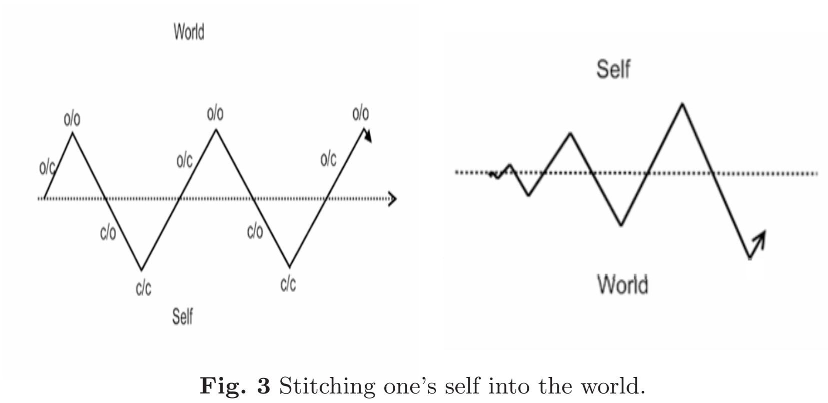

Fig. 3 Stitching one's self into the world.

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, and though the cycle of cognition that is the beating heart of conscience proceeds uninterrupted, the agent "without a conscience" has a heart only for its own self.

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. The habitual employer of the open mode, on the other hand, suffers and is in fact increasingly burdened as more and more persons fall to desperate situations in the wake of the selfish stampede for success.

In either case, the lesson is that agents shape their environments through their actions[16]. The agent shapes the world through action, thereby setting 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 related on this picture. As the agent opens to the world, the agent takes up the understanding of this situation, and carries it into the next situation, and so on. Thus, in opening and in closing to the world, the agent becomes the product of the terms generated. This process is illustrated in Figure 3.

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

[16] The student of Philosophy will recall that this is the essential message and the primary motivation behind J.S. Mill's Utilitarianism, and may also recall the central role of conscience therein.

habitually open mode, which leads to what existentialist have called the "beautiful soul" and that phenomenologists have called "authenticity".

At this point, the role of conscience in personal freedom, freewill, can be clarified. As is shown in the preceding figure, and as is alluded to in the preceding discussion, the role of conscience in freedom is that it serves as the mechanism which makes the freedom of self-determination a real possibility. Conscience is not the seat of something more, some radical freedom, 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. Instead, conscience is a steering wheel of sorts, a gentle handle on personal self-transformation and, perhaps, even personal transcendence, though any further discussion on these issues is beyond the immediate scope of this paper[17].

The question we are left with as we turn to consider what sort of moral framework emerges from the model developed thus far has less to do with what one will do in a given situation, and more to do with who one wishes to become as the product of his or her experience from either opening or closing to situations as they change. This leads us to the final section, and returns us to recent considerations on the possibility of autonomous moral agents.

3

The job of conceiving of autonomous moral agents is difficult enough, but the task becomes more difficult for the fact that human moral agency is not so well understood. Even in cases where one would think the matter settled, such as an application of traditional moral theory to the conception of moral agency, there is the additional problem of misinterpretation of moral theory that must be dealt with before one can attend to the issue of agency design.

Consider Kant's moral theory in this light. According to some commentators, the role of conscience in Kant's moral theory is merely that of the traditional voice of conscience, warning against immoral action. Conscience is simply recast as the representative voice of the categorical imperative. On this account, conscience rises to awareness when one is considering an action which would violate the categorical imperative[24]. Still other appropriations of Kant's moral theory, specifically into discussions of on the possibility of autonomous moral agents, fail to consider conscience at all [22].

However, such accounts are not consistent with Kant's greater moral theory. 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 person and, in fact, in a process which, mirroring Stevan Harnad's famous "symbol grounding" problem, grounds moral action through Kant's infamous "goodwill". Issues of space forbid a full exposition of these claims. So, in this section, I will simply lay out a cursory interpretation of Kant's moral theory

[17] These issues are fully explored in *Conscience, the mechanism of morality.*

along these lines, illustrate how it might arise through the proper functioning of an ACTWith endowed agent per the discussion in the previous two sections, and finally redraw Kant's categorical imperative in terms consistent with both.

Let us first consider what Kant means by morality, not forgetting the example of the poor man from the last section. In Section 35 of *The Metaphysics of Ethics*, Kant tells us that "[...] although it is no direct duty to take a part in the joy or grief of others, yet to take an active part in their lot is [...]" and that we ought not "avoid the receptacles of the poor, in order to save ourselves an unpleasant feeling, but rather to seek them out". 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". In these lines, Kant paints a picture of an affectively motivated moral agent, compelled as a "spring" overcoming reason in discharging of what he terms one's moral duty. Just what this moral duty is will become clear in a moment. And, what is this "spring"? It is conscience, not misunderstood as mere warning light for the categorical imperative, itself understood as a purely rational directive, but instead understood as the spring that motivates a person according to the logic of moral affect.

The affect central to Kant's moral theory is goodwill. What is good will? Earlier, 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 [...]" And, later, in the second section "That, we now know, is a good will whose maxim, if made law universal, would not be repugnant to itself". Thus, it is good will both that one aspires to (insofar as one wishes to be moral) and that guides action along the way. Here, it is important to note that repugnance is another word for disgust, both of which are not concepts belonging to reason, where typical misinterpretations on Kantian ethics place the locus of moral motivation (in rationality), instead.

How does good will work to motivate to moral ends via moral actions? By Kant's account, goodwill alone is not enough. One must also have in mind some exemplar, some other embodied agent, whether real or ideal, in light of which one may, at least initially, model ones actions, and thus eventually one's self. The emotion that signifies the importance of these examples is reverence, and in fact the object of reverence serves as the measuring stick for one's own moral worth. Kant tells us, in the notes to chapter 1, that "What is called a moral interest, is based solely on this emotion". And what is reverence? Without prying any further detail directly from Kant's own writings in support of the claim, it can be understood, in contemporary terms, to involve the employment of mirroring capacities of the human body to emulate, and so train, one's self to adopt and thus become like another human being, whether that being be, on Kant's account, real or ideal. Moral interest, thus, is fundamentally to become the best person one can become.

In these two concepts, reverence and goodwill, the opening and closing functions of the ACTWith model are plotted onto Kant's moral theory. So, where is conscience in all of this? Conscience is the binder of the two. In a section of the Metaphysics of Ethics interestingly entitled "Prerequisites towards constituting man a moral agent", Kant affirms that one's understanding is the limit whereby he or she can determine right or wrong, writing that "obligement can extend only to the illuminating his understanding as to what things are duty, what not". And this returns us to the notion of moral duty, and to the question what is this "spring implanted in us by nature" that motivates a person to seek to fulfill this duty. Both the duty that is attached to action, and the spring that motivates to one's highest potential as a person, to become worthy of reverence through the exhibition of goodwill, are the subjects of conscience. To this end, 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.

In other words, one's highest potential is to be conscientious, and one's primary duty in action is to maximize this potential through conscientiousness, the habitual act thereof maximizing one's understanding, and so expanding one's potential to recognize his or her obligation to others in the fulfillment of moral duty. It is a cycle. And, it is easy to see that this process leads directly to the "beautiful soul"[18].

Finally, shortly after the preceding statement, Kant spells out this duty for conscientious moral agents when passively serving as models and guides for others, according to the same logic of disgust and mirroring:

> 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 in 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.

Altogether, we have a portrait of Kantian moral theory which can be understood as a direct extension of the mechanisms at work in the ACTWith model. Accordingly, it serves to reconsider the categorical imperative in light of these results. Arguably, the most famous form of the categorical imperative is the following, and the one which Kant himself prefers as he restates

[18] 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.

it in chapter 2 of *The Metaphysics of Ethics*: "Act according to that maxim which thou couldst at the same time will an universal law". In light of the present results, especially in view of the role of conscience in the preceding appropriation of Kant's moral theory, this imperative can be rewritten in the following forms:

1. Do not become through action (or inaction) an object of self-disgust.
2. And, conversely: Do become through action (or inaction) an object of reverence.
3. And, most simply: Do not put another into a situation that you would not seek for your own[19].

4 Conclusion

This paper has put forward a model of moral cognition consistent both with neurological insights into human motivation to moral action and to becoming a moral person. What are the implications of this proposal? Ideally, it serves in two ways. One, it may redirect focus in the development of autonomous moral agents away from the post-hoc introduction of ethical systems or principles, either as strictures or as measures of moral performance, and toward the development of morally productive architectures from the ground up. As technology develops, limitations to applications increasingly derive from the conceptions which drive and inspire these applications rather from the technology, itself. In terms of moral agents, thus, it is up to the moral philosopher to prefigure these potential applications by providing frameworks of the broadest possible scope with the greatest possible explanatory power. The future of the development of autonomous moral agents, in my mind, depends on this. The ACTWith model proposed here is intended to serve as a starting point in exactly this way.

Two, it may open the way for computational, control, and systems theories of moral agency to be employed increasingly as tools in the analysis and augmentation of human moral conduct. The flow of information from man to machine is bi-directional. It goes both ways. As these models are developed, they require testing and evaluation, and the only method available is against direct human experience. Further, in the testing, we human beings stand to learn something about ourselves that may have lain hidden without the mediation of the models under review.

Finally, it is my hope that the ACTWith model serves as an introspective guide for the moral practice of actual, living people whose interests rest alongside that put forward by Immanuel Kant and so many other moral philosophers before and since: to become, through reflection, and perhaps through the use of what may be called "moral mediators" in the spirit of Lorenzo Magnani's "epistemic mediators", the best people that they can possibly become.

[19] Which might invite a violation of either 1 or 2.

References

1. Ames van, M.: Conscience and calculation. International Journal of Ethics 47, 180–192 (1937)
2. Bailey, A.R.: The strange attraction of sciousness: William james on consciousness. Transactions of the Charles S. Peirce Society 34, 414–434 (1998)
3. Barsalou, L.W.: Perceptual symbol systems. Behavioral and Brain Sciences 22, 577–660 (1999)
4. Beiswanger, G.: The logic of conscience. The Journal of Philosophy 47, 225–237 (1950)
5. Boutroux, E.: The individual conscience and the law. International Journal of Ethics 27, 317–333 (1917)
6. Boutroux, E.: Liberty of conscience. International Journal of Ethics 28, 59–69 (1917)
7. Brooks, R., Stein, L.: Building brains for bodies. Autonomous Robots 1, 7–25 (1994)
8. Clark, A.: Embodiment and the philosophy of mind. Current Issues in Philosophy of Mind 43, 35–52 (1998)
9. Dean, R.: Does neuroscience undermine deontological theory (2010), doi:10.1007/s12152-009-9052-x
10. Eliasmith, C.: How we ought to describe computation in the brain (2010), http://www.arts.uwaterloo.ca/~celiasmi/cv.html (last accessed February 15, 2010)
11. Gallese, V., Keysers, C., Rizzolatti, G.: A unifying view of the basis of social cognition. Trends in Cognitive Sciences 8, 396–403 (2004)
12. Goldman, A.: Hurley on simulation. Philosophy and Phenomenological Research 77, 775–788 (2008)
13. Hurley, S.: Understanding simulation. Philosophy and Phenomenological Research 77, 755–774 (2008)
14. Klein, D.B.: The psychology of conscience. International Journal of Ethics 40, 246–262 (1930)
15. Lavazza, A., De Caro, M.: Not so fast: On some bold claims concerning human agency (2010), doi:10.1007/s12152-009-9053-9
16. Natsoulas, T.: The sciousness hypothesis - part i. The Journal of Mind and Behavior 17, 45–66 (1996)
17. Natsoulas, T.: The sciousness hypothesis - part ii. The Journal of Mind and Behavior 17, 185–206 (1996)
18. Olson, R.G.A.: Naturalistic theory of conscience. Philosophy and Phenomenological Research 19, 306–322 (1959)
19. Ramachandran, V.: A Brief Tour of Human Consciousness. Pearson Education, New York (2002)
20. Reid, M.D.: Memory as initial experiencing of the past. Philosophical Psychology 18, 671–698 (2005)
21. Sun, R.: The Duality of Mind: A Bottom-Up Approach to Cognition. L. Erlbaum and Associates, New Jersey (2002)
22. Tonkens, R.: A challenge for machine ethics. Minds & Machines 19, 421–438 (2009)
23. Umilta, M., Kohler, E., Gallese, V., Forgassi, L., Fadiga, L., Keysers, C., Rizzolatti, G.: I know what you are doing: A neurophysiological approach. Neuron. 31, 155–165 (2001)

24. Velleman, J.D.: The voice of conscience. Proceedings of the Aristotelian Society 99, 57–76 (1999)
25. Ward, B.: The content and function of conscience. The Journal of Philosophy 58, 765–772 (1961)
26. William, W.: Some paradoxes of private conscience as a political guide. Ethics 80, 306–312 (1970)
27. Wilson, E.: Consilience: The Unity of Knowledge. Random House, New York (1998)
28. Wright, W.K.: Conscience as reason and emotion. Philosophy Review 25, 676–691 (1916)