Particularism and Generalism: How AI can Help us to Better Understand Moral Cognition

Marcello Guarini

Department of Philosophy, University of Windsor
401 Sunset, Windsor, Ontario. Canada. N9B 3P4
http://www.uwindsor.ca/guarini
mguarini@uwindsor.ca

Abstract

Particularism and Generalism refer to families of attitudes towards moral principles. This paper explores the suggestion that neural network models of cognition may aid in vindicating particularist views of moral reasoning. Neural network models of moral case classification are presented, and the contrast case method for testing and revising case classifications is considered. It is concluded that while particularism may have some legitimate insights, it may underestimate the importance of the role played by certain kinds of moral principles.

1. Particularism and Generalism

Much ink has been spilled on the nature of moral reasons. Some philosophers have defended Particularism, while others have defended Generalism. These terms can be misleading since there are a number of different theses that appear to go under the heading of 'Particularism.' Since Generalists are taken as denying what Particularists claim, getting clear on some of the different positions referred to as particularistic will also help us to clarify some possible generalist commitments. The first part of this paper will clarify some of the different positions that have come under the heading of Particularism. Part Two will raise some questions for particularism and present two artificial neural network models in an attempt to explore what possible answers might be to those questions. Part Three will discuss how the construction of such models may lead to insights that require us to move beyond the boundaries of the Particularism-Generalism debate, leading us to a better understanding of the space of possibilities for the nature of moral reasons. Part Four will discuss the relationship between this work and other work.

Particularisms

Particularism is often defined in terms of an attitude towards moral principles (Dancy 1993 & 2000). Thus, differing attitudes towards moral principles and different conceptions of moral principles lead to different versions of Particularism. Let us begin by examining the different types of moral principles Particularists tend to consider. First, principles may be conceived as exceptionless rules that (a) specify sufficient conditions for what makes a state of affairs (which are taken to include actions) or an entity (which includes persons) appropriately described by predicates such as good, bad, right, wrong, impermissible, permissible, acceptable, unacceptable and so on; (b) explain or otherwise shed some light on why the principle applies when it does, and (c) are serviceable as premises in moral deliberation. Call this the *exceptionless standard* conception of a principle. The exceptionless standard could state that, for example, all actions of a specific type are to be treated in a certain way – "Any act involving killing is morally unacceptable." The reasons for adding *(b)* and *(c)* to this conception of a principle is that particularists will generally concede that the moral supervenes on the non-moral; in other words, particularists generally agree that there can be no moral or prescriptive difference between two entities or states of affairs unless there is also non-moral or descriptive difference between the two entities or states of affairs (Hooker and Little, 2000). If we concede supervenience, then it might be argued that there may always be some exceptionless moral principle(s) provided we are prepared to countenance one or more very long and complex moral principles. However, a principle that would take 1000 encyclopedia volumes to fully articulate would be so complex that (i) it may not shed any light for the average human cognizer on why the principle applies, and (ii) it would not be serviceable as a premise in moral deliberation.

Of course, principles need not be conceived of as exceptionless standards. Rather, they may be seen as stating what sorts of predicates contribute to moral deliberation without trying to state sufficient conditions for when a state of affairs or an entity is appropriately described by some moral predicate. For example, it might be asserted that "Killing always contributes to the wrongness of an action." It is consistent with this contributory principle that an act of killing may be morally

acceptable. For example, it might be said that killing in defense of the innocent may be acceptable. It may just be that while killing contributes to the wrongness of the act, other considerations (preserving the lives of innocents) would contribute to the rightness of the act, and the factors contributing to wrongness are outweighed by the factors contributing to rightness. In other words, all things considered, an action may be right (or permissible, or acceptable…) even if it contains wrong-making (or impermissible-making, …) features.

Two conceptions of moral principles have been identified; let us now examine some of the attitudes that particularists may take towards those principles. Moral principles do not exist – following McKeever and Ridge (2005), we can call this attitude Principle Eliminativism. There are different ways to formulate this position. One can be an eliminativist with respect to exceptionless standards or contributory principles or both. It is also possible to take approaches to principle eliminativism that vary in scope. For example, one might say that there are no moral principles (exceptionless, prima facie, or both) whatsoever. However, one might assert that there are moral principles (exceptionless, prima facie, or both) but only in *some* domains, while in other domains there are no moral principles.

Principle abstinence is another attitude towards principles. It is the view that while moral principles might exist, our moral reasoning will be of higher quality if we avoid invoking principles. The idea, roughly, is that our moral life and reasoning about it is so complex that focusing on principles will tend to over simplify situations and lead us astray. Moral reasoning, allegedly, would proceed better by abstaining from the use of moral principles.

Two views of principles have been identified, and two general attitudes towards principles have been identified. Different combinations of these attitudes are possible. For example, if one is a Principle Eliminativist with respect to exceptionless and prima facie principles, one will likely have a strong preference for Principle Abstinence. Moreover, it is possible for one to endorse Principle Abstinence and reject Principle Eliminativism of all kinds. For example, one might be of the view that while principles exist, they are not generally helpful in moral reasoning, so they should be avoided. It is not being stated here that eliminativism and abstinence are the only possible attitudes towards principles, but they are among the better known views. The focus of this paper will be on a form of Principle Eliminativism asserting that at least in some domains, neither exceptionless nor prima facie principles exist.

So far, we have identified two types of principles and two different attitudes that particularists may take towards each type of principle. Particularists may also vary on the scope of their particularism. Moral judgments are made about either *entities* or *states of affairs*. Among the entities subject to moral evaluation are persons, countries, corporate entities or societies:

> Jack and Jill are good people.
> Canada is a just country.
> Enron is an irresponsible corporation.
> The Rotary Club does good work.

Actions and other states of affairs can also be subject to moral evaluation:

> Jack ought not to have spoken to Jill in that way.
> That 10% of the population controls 90% of the wealth is unjust.

While there is a relationship between judgments about people and judgments about corporate entities, it does not follow that judgments about corporate entities is redundant or eliminable. For example, Jack and Jill may be good people, and they may have worked for Enron and belonged to the Rotary Club. Enron may still be called a corrupt company even if Jack and Jill were perfectly honest employees, and the Rotary Club may do good work even if Jack and Jill were passive members and never did any of that work. While there is also a relationship between judgments about persons (on the whole) and judgments about specific actions, once again, it is not as if judgments about persons are redundant or eliminable. When we judge a person, we judge their character *on the whole*, their dispositions to behave in certain ways. One morally questionable *action* does not automatically make someone a bad *person*. For example, if Jill is a wonderful human being who has never stolen anything in her life, and one day she is caught stealing a pen from work, it does not follow that she is a bad person or that she has a bad character. Recognizing that Jill has a good character (all things considered) her supervisor will likely *not* dismiss her.

It is useful to be clear on the point that there are different objects of moral evaluation since principles may have an important role to play in assessing all entities and states of affairs, neither entities nor states of affairs, entities but not states of affairs, states of affairs but not entities, some entities and some states of affairs, or other combinations of entities and states of affairs. The point is that one may think that principles play an important role with respect to some objects of evaluation and not others. In other words, one may be a generalist with respect to some objects of evaluation and a particularist with respect to others.

2. Challenges of Particularism

A number of interesting questions arise for different forms of particularism.

1. In those domains where Particularism is alleged to be true, how do we learn to classify cases if we are not grouping them under common principles?
2. How do we generalize from the cases we have learned to new cases?
3. (a) How do we come to recognize that our initial classification of cases needs revision? (b) How do we carry out that revision?

The second of these questions is pressing since it appears cognitively implausible that each situation we learn to classify as morally permissible or impermissible is *completely* different from every other case we learn to classify. The idea that intelligent beings (natural or artificial) could exhibit the kind of real-time classification prowess that adult humans generally do while functioning as massive look-up tables is implausible. There are too many possible different cases that we know how to classify, and our ability to classify is often quick and effortless. If intelligent beings are functioning as look-up tables, this would presuppose that any case that is encountered is already stored on the table. On its own, this assumption is cause for concern, but the concern becomes greater when we realize that no one has a plausible model for searching such a table (without using substantive moral principles to guide the search) in real time. If cases are not being grouped together under principles, then how do we generalize to newly encountered cases? Indeed, if we are not using principles of any kind or a look-up table, then how do we carry out our original classification of cases? Presumably, there is some connection between how we classify cases we are presented with in our education and how we learn to generalize to new cases. If principles and look-up tables are not involved in generalizing to new cases, it is hard to see why they would play a significant role in the original learning of cases. Allowing principles to play a significant role in learning would be a serious concession for particularism. Moreover, storing all cases and classifications encountered during the learning phase in a massive look-up table would not appear to be useful for generalizing. So how do intelligent beings learn to classify cases? Finally, assuming we have some sort of model for classifying cases and generalizing to new cases, how do we extend it to revise our initial classification of cases? After all, it is a mark of intelligence that people question what they have learned and, where appropriate, revise their initial classifications. Particularists would appear to be precluded from talking about a clash or conflict in principles leading to a need for revision since that would make them appear to reintroduce the importance of principles. But if principles conflicting does not lead to a need to revise our classification of cases, then what does? And how do we carry out revisions if principles are not involved and a look-up table is too inefficient?

Jonathan Dancy (1998) has suggested that some of the concerns pertaining to learning and generalizing without principles might be profitably explored with neural network models of cognition. In the next section, two neural network models for classifying cases, generalizing to new cases, and reclassifying the original cases will be presented. The models are crude, however their point is not to tell the final story of this matter. Rather, it will be to show that while particularism may be more plausible then it might appear at first, important considerations suggest that it may not be the whole story about moral reasoning. Moreover, the very lines of the debate between particularists and generalists will become blurred.

Moral Case Classifiers

Artificial Neural Networks (ANNs): Feedforward Net. The first ANN considered in this section is a three layer, fully interconnected feed forward net. See figure 1. It was trained on cases that described instances of either killing or allowing to die. Every case involves an actor (the person doing the killing or allowing the dying), an action (killing or allowing to die), and a recipient (the person being killed or allowed to die). With the feed forward (FF) classifier, it is possible to specify one intention and one consequence. The first table in the appendix lists the intentions and consequences used in the simulation.

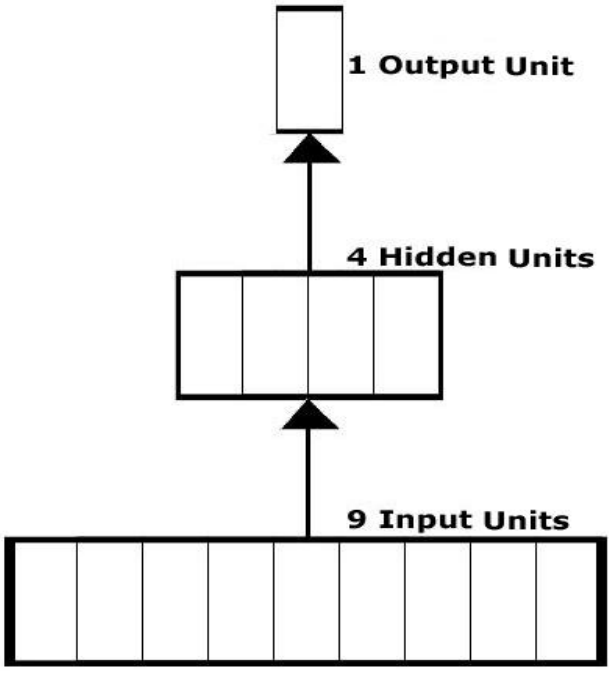

Figure 1

The network was trained on 22 cases and tested on 64 cases. A complete list of training and testing cases is included in the appendix. Training cases included the following.

Input: Jill kills Jack in self-defense. Output: Acceptable.
Input: Jack allows Jill to die, extreme suffering is relieved. Output: Acceptable.

Interestingly enough, while the training set did not include any "suicide" cases (Jack kills Jack, or Jill kills Jill), the network generalized and replied plausibly to these (as well as other cases). Moreover, the training set did not include a single case that contained both a motive and a consequence. In spite of this, the trained net generalized well to cases that had both a motive and a consequence.

There are a variety of different strategies that human agents use to examine their views on moral cases to consider the possibility of revision. One of these strategies makes use of contrast cases. To see an example of how this might work, consider a (modified version) of a case provided by Judith Thomson (1971). There is a world famous violinist who is dieing of a kidney ailment. You're the only person around whose kidneys could filter his blood. The society of music lovers kidnaps you, knocks you unconscious, hooks you up to the violinist, and you awake to discover that your kidneys are filtering his blood. A doctor apologizes, says that the hospital had nothing to do with this, and informs you that you are free to disconnect yourself and walk away. Doing so, without reconnecting yourself, means that the violinist will die within a week or so. You would have to stay connected for about nine months before the appropriate machines could be brought in to keep him alive. Thomson has (as do most people) the intuition that it is morally permissible to disconnect yourself and walk away, even if not reconnecting yourself means the violinist will die. Thomson suggests that this case is analogous to a woman who has become pregnant as the result of rape and is seeking an abortion. In both cases, one life has been made dependent on another through force, and if it is morally acceptable not to sustain the violinist, then it is morally acceptable not to sustain the fetus. There are a number of ways to challenge this analogy. One strategy is to find some difference between the cases that requires treating them in different ways. For example, someone might say that in the violinist case, one does not kill the violinist by disconnecting oneself and walking away; one merely allows him to die. In the case of abortion (even is cases where the pregnancy resulted from rape), one is killing. To this, it might be added that there is a morally relevant difference between killing and allowing to die – where *killing* is thought to be impermissible and *allowing to die* thought to be permissible. How does one challenge this distinction? The method of contrast cases is one way to challenge it. Let us examine how it works.

To test a distinction, one finds two cases that are identical except for the feature being tested, and one sees if varying that feature makes a difference. For example, consider the following cases. **Case A**: Jack is insane and is beating up on Jill for no good reason and wants to kill her by shooting her, so Jill kills Jack in self-defense. **Case B**: Jack is insane and is beating up on Jill for no good reason and wants to kill her by shooting her, but the gun is facing the wrong way, so Jill allows Jack to accidentally shoot himself (allowing him to die). The intuition of many is that (other things being equal) in both cases, Jill behaved in a morally permissible manner, and that, at least in this pair of cases, killing versus allowing to die is a factual difference that makes no moral difference. In the case coding scheme for the network discussed above, the two cases in this paragraph could be coded as follows:

Case A. Input: Jill kills Jack in self-defense. Output: permissible. (Training case No. 1 in the appendix.)

Case B. Input: Jill allows Jack to die in self-defense. Output: permissible. (Testing case No. 5 in the appendix.)

One set of contrast cases is not decisive, but if one cannot find any contrast cases to suggest that a feature is relevant to moral evaluation, then that suggests that a revision regarding that distinction may be in order.

Let us say we want to model the views of an individual who believes that (i) it is morally permissible to disconnect and not reconnect yourself to the violinist, and (ii) it is morally impermissible to have an abortion in cases where the pregnancy results from rape. The respective representations for the two cases just mentioned are as follows:

Case C. Input: Jill allows Jack to die; freedom from an imposed burden results. Output: permissible. (Training case No. 13 in the appendix.)

Case D. Input: Jill kills Jack; freedom from an imposed burden results. Output: impermissible. (Training case No. 12 in the appendix.)

The network described above can be trained so that the inputs for **Cases A** through **D** yield the stated outputs. A second net (figure 2) can be trained that takes as its inputs both the outputs and inputs of the first net. This second net (or Meta-net) takes pairs of cases together with their classifications (acceptable or unacceptable) as inputs, and if the two cases are identical in all respects except one, and if the initial classifications of these cases differ, then the output neuron fires to indicate that this pair of cases is a pair of contrast cases. Meta-net flags

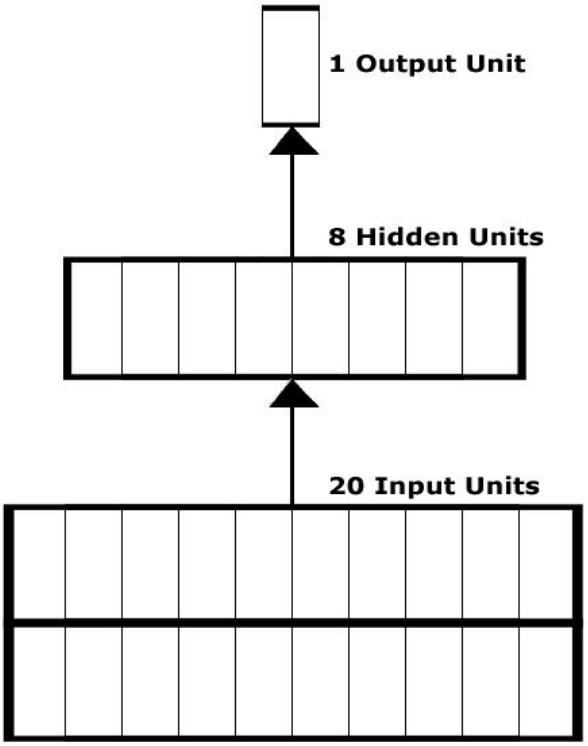

Figure 2

pairs of cases where one feature appears to make a relevant difference between the two cases.

If a distinction is purported to make a difference in *only* one pair of cases, then it is hard to believe that distinction carries any weight. This can be seen in the common practice of looking for other pairs of cases where a distinction purportedly makes a difference. If such cases cannot be found, then a revision may be in order with respect to the original pair of cases. The training cases were designed so that the distinction between killing and allowing to die only made a difference in **Case C** and **Case D**. Other pairs of cases involving killing and allowing to die as the only difference can be tested with Meta-net, and in those cases, the distinction between killing and allowing to die makes no difference. Finding this, we might then revise our judgment on **Case D** to *permissible*. After this change is made in the training set, the FF net can be trained to get the new answer and preserve the answers on the other cases.

A Simple Recurrent Network. Real situations often have more than one motive or one consequence. The only way to accommodate multiple motives and consequences with a simple feed forward net is to keep expanding the size of the input layer. A simple recurrent net (see Elman 1990 for a discussion of this type of net) on the other hand, can accommodate multiple motives and consequences more straightforwardly. The second ANN model I want to consider in this section is simple recurrent (SR) classifier of cases (figure 3). Unlike the FF classifier, the SR classifier takes the input of a case sequentially. For example, the FF classifier will receive as input all of *Jill kills Jack in self-defense* at time t_1. The SR classifier receives *Jill* as input at t_1, *kills* at t_2, *Jack* at t_3, and *in self-defense* at t_4.

When the SR classifier was trained on exactly the same cases as the FF classifier, it generalized just as well (and trained in fewer epochs than the FF classifier). Moreover, even though it did not have a single example of multiple motives or consequences in the training set, it was able to generalize to some cases involving multiple motives and consequences. For example, it provided the following input-output mappings.

> Input: Jill kills Jack in self-defense; freedom from imposed burden results. Output: acceptable.
> Input: Jill allows Jack to die to make money; many innocents suffer. Output: unacceptable.
> Input: Jack kills Jill out of revenge; many innocents suffer. Output: unacceptable.
> Input: Jack allows Jill to die out of revenge; many innocents suffer. Output: unacceptable.

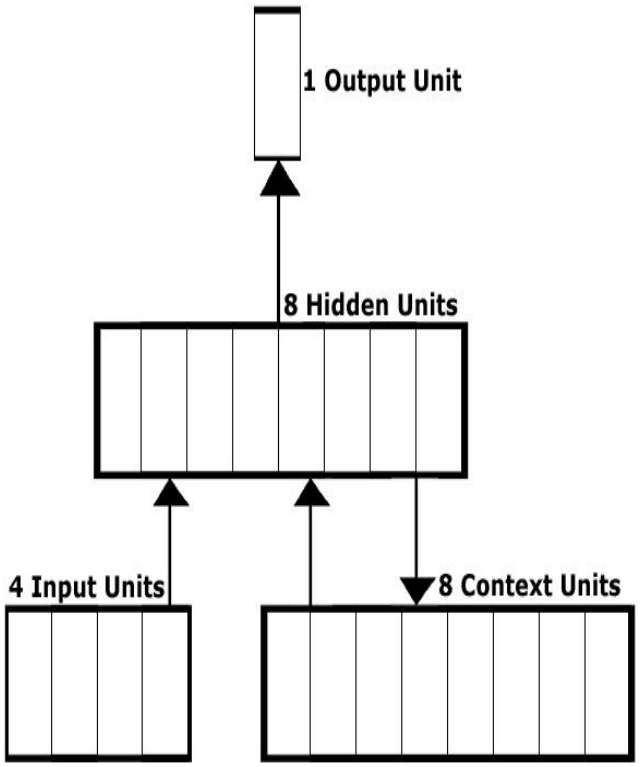

Figure 3

However, the SR classifier frequently errs on cases with multiple motives and consequences unless the training set is expanded. More work is required to improve the performance of the SR classifier, and this will likely involve increasing the size of the training set. The SR classifier also has problems with cases where an acceptable motive is mixed with an unacceptable consequence, or an unacceptable motive is mixed with an unacceptable consequence. I am assuming that in order for an action to be overall acceptable, both the motivation and the consequence must be acceptable. Once again, expansion of the training set would appear to be the solution since it does not currently contain cases where there is a mix of acceptable and unacceptable components. Thus far, we have not had success constructing a meta-net for the SR classifier, but work continues on this front.

3. Assessment

Can any of the above be seen as support for particularism? Well, yes and no. To begin, let us consider two different ways a system or entity may be said to follow a rule or law. (Laws will be treated as a type of rule.)

R1: The Earth is following the law of gravity as it orbits the sun.

R2: The judge is following the law of his jurisdiction in rendering his verdict.

In **R2**, we might say that a rule is being *consulted*, but in **R1** we would not say that the Earth consulted the law of gravity. Nor would we say that the Earth is *executing* a law. Mere agreement with a rule or law (as in **R1**) is not the same as executing or consulting a rule (which we find in **R2**). This distinction is relevant since it can plausibly be argued (Guarini 2001) that ANNs of the types discussed in the previous section are not executing or consulting exceptionless rules of the type we would express in natural language. To be sure, the system that trained the net was *executing* the backpropagation algorithm, but that is not the same thing as executing or consulting a rule like, "Killing always contributes to the wrongness of an action." It is rules of this latter type that do not appear to have been executed or consulted in the training of the net. However, without such rules, the net not only trained but generalized well on the test cases. Let us assume that such nets could be scaled-up to handle the vast number of cases that humans can handle. This assumption is being made *not* because I think it is plausible, but simply to see what would follow from it. At best, we could say that the learning of classifications on training cases and generalization to new cases does not involve consulting or executing moral rules; it does not follow that moral rules are not true. For example, consider the following rules:

R3: Killing always contributes to the moral wrongness of an act.

R4: A bad intention is sufficient for making an act morally unacceptable.

Note: **R3** does not say that an act that involves killing is always wrong. It says that killing always contributes to the wrongness of an act, but it is left open that some other feature (such as saving the lives of many innocents) may outweigh the wrongness contributed by killing and, all things considered, make an act acceptable even though it involves killing. It could be argued that the behaviour of the ANNs in the previous section is *in agreement* with **R3** and **R4**, and **R3** and **R4** may very well be true, and this in spite of the fact that these nets do not *execute* or *consult* **R3** and **R4**. This distinction between being in agreement with a rule and consulting or executing a rule muddies the distinction between particularism and generalism, for it allows us to see how particularists *might* be right that we can learn and generalize without making use of (in the sense of consulting or executing) rules, but it also allows us to see how a generalist might then go on to insist that moral rules are, nonetheless, true (and that a system's behaviour may even be in agreement with such rules).

Thus far, comments have been made on the first two of the three questions raised at the beginning of section two, **Challenges of Particularism**. Some remarks are needed on the third concern (which is really two questions): how do we recognize that our initial classification of cases needs revision, and how do we carry out that revision? Even if neural nets could be scaled up and continue to learn and classify moral situations without consulting or executing moral rules, such rules may be involved in the process of reflecting on the initial classification of cases. For example, someone who is presented with the violinist case discussed above (**Case C**) may put the following rule to the test.

R5: Killing always counts against doing an act, while allowing someone to die does not count against doing an act.

The process of reflecting on initial classifications may require the representation of contributory rules or principles. This is not to say that all reflection would involve the representation of such rules, but at least some reflection in humans does appear to require the explicit representation or consultation of such rules. Sometimes the representation of such rules will lead to their acceptance, and sometimes to their rejection. It is important to note that even in those cases where the explicit consultation of a rule leads to the rejection of the rule and subsequent revisions in the initial classification of

cases (which is what we considered in the previous section), *that rule played an important role in reasoning*. Moreover, the rejection of an explicitly consulted rule may lead to the acceptance of another rule, such as the following.

R6: Killing and allowing to die always count against doing an act.

Understood as a contributory rule, **R6** does not mean that any act involving killing or allowing to die is always unacceptable since some other feature of the situation may outweigh the unacceptability of killing or allowing to die.

Thus far, the term "learning" has been reserved for the initial classification of cases, and "revision" has been reserved for the reflective process of changing the initial classification of cases. However, when humans revise their views on cases, they are often said to have *learned* something. This is a perfectly common and acceptable use of the term "learned." The restriction of "learning" and its variants in this paper to the initial classification of cases has been strictly for the purpose of making clearer different moments in moral cognition. However, the restriction is artificial, and no attempt is being made to assert that learning involves only the initial classification of cases.

4. Context and Future Work

Perhaps the most obvious limitation of the type of neural network models discussed herein is their inability to generate arguments as outputs. Generating an argument requires the ability to produce a sequence of sentences. Above, both a net and a meta-net were considered, but an outside agent was governing the application of the meta-net and engaging in the reasoning that lead to the application of the meta-net. A more developed model would include the ability to represent to itself rules like R5 and R6, and the ability to reason about such rules to engage in the reclassification of cases. Ideally, the model would also be able to produce its reasons in the form of an argument, citing rules where needed. Moreover, the ability to receive arguments as input and process them – including claims about rules – would also be required. An account of the classification of cases is only part of the story of moral reasoning, re-classification under the pressure of objections is also an important part, and this latter part is mediated by language use in ways that existing neural nets have a difficult time handling (though these are still early days in such research).

The views expressed herein are compatible with many of the views expressed in some of Bruce McLaren's work (2003). As McLaren rightly points out, what we want is not just the ability to classify cases, receive arguments, and make arguments, but also the ability to come up with creative suggestions or compromises (in situations that allow for them). However, most work in case-based reasoning in AI starts with an initial set of cases – the initial case base – and treats new cases by comparing them, in some way or other, with one or more cases in the initial case base. Much work in the Law and AI literature (think of the tradition arising from Ashley 1990) proceeds in this way, and McLaren's work grows out of that tradition. The models presented in this paper are quite limited and do not have many of the virtues possessed by models developed in the aforementioned tradition; however, part of the point of this paper has been to motivate the need for constructing models for how we reason about and revise an initial case base. Initial case bases tend to be treated as "given" in two respects. First, the system does not have to learn how to treat the initial cases; their proper treatment is simply built into the system. Second, the system does not have any way of dealing with challenges to the initial case base; it is assumed that the system will not have to defend or modify its treatment of those cases. It is a mark of intelligent beings that they do not treat cases as given in either of these two senses. More work is needed on expanding existing models or constructing new models that move beyond the two-fold givenness of the initial case base. It is in considering how to revise cases that principles (whether they are called contributory, open textured, extensionally defined, *et cetera*) will likely loom large. Consulting or rendering such principles explicit appears to be an important part of moral reasoning. While particularism in some qualified sense *may* be part of the story in modeling moral reasoning, it is likely not the whole story. Making contributory principles explicit is an important part of moral reasoning.

Principles may also be an important part of understanding moral psychology. When someone is conflicted about killing a person to save many other innocent persons, that may be because there is (a) a general obligation not to kill, and (b) a general obligation not to let harm come to the innocent. When it is not possible to satisfy both obligations, we are conflicted, and rightly so. This does not mean that we need to be paralyzed into inaction; one can go on to argue that one obligation may trump the other, but it does not follow that the obligation being trumped has no force. Indeed, it is precisely because both obligations do retain force that we are *rationally* conflicted.

Continuing on the theme of moral psychology, it should be noted that this paper has assumed that both intentions and consequences are an important part of moral reasoning. That the consequences of an action are an important part of moral reasoning requires that an intelligent agent be able to reason about the causal effects of his, her, or its actions. This is a rather obvious point. A point that may not be as obvious is that an intelligent agent has to be able to reason about the intentions of other agents. Much literature in

philosophy and psychology has been devoted to so called "mind reading" (which is not intended in the psychic sense but simply as a way of indicating that we often can figure out what is on the minds of other agents). This literature is relevant to Machine Ethics since the moral assessments of intelligent agents makes use of the states of mind of beings whose actions are being assessed. Say that Huey's hand hits Louy in the stomach because Huey has a neurological disorder that causes him to twitch uncontrollably; say Dewy hits Louy in the stomach because Dewy wants to wipe the grin off of Louy's face. (Imagine the Louy is grinning because he recently won a grant.) In both cases we have the same behaviour (a hand hitting someone in the stomach) but Dewy's motive is nasty, and Louy would be right to treat Huey and Dewy differently. The work in this paper *presupposes* that some solution is forthcoming to the problem of figuring out the states of minds of other agents. Again, this is another area in which much work needs to be done.

Finally, there is the issue of what we are reasoning about. As was mentioned earlier, we can reason about actions or other types of states of affairs, and we can reason about persons or other types of entities (clubs, businesses, …). While this paper has focused primarily on actions, many of the points made about moral reasoning as it pertains to action *may* apply to the other possible objects of moral reasoning. This is a matter requiring further investigation. Since the moral reasoning of intelligent beings is about more than actions, adequate models of such intelligence will need to include, but also go beyond, the consideration of actions.

5. Acknowledgements

I thank the Social Sciences and Humanities Research Council of Canada for financial support during the research and writing of this paper. I also thank Pierre Boulos for comments on an earlier version of this work; Sulma Portillo for assistance in coding the simulations and proof reading this work; Terry Whelan for assistance in proof reading, and Andy Dzibela for assistance in putting together the figures.

Appendix

Case Terms for Inputs

Agents	Actions	Motives	Consequences
Jack	Kills	Self-Defense	Freedom (of the actor/agent) from imposed burden
Jill	Allows to die	To make money	Extreme suffering (of the subject/recipient) is relieved
		Revenge	Lives of many innocents (other than the actor and subject) are saved
		Eliminate Competition	Many innocents (other than the actor and subject) die
		Defend the innocent	Many innocents (other than the actor and subject) suffer

Outputs: A=morally acceptable; U=morally unacceptable

Initial Training Cases

No.	Input: Case Description	Output
1	Jill kills Jack in self-defense	A
2	Jack kills Jill in self-defense	A
3	Jack allows Jill to die in self-defense	A
4	Jill kills Jack to make money	U
5	Jack kills Jill to make money	U
6	Jack allows Jill to die to make money	U
7	Jack kills Jill out of revenge	U
8	Jill allows Jack to die out of revenge	U
9	Jack kills Jill to eliminate competition	U
10	Jill allows Jack to die to eliminate competition	U
11	Jill kills Jack to defend the innocent	A
12	Jill kills Jack; freedom from imposed burden results	U
13	Jill allows Jack to die; freedom from imposed burden results	A
14	Jack allows Jill to die; freedom from imposed burden results	A
15	Jack kills Jill; many innocents suffer	U
16	Jill kills Jack; lives of many innocents are saved	A
17	Jill allows Jack to die; lives of many innocents are saved	A
18	Jack allows Jill to die; lives of many innocents are saved	A
19	Jill kills Jack; many innocents die	U
20	Jack allows Jill to die; many innocents die	U
21	Jill kills Jack; extreme suffering is relieved	A
22	Jack allows Jill to die; extreme suffering is relieved	A

Initial Testing Cases

No.	Input: Case Description	Output
1	Jill kills Jack	U
2	Jack kills Jill	U
3	Jill allows Jack to die	U
4	Jack allows Jill to die	U
5	Jill allows Jack to die in self-defense	A
6	Jill allows Jack to die to make money	U
7	Jill kills Jack out of revenge	U
8	Jack allows Jill to die out of revenge	U
9	Jill kills Jack to eliminate competition	U
10	Jack allows Jill to die to eliminate competition	U
11	Jack kills Jill to defend the innocent	A
12	Jill allows Jack to die to defend the innocent	A
13	Jack allows Jill to die to defend the innocent	A
14		
15	Jack kills Jill; freedom from imposed burden results	U
16	Jill kills Jack; many innocents suffer	U
17	Jill allows Jack to die; many innocents suffer	U
18	Jack allows Jill to die; many innocents suffer	U
19	Jack kills Jill; lives of many innocents are saved	A
20	Jack kills Jill; many innocents die	U
21	Jill allows Jack to die; many innocents die	U
22	Jack kills Jill; extreme suffering is relieved	A
23	Jill allows Jack to die; extreme suffering is relieved	A
24	Jill kills Jack in self-defense; freedom from imposed burden results	A
25	Jack kills Jill in self-defense; freedom from imposed burden results	A
26	Jill kills Jack in self-defense; lives of many innocents are saved	A
27	Jill allows Jack to die in self-defense; lives of many innocents are saved	A
28	Jack allows Jill to die in self-defense; lives of many innocents are saved	A
29	Jill kills Jack to defend the innocent; lives many innocents are saved	A
30	Jill allows Jack to die to defend the innocent; lives of many innocents are saved	A
31	Jack kills Jill to defend the innocent; lives of many innocents are saved	A
32	Jack allows Jill to die to defend the innocent; many innocents are saved	A
33	Jill kills Jack to make money; many innocents suffer	U
34	Jack kills Jill to make money; many innocents suffer	U
35	Jill allows Jack to die to make money; many innocents suffer	U
36	Jack allows Jill to die to make money; many innocents suffer	U
37	Jack allows Jill to die out of revenge; many innocents die	U
38	Jill allows Jack to die out of revenge; many innocents die	U
39	Jill kills Jack in self-defense; extreme suffering is relieved	A
40	Jack kills Jill in self-defense; extreme suffering is relieved	A
41	Jill allows Jack to die in self-defense; extreme suffering is relieved	A
42	Jack allows Jill to die in self-defense; extreme suffering is relieved	A
43	Jill kills Jack out of revenge; many innocents suffer	U
44	Jack kills Jill out of revenge; many innocents suffer	U
45	Jill allows Jack to die out of revenge; many innocents suffer	U
46	Jack allows Jill to die out of revenge; many innocents suffer	U
47	Jill kills Jill	U
48	Jack kills Jack	U
49	Jill allows Jill to die	U
50	Jack allows Jack to die	U
51	Jill kills Jill; many innocents die	U
52	Jill kills Jill; lives of many innocents are saved	A
53	Jack kills Jack; many innocents die	U
54	Jack kills Jack; lives of many innocents are saved	A
55	Jack allows Jack to die; lives of many innocents are saved	A
56	Jill allows Jill to die; lives of many innocents are saved	A
57	Jill kills Jill in self-defense; extreme suffering is relieved	A
58	Jack kills Jack in self-defense; extreme suffering is relieved	A
59	Jill kills Jill in defense of the innocent; lives of many innocents are saved	A
60	Jack kills Jack in defense of the innocent; lives of many innocents are saved	A
61	Jack kills Jill out of revenge; lives of many innocents are saved	A
62	Jill kills Jack out of revenge; lives of many innocents are saved	A
63	Jack kills Jill to make money; lives of many innocents are saved	A
64	Jill kills Jack to make money; lives of many innocents are saved	A

Both the FF and SR nets erred on cases 61-64. Note that in the training set, there were no cases containing both acceptable and unacceptable components. As indicated in the text, it is being assumed that in order for a case to be overall acceptable, both the motive and consequence need to be acceptable.

Specifications for the Feed Forward Classifier
Trained and tested on the above cases.
Number of input units: 9.
Number of hidden units: 4.
Number of output units: 1.
Learning rate: 0.1.
Momentum: 0.9.
Escape Criterion: sum of squared errors ≤ 0.1.
Epochs to train: 3360.

Specifications for Meta-Net
Due to space restrictions, the training and testing pairs for this net could not be included.
Number of input units: 20. 10 units for each case, where 9 units describe the case, and one unit describes its classification (acceptable or unacceptable). The net takes as input two cases and a classification for each (where that classification was delivered as output by the FF classifier).
Number of hidden units: 8.
Number of output units: 1.
Learning rate: 0.1.
Momentum: 0.9.
Escape Criterion: sum of squared errors ≤ 0.1.
Epochs to train: 540. There are still problems with generalization. The net delivers false positives: some cases are classified as contrast cases when they should not be.

Specifications for the Simple Recurrent Classifier
Trained and tested on the above cases.
Number of Input units: 4.
Number of Context units: 8
Number of hidden units: 4.
Number of output units: 1.
Learning rate: 0.01.
Momentum: 0.9.
Escape criterion: count the number of times the sum of squared error for a sequence is ≥ 0.05. When the number of sequences satisfying the preceding condition = 0, stop training. An event consists of one term ("Jack" or "allows to die" . . .), and a sequence consists of a complete case ("Jack kills Jill in self defense").
Epochs to train: 633.

Simulation Software
All simulations were run on the PDP++ simulator (version 3.1). The generalized delta rule for backpropagation was used for all training. See O'Reilly and Munakata (2000) for discussion of the simulator and possible applications. The simulator can be freely obtained from the following site:
http://www.cnbc.cmu.edu/Resources/PDP++//PDP++.html

References

Ashley, Kevin D. 1990. *Modelling Legal Argument: Reasoning by Cases and Hypotheticals*. Cambridge, MA & London, England: MIT Press, a Bradford Book.

Dancy, J. 1993. *Moral Reasons*. Oxford: Blackwell.

Dancy, Jonathan. 1998. Can a Particularist Learn The Difference between Right and Wrong? In *Proceedings from the 20th World Congress of Philosophy, Volume I: Ethics*, pp. 59-72, Klaus Brinkmann, ed. Bowling Green, Ohio: Philosophy Documentation Center.

Dancy, J. 2000. The Particularist's Progress, in *Moral Particularism*, Hooker, B., and Little, M. eds. Oxford: Oxford University Press.

Elman, Jeffery. 1990. Finding Structure in Time. *Cognitive Science* 14: 179-211.

Guarini, M. 2001. A Defence of Connectionism and the Syntactic Objection. *Synthese* 128: 287-317.

Hooker, B., and Little, M. eds. 2000. *Moral Particularism*. Oxford: Oxford University Press.

McKeever, S., and Ridge, M. 2005. The Many Moral Particularisms. *Canadian Journal of Philosophy* 35: 83-106.

O'Reilly, R., and Munakata, Y. 2000. C*omputational Explorations in Cognitive Neuroscience*. Cambridge, Mass.: The MIT Press, a Bradford book.

Thomson, J. J. 1971. A Defense of Abortion. *Philosophy and Public Affairs* 1: 47-66.