Computational Neural Modeling and the Philosophy of Ethics

Reflections on the Particularism-Generalism Debate

Marcello Guarini

Introduction

There are different reasons why someone might be interested in using a computer to model one or more dimensions of ethical classification, reasoning, discourse, or action. One reason is to build into machines the requisite level of "ethical sensitivity" for interacting with human beings. Robots in elder care, nannybots, autonomous combat systems for the military – these are just a few of the systems that researchers are considering. In other words, one motivation for doing machine ethics is to support practical applications. A second reason for doing work in machine ethics is to try to better understand ethical reasoning as humans do it. This paper is motivated by the second of the two reasons (which, by the way, need not be construed as mutually exclusive).

There has been extensive discussion of the relationship between rules, principles, or standards, on the one hand, and cases on the other. Roughly put, those stressing the importance of the former tend to get labeled generalists, whereas those stressing the importance of the latter tend to get labeled particularists. There are many ways of being a particularist or a generalist. The dispute between philosophers taking up these issues is not a first-order normative dispute about ethical issues. Rather, it is a second-order dispute about how best to understand and engage in ethical reasoning. In short, it is a dispute in the philosophy of ethics.[1] This paper will make use of computational neural modeling in an attempt to scout out some underexplored conceptual terrain in the dispute between particularists and generalists.[2]

[1] The expression "meta-ethics" could be used in place of "philosophy of ethics." However, some hold on to a restricted conception of meta-ethics, associating it with the methods and approaches of analytic philosophers of language (especially of the first half of the twentieth century). To avoid any misunderstandings, I have used the expression "philosophy of ethics" to indicate any second-order inquiry about first-order ethics. Jocelyne Couture and Kai Nielsen (1995) provide a very useful survey of the history of meta-ethics, including its broader more recent uses.

[2] Whereas Horgan and Timmons (2007 and 2009) characterize their position as "core particularism," I read it as an attempt to search out the underexplored middle ground between the more

https://doi.org/10.1017/CBO9780511978036.023 Published online by Cambridge University Press

The next section will lay down some terminology that will be used throughout the rest of the paper. Part three will lay out some of the logically possible options available with respect to learning; part four will lay out some of the options available with respect to establishing or defending the normative status of ethical claims. Part five will provide a preliminary analysis of some of the options available to particularists and generalists, so that in part six we can look at and analyze neural networks trained to classify moral situations. Parts six and seven will explore some of the middle ground between the more thoroughgoing forms of particularism and generalism. Nothing in this paper should be read as an attempt to administer knock-down blows to other positions. I quite deliberately bill this work as exploratory. There are empirical assumptions at work in discussions between particularists and generalists, and it is early days still in understanding the strengths and weaknesses of various computational models and in empirical research on human cognition. Clarifying what some of the options are and showing how computational neural modeling may help us to see options that may have otherwise gone unconsidered are the main goals of the paper.

Some Terminology

As alluded to in the introduction, there are many forms of particularism and generalism. They can be understood in terms of the approach they take toward principles. One useful distinction between different types of principles is that between the exceptionless or total standard and the contributory standard.[3] The total standard provides a sufficiency condition for the application of an all-things-considered moral predicate. For example, "Killing is wrong" can be interpreted as providing a sufficiency condition for applying the predicate "wrong," all things considered. This would suggest that killing is wrong in all circumstances. Alternatively, the same claim could be interpreted as a contributory standard. The idea here would be that killing contributes to the wrongness of an action, but other considerations could outweigh the wrongness of killing and make the action, all things considered, morally acceptable. To say that killing contributes to the wrongness of an action is not to say that in any given case, all things considered, the action of killing is wrong. In other words, the contributory standard does not supply a sufficiency condition for the application of an all-things-considered moral predicate in a given case.

Standards, whether total or contributory, can be classified as thick or thin. In a thick standard, a moral predicate is explicated using, among other things, another moral predicate. In a thin standard, a moral predicate is explicated without the use of other moral predicates. "If you make a promise, you *ought* to keep it" is

thoroughgoing versions of particularism and generalism (because they try to preserve some of the insights of particularism without denying some role for generality).

[3] McKeever and Ridge (2005) provide a brief and very useful survey of the different types of standards and the different types of particularism and generalism.

https://doi.org/10.1017/CBO9780511978036.023 Published online by Cambridge University Press

thin because what you ought to do is explained without the use of another moral predicate. "If you make a promise you *ought* to keep it, unless you promised to do something *immoral*" is thick.[4]

Jonathan Dancy (2006) is arguably the most thoroughgoing particularist around. He rejects the need for all standards, whether total or contributory, thick or thin. Not all particularists go this far. Garfield (2000), Little (2000), and McNaughton and Rawling (2000) all consider themselves particularists and find acceptable the use of thick standards; what makes them particularist is that they reject thin standards. Generalists like Jackson, Petit, and Smith (2000) insist on thin standards. Being open to both thick and thin standards would be to occupy a middle ground between many particularists and generalists. Guarini (2010) is an example of this sort of position. As we will see in parts six and seven, there may be other ways to occupy a middle ground.

Some Options with Respect to Learning

This section will ask a question (Q) about learning (L), and some possible answers (A) will be outlined. The purpose here is not to catalog every possible answer to the question, but to give the reader a sense for what different answers might look like. The same will be done in the next section with respect to understanding the normative statuses of cases and rules. After doing this, we will be in position to explore a key assumption of some of the answers.

LQ: With respect to learning, what is the relationship between cases and rules?

There are a number of possible answers to this question. The answer of the most unqualified of particularists would be as follows.

LA1: Rules do not matter at all. They are simply not needed. This view applies to both total and contributory standards, whether thick or thin.

We can imagine variations on LA1 where contributory standards are considered important but not total standards (or vice versa), but as I have already stated, it is not my goal here to catalog all possible replies.

LA2: During the learning process, we infer rules from cases.

Whether LA2 is particularist or generalist will depend on how it is developed. Consider two variations.

LA2A: During the learning process, we infer rules from cases. These rules, though, do not feed back into the learning process, so they play no essential role in learning. They are a kind of summary of what is learned, but they are not required for initial or further learning.

LA2B: During the learning process, we infer rules from cases. These rules do feed back into the learning process and play a central role in the learning of further cases and further rules.

[4] This is a very quick explanation. "Particularism, Analogy, and Moral Cognition" contains a more detailed discussion of thick and thin standards, including a distinction between a cognitively constrained conception of these terms and a more purely metaphysical conception.

https://doi.org/10.1017/CBO9780511978036.023 Published online by Cambridge University Press

Clearly, LA2a is thoroughly particularist. There is a way for LA2b to be particularist (but not to the extent of LA2a): Insist that the rules being learned and feeding back into the learning process are all thick. If it turns out that the rules feeding back into the learning process are all thin, then we have a generalist account of learning. An even stronger generalist account is possible, but this takes us beyond LA2, which assumes that we do not start with substantive rules. Let us have a brief look at this more ambitious generalist position.

LA3: When we start learning how to classify cases, we are using innate rules. We infer further rules when exposed to enough cases, and these further rules feed back into the learning process and play a central role in the learning of further cases and further rules.

Again, there are different ways in which this position might be developed. Provided the innate rules are thin, substantive rules, the position is a very thoroughgoing form of generalism. Even if the rules are not substantive but constitute a kind of grammar for learning to classify moral situations, it would still be generalist. If the innate rules are thick, then we have a particularist position. Variations on innatism (of which LA3 is an instance) will not be explored in any detail herein, so I will not comment on it at length here. LA3 was introduced to provide a sense for the range of options that are available.

Let us return to the variations on LA2. Perhaps some of the rules that feed back into the learning process are thick, perhaps some are thin (which would be a variation of LA2b). If that were so, then a hybrid of particularism and generalism would be true (at least with respect to learning). Other hybrids are possible as well. For example, perhaps some of the rules we learn are thin and feed back into the learning process (a generalist variation on LA2b), and perhaps some rules we learn function as convenient summaries but do not feed back into the learning process (LA1). Again, I am not going to enumerate all the possible hybrid approaches that might be defended. As we will see, there are other possibilities.

Some Options with Respect to Normative Standing or Status

Let us pursue the strategy of question and answers with respect to normative (N) statuses.

NQ: With respect to the normative standing or status of cases and rules, what is the relationship between cases and rules? Let us take moral standing or status to refer to things like moral acceptability, permissibility, rightness, goodness, obligatoriness, virtuousness, and supererogatoriness (or their opposites).

NA1: Rules do not matter at all. When establishing moral standing or status, we need not appeal to substantive rules of any sort.

NA2: All morally defensible views on cases must be inferred from one or more valid thin general rules.

NA1 is very strong form of particularism, and NA2 is a very strong form of generalism. A position somewhere between these polar opposites is possible.

https://doi.org/10.1017/CBO9780511978036.023 Published online by Cambridge University Press

NA3: Sometimes the moral standing or status of a rule is established by reasoning from a particular case, and sometimes the standing or status of a case is appropriately established or revised by reasoning from one or more general thin rules.

NA3 is a hybrid of NA1 and NA2.

Hybrid positions sometimes seem strange or unprincipled, or perhaps blandly ecumenical. It is, of course, completely legitimate to ask *how* such positions are possible. What is moral cognition that sometimes we overturn a view on a case by appealing to a rule, and sometimes we overturn a view on a rule by appealing to a case? Someone defending a hybrid position with respect to the normative status of cases and rules should have something to say about that. Someone defending a hybrid view on learning should also have something to say about the nature of moral cognition that makes that sort of learning possible.

Preliminary Analysis of the Answers to LQ and NQ

Let us have a closer look at the first two answers (NA1 and NA2) to the normative question (NQ). NA1 and NA2 are opposites, with NA1 claiming that rules do not matter and NA2 claiming that they are essential. We could easily see how proponents of either view may be shocked by proponents of the other. The particularist takes it that cases are primary, and the generalist takes it that thin rules are primary. The debate between these two types of positions could come down to answering a question like this: With respect to the justification of rules and cases, which is primary or basic? The question *assumes* that one of the two – rules or cases – has to be more basic than the other under all circumstances. If it seems to you like that must be right, then hybrid views like NA3 are going to seem downright puzzling or unprincipled. I want to suggest that it is not obvious whether either one needs to be more basic than the other under all circumstances.

We could engage in the same line of questioning with respect to the learning question (LQ) and the possible answers to it. Some answers might simply assume that one of either rules or cases might be central or essential to learning whereas the other is not. Hybrid positions would seem odd or unprincipled to proponents of such views. Again, I want to suggest that a middle ground is possible. For comparison, consider the following question: Which is more basic to a square, the edges or the vertices? The question is downright silly because it assumes something that is clearly false, that one of either edges or vertices is more basic or important to forming a square than the other. You simply cannot have a square without both edges and vertices. What if the relationship between particular cases and generalities in learning is something like the relationship between edges and vertices in a square? (That is, we need both, and neither can be said to be more basic than the other.) The next section will begin the exploration of this possibility using artificial neural network simulations of moral case classification.

https://doi.org/10.1017/CBO9780511978036.023 Published online by Cambridge University Press

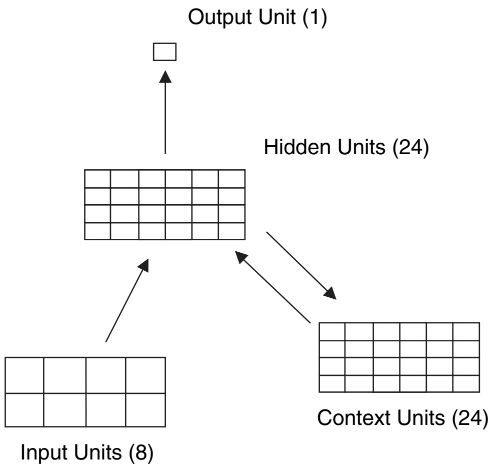

Figure 18.1. Simple Recurrent Network (SRN). Input, Hidden, and Output layers are fully interconnected. Activation flow between hidden and context units is via one-to-one copy connections.

Learning to Classify Moral Situations

The neural networks discussed in this section are all simple recurrent networks (SRNs). The networks all possess exactly the same topology: eight input units, fully interconnected with 24 hidden units, each of which has both a one-to-one connection to the 24 context units and a connection with the one output unit. (See Figure 18.1.) All networks were trained with the generalized delta rule for back-propagating error. The same set of 33 training cases was used for each network.[5] More than 230 testing cases were used. All training and testing cases were instances of either killing or allowing to die. All training cases involved either Jack or Jill being the actor or the recipient of the action, and all training cases involved the specification of at least one motive or one consequence. Table 18.1 provides a sample of the inputs in natural language and the target outputs in parentheses.

One way of presenting information to the network is such that the outputs are ignored until the entire case has been presented. For example, the vector for *Jill* is provided as input, processed at the level of hidden units, copied to the context units, and the target output is 0. Next, the vector for *kills* is fed in as input and sent to the hidden units together with information from the context units; the results are processed and copied back to the context units, and the target output is 0. Next, the vector for *Jack* is provided as input and sent to the hidden units together with

[5] The training cases used in this paper correspond to both training batches A and B in Guarini (2010). A sample of 67 testing cases can also be found in this other work. All training and testing cases are available from the author.

https://doi.org/10.1017/CBO9780511978036.023 Published online by Cambridge University Press

Table 18.1. *Sample cases table (1 = permissible; –1 = impermissible)*

Sample training cases	**Sample testing cases**
Jill kills Jack in self-defense (1)	Jill allows Jack to die in self-defense (1)
Jack allows Jill to die to make money (-1)	Jill kills Jack out of revenge (-1)
Jill allows Jack to die; lives of many innocents are saved (1)	Jill allows Jack to die to make money (-1)
Jack kills Jill to eliminate competition and to make money; many innocents suffer (-1)	Jack kills Jill to defend the innocent; the lives of many innocents are saved (1)
Jack kills Jill out of revenge and to make money; many innocents suffer (-1)	Jill kills Jack to defend the innocent and in self-defense; freedom from imposed burden results, extreme suffering is relieved, and the lives of many innocents are saved (1)

information from the context units; the results are processed and copied back to the context units, and the target output is 0. Next, the vector for *in self-defense* is provided as input and sent to the hidden units together with information from the context units; the results are processed and the target output is 1. That is one way to train the network. Another way is to classify what I call the subcase or subcases that may be present in a longer case. Table 18.2 shows the difference between a case that is trained with the subcases unclassified and the same case trained with the subcases classified. The first column provides the order of input; the second column provides the natural language description of the input; the third column provides the target output when subcases are unclassified, and the final column provides the target output with the subcases classified. An output of 0 indicates uncertain.

Let us consider two simple recurrent networks, SRNa and SRNb. The networks themselves are identical, but SRNa is presented with the training cases such that the subcases are unclassified, and SRNb is presented with the training cases such that the subcases are classified. More specifically, SRNb is trained such that both subcases of the form

x kills y
and
x allows to die y

are classified as impermissible. Using a learning rate of 0.1 and 0.01, SRNa failed to train (even with training runs up to 100,000 epochs), and SRNb trained in a median of 17 epochs using a learning rate of 0.1. Notice that our inputs do not include any sort of punctuation to indicate when the case has ended. If we add the equivalent of a period to terminate the case, then we can get SRNa to train with a learning rate of 0.01 in a median of 2,424 epochs. Clearly, training subcases has its advantages in terms of speed of learning.

https://doi.org/10.1017/CBO9780511978036.023 Published online by Cambridge University Press

Table 18.2. *Unclassified and classified subcases*

Order	**Input**	**Output: subcase unclassified**	**Output: subcase classified**
1st	Jill	0	0
2nd	kills	0	0
3rd	Jack	0	–1
4th	in self-defense	0	1
5th	freedom from imposed burden results	1	1

Let us see what happens if we complicate training by subcases a little more. Say we take an SRN topologically identical to SRNa and SRNb, and we train it on the same set of cases, but when we train by subcases this new network, SRNc, is asked to classify all subcases of the form

x kills y
as impermissible, and all cases of the form
x allows to die y
as permissible.

This complicates the training somewhat, but it is still training by subcases. Using a learning rate of 0.1, SRNc failed to train under 100,000 epochs. Although training by subcases has its advantages (as seen in SRNb over SRNa), complicating the subcases requires complicating the training a bit further. It is possible to get SRNc to train using a learning rate of 0.1, but the technique of staged training needs to be invoked.[6] Training of SRNc is divided into two stages. There are 34 training cases, but during the first stage, only 24 cases are presented to the network; during the second stage, all 34 cases are presented to the network. The 24 cases used in the first stage each have exactly one motive or one consequence, but not both (just like the first three cases in Table 18.1.) The subcases are trained, and the network does train successfully on this smaller, simpler set of cases using a learning rate of 0.1. After SRNc trained on the simple subset, the second stage involves presenting the entire set of 34 training cases, which includes the original simple 24 cases as well as 10 cases with multiple motives or multiple consequences. The fourth and fifth cases in Table 18.1 are examples having more than one motive or consequence. If we sum the total number of epochs for both stages of training, the median number of epochs required to train SRNc is 49. If we use the staged training approach for SRNa with a learning rate of 0.1, it still fails to train with or without stoppers. This suggests that the success in training SRNc is partly due to staged training and partly due to classifying

[6] See Elman 1990 for the pioneering work on staged training of simple recurrent networks.

https://doi.org/10.1017/CBO9780511978036.023 Published online by Cambridge University Press

the subcases. After all, if we used staged training in SRNa and SRNc, the only difference between the two is that SRNc classifies subcases and SRNa does not, yet SRNc trains and SRNa does not.

It is pretty clear that none of the SRNi have been provided with explicit, substantive moral rules as input, whether total, contributory, thick, or thin. However, the case can be made that the behavior of the SRNi is in agreement with contributory standards. There is a distinction that can be made between following a rule as executing or consulting a rule – think of a judge explicitly consulting a statue – and following a rule as simply being in agreement with a rule – think of the planets being (roughly) in agreement with Newton's universal law of gravitation. There are at least two pieces of evidence suggesting that the SRNi are in agreement with contributory standards. First, there is the dummy or blank vector test. If we take a trained network and use vectors it has never seen before to give it the equivalent of

> Jack _____ Jill in self-defense,
> an output of permissible is still returned. If we feed a trained network
> Jill _____ Jack; many innocents die,

an output of impermissible is returned. This is some reason for saying that the networks treat acting in self-defense as contributing to the acceptability of an action, and actions that lead to the deaths of many innocents are treated as contributing to the impermissibility of an action.

There is a second type of evidence that supports the view that the SRNi have behavior in agreement with contributory standards. We can see the final vector for each case produced at the level of hidden units as the network's internal representation of each case. We could plot the value of each hidden unit activation vector in a 24-dimensional state space, but we cannot visualize this. Alternatively, we can do a cluster plot of the vectors. Assuming that we can treat the distance between the vectors as a similarity metric – the further apart two vectors/cases are in state space, the more different they are; the closer two vectors are, the more similar they are – a cluster plot can give a sense of the similarity space the network is working with once it is trained. Figures 18.2 and 18.3 provide cluster plots for the training cases and outputs for SRNb and SRNc respectively (after they have been trained).[7] Notice that even though the outputs for each training case are the same, the cluster plots are quite different. To be sure, we do not get *exactly* the same cluster plot for SRNb every time it is trained on a randomly selected set of weights, and the same is true of SRNc. That said, the tendency for SRNb to group together or treat as similar cases involving killing and allowing death, and the tendency of SRNc to not group together such cases is robust. In other words,

[7] A Euclidean metric is used for distance in these plots. Other metrics are possible, and there is room for argument as to which sort of metric is best to use. However, that is a topic for another paper.

https://doi.org/10.1017/CBO9780511978036.023 Published online by Cambridge University Press

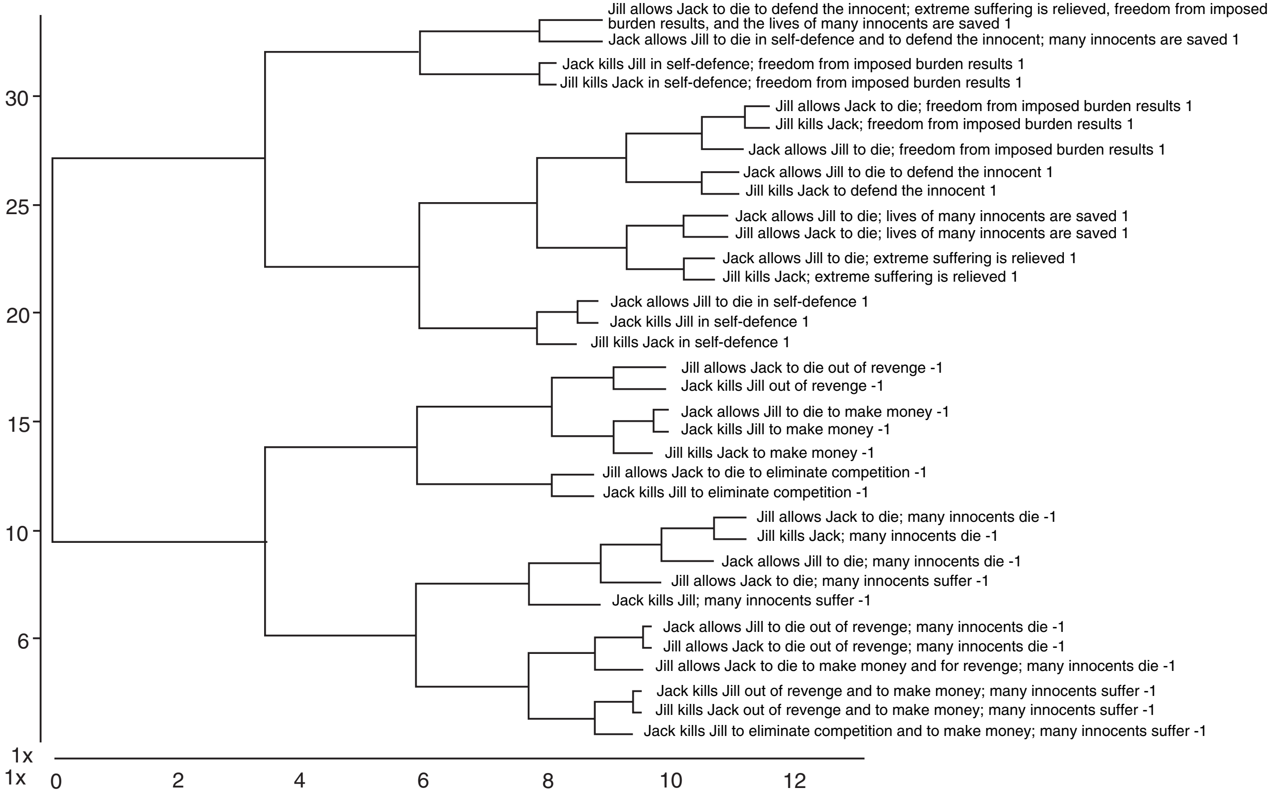

Figure 18.2. SRNb post training cluster plot of hidden unit activation vectors for 34 training cases.

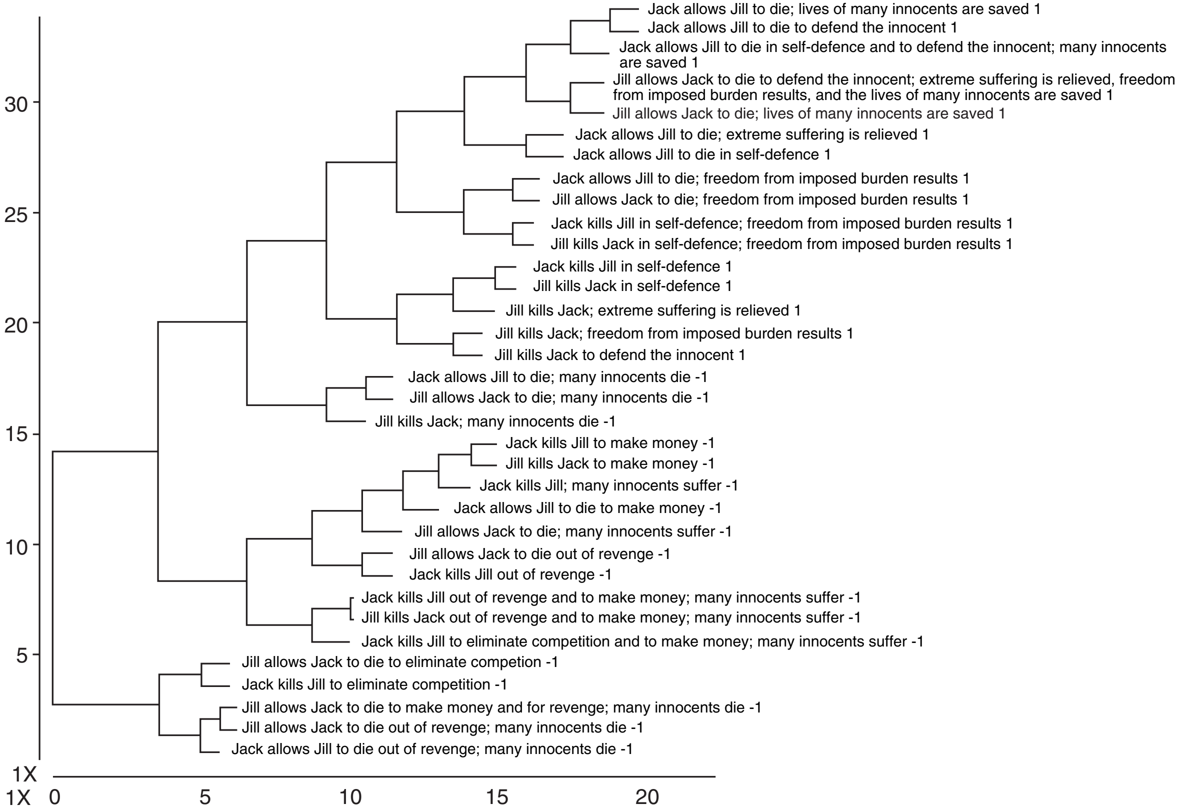

Figure 18.3. SRNc post training cluster plot of hidden unit activation vectors for 34 training cases.

if we take two cases that have the same motive(s) or consequence(s) but differ with respect to one involving killing and one involving allowing a death, such cases are more likely to be grouped together in SRNb than in SRNc. This is not surprising because in SRNb the subcases for both killing and allowing to die were treated in the same way, but in SRNc they are treated differently. Killing and allowing death are making different contributions to the similarity spaces for SRNb and SRNc because the subcases were classified differently. These plots are one more piece of evidence that a network's behavior can be in agreement with contributory standards even if the network was not supplied with such standards as input.

Are the networks we have been considering particularist or generalist? I want to suggest that this sort of dichotomy is not useful. To be sure, the networks have not been provided with substantive moral rules as input, and yet they train (with varying levels of success). Score one for the particularist. However, there is evidence that, the preceding notwithstanding, what the network learns has a general character to it. Score one for the generalist. When considering replies to the learning question (LQ), we considered LAa and LAb, where the former suggested that any rule learned did not feed back into the learning process, and the latter suggests that learned rules do feed back into the learning process. All of this makes us think of a learned rule in terms of an explicit, functionally discrete representational structure (something like a sentence) that either does or does not become casually efficacious in learning. The results in this section suggest that these are not the only logically possible options. Something general may be learned even if that generality is not given a classical, discrete representational structure.[8] Because it does not have such a structure, it is not even clear what it would mean to say that the generality feeds back into the learning process. That way of putting things assumes a framework for modeling thought that is committed to a very high degree to functionally discrete representations. What the network is doing when it is being trained is comparing the actual output on particular cases with the expected output and then modifying synaptic weights so that actual output on particular cases comes closer to the expected output. *There is never any explicit examination of a moral rule or principle*. In spite of this, the evidence suggests that generalities are mastered, and the simple generalities learned have an influence on the network. For example, consider the staged training of SRNc. Learning on the whole training set straight off led to failure. Training it on a subset, where it learned some generalities, and then training it so that it could master the rest of the cases was successful. So focusing on some cases and learning some simple generalities (implicitly) can make it easier to learn how to classify other cases. It is not at all obvious that anything like a functionally discrete representation is feeding back into the learning process, but there is reason to think learning some (implicit) generalities may be an important part of the

[8] No doubt someone like Jerry Fodor would balk at the potential of using such approaches in developing cognitive models. See Guarini (2009) for a response to some of the concerns raised by Fodor (2000).

https://doi.org/10.1017/CBO9780511978036.023 Published online by Cambridge University Press

learning process. Generalities may be at work even if they are not at work in some functionality discrete form. That generalities are at work is a point that a generalist would very much like and stress; that the generality is nowhere represented in a functionally discrete (language-like) form is a point the particularist would surely stress. A particularist might also stress that if we ignore the ordering of cases – start with the simpler ones, and then move to more complex cases – we simply will not be able to get certain networks to train (or to train efficiently), so thinking about the cases matters quite a bit. Mastering generalities helped SRNc to learn, but (a) the order of the presentation of the cases was crucial to the success of the learning, and (b) staged ordering contained no representation of generalities. In a situation like this, to ask which is more important to learning, generalities or cases, seems not very helpful at all. Saying that the cases are more important because their order of presentation matters is to miss that the network needed to master simple *generalities* before moving on to more complex cases; saying that generalities matter more misses that the network never explicitly operates on generalities and that the ordering of *cases* is central to mastering the relevant generalities.

The sort of rapid classification task (in a trained network) we have been considering is not something that would be classically conceived of as a reflective process. This is a point to which we will return. I mention it here to stress that the sort of learning we have been considering is not the only kind of learning. The consideration of forms of learning more thoroughly mediated by language and by means of inferential processes surely raises more issues (that cannot be adequately explored herein).

Establishing or Defending Normative Statuses

In this section we will consider how it is possible that sometimes cases can lead to the revision of principles, and sometimes principles can lead to the revision of cases. We will start with an admittedly contrived dialogue designed to show how difficult it is to establish in a non-question-begging manner that one of either cases or rules must be prior from the normative point of view. I do not know how to falsify the view outright that one of either cases or rules must be prior, so I will settle for showing the shakiness of the dialectical ground on which such views stand. Then, I will turn to examining how it could be that sometimes cases force revisions to rules, and sometimes rules to cases.

Consider the following dialogue:

Generalist: Never mind all that stuff about how we learn cases, now we are talking about how we defend what is actually right or wrong. For that, a case must always be deduced from a correct total standard.

Particularist: Why *must* it be like that? We often reject principles on the basis that they provide the wrong answer about specific cases. This suggests that with respect to establishing the normative status of a situation, cases are more basic than principles.

https://doi.org/10.1017/CBO9780511978036.023 Published online by Cambridge University Press

Generalist: But when we reject a principle P based on some case C, we are assuming some other principle, P2, is correct. It is based on that P2 that C must have the normative status it has.

Particularist: Why say we are assuming such a thing? That just begs the question against my position.

Generalist: Okay then, because you argued that cases can be used to overturn principles, how about the possibility that principles can be used to overturn cases. That happens sometimes. Doesn't that show that at least sometimes principles are more basic than cases?

Particularist: It does not. In the cases you mention, the principle cited is simply a kind of summary, a reminder of a set of related cases. In the end, it is really the cases that are doing the normative work, not the principles. Any principle that is cited is really presupposing normative views on cases.

Generalist: Hang on, when you started out, you said that cases can be used to overturn principles, and you objected to my claim that when this happens we are assuming the correctness of some principle that normatively underwrites the case being used. Now you are basically making the same move in your favor: You are saying that when a principle is used to overturn a view on a case, we are assuming the normative appropriateness of other cases underwriting that principle. How is it that you are permitted to make this sort of move and I am not?

Resolving the standoff in the preceding dialogue is difficult because any attempt to simply insist that a principle is normatively underwritten by cases may be countered by the insistence that cases are normatively underwritten by principles. The way this exchange is set up, the generalist assumes that cases presuppose some principle in order to have a specified normative status, and the particularist assumes that principles presuppose cases that have some specified normative status. They are both assuming that one of either cases or principles are more basic than the other when it comes to establishing normative status. Let us have a look at how it might be possible that neither cases nor principles are more basic than the other under all circumstances.

There is a difference between our pre-reflective,[9] non-inferential (or spontaneous or intuitive[10]) classificatory prowess and inferential, reflective reasoning. Thus far, we have been considering pre-reflective classificatory abilities. Reflective reasoning can involve explicit comparisons of cases with one another, explicit examination of principles and cases, and consciously drawing inferences about cases, principles, or the relationship between the two. To the extent that

[9] The prefix "pre" (as is the prefix "non") is potentially misleading when attached to "reflective." What is a non-inferential, pre-reflective process at time t_0 may be scrutinized by reflective processes at time t_1, leading to different non-inferential, pre-reflective assessments at time t_2. By referring to an assessment or any process as "pre-reflective," there is no attempt to suggest that the process has in no way been informed or influenced by reflective processes.

[10] I do not mean "intuitive" in a technical, philosophical sense (i.e., what someone like Kant was referring to when he discussed intuition). Rather, it is being used in something closer to the colloquial sense of an immediate (non-inferential) response.

https://doi.org/10.1017/CBO9780511978036.023 Published online by Cambridge University Press

contributory standards are at work in the networks considered earlier, they are at work implicitly or pre-reflectively. When engaged in reflective work, we often try to articulate what we take to be the similarities between cases, and proposing and defending contributory standards may play an important role in that process. Further examination of our pre-reflective views may lead us to revise the reflectively articulated standards, and the standards we reflectively articulate may lead us to revise our pre-reflective views on cases and may even lead to significant reconfigurations of our pre-reflective grasp of moral state space. Crucial to what is being referred to as a pre-reflective or an intuitive grasp of moral state space is that it is not (explicitly or consciously) inferential.

Let us return to the issue of whether we must say that one of either rules or cases is normatively prior to the other. We should keep in mind that arguments turning on claims like "without rules, we could not learn to generalize to new cases" are part of the psychology of moral reasoning. It is an empirical question concerning how it is possible for us to learn or not learn. If we subscribe to some form of *ought implies can*, then empirical constraints become relevant to establishing how we ought to reason. That said, it is not entirely obvious exactly how the empirical work on how it is possible for us to reason will turn out. Moreover, even if it turns out that explicit rules are absolutely essential to *learning*, it does not follow without further argument that rules are *normatively* prior to cases.[11] One piece of evidence that neither rules nor cases are exclusively prior is that each is sometimes used to revise the other. A few words are in order with respect to showing how this is possible.

On the model being considered, the initial classification of cases is done in a rapid, non-inferential manner. The SRN classifiers are toy models of that sort of process. Other (reflective) processes can then examine the work done by the pre-reflective processes. That citing general considerations can lead to the revision of cases is not surprising if we recognize that (a) there are generalities informing how we classify cases, and (b) the size of the case set whose members we are expected to rapidly classify is vast. Given the number of cases involved, it should be no shock if some simply do not line up with the generalities that tend to be at work, and pointing out that some cases do not line up with the general tendencies at work in related cases is an effective way of shifting the burden of proof. Moreover, general theoretical considerations may be cited in favor of contributory standards. For example, a case might grievously violate someone's autonomy, and someone might cite very general considerations against the violation of autonomy (such as autonomy being a condition for the possibility of morality). This sort of general consideration may lead to the revision of a particular case.[12]

[11] Someone may well want to argue that rules may be required for learning to proceed in an efficient manner, but cases are the source of the normative status of any rules we may learn. Put another way, someone might claim that rules merely play a pedagogical role, not a justificatory role.

[12] Although I will not explore disagreements between those who argue that morality is objective and those who argue that it is subjective (and those who think it is a little of both), I want to make it clear that I am trying to be as neutral as possible on these disputes for the purposes of this paper.

https://doi.org/10.1017/CBO9780511978036.023 Published online by Cambridge University Press

The geometric model can accommodate the views we have been discussing quite straightforwardly. If we are learning how to partition a state space to classify situations, given sufficiently many cases, partitions that capture generalities of some sort while leaving some cases out of line with the generalities would not be surprising. If generalities are constitutive of the location of cases in state space, then arguments that appeal to generalities could be expected to be effective at least in some contexts.

However, that the appeal to generalities will not always lead to straightforward answers on particular cases is also unsurprising if we recognize that there may well be a variety of contributory considerations, and these considerations can combine in many different ways. The importance of some general considerations have to be weighed against the importance of other general considerations, and it is often difficult to do this in the abstract; it is not until we descend to the level of particular cases that we understand the implications of weighing some contributory considerations more heavily than others. Again, the model we have been considering renders this unsurprising. Given that our reflective processes are often not very good at working out the consequences of various rules or their combinations, we should not be shocked if we reflectively generate a set of rules R such that someone can conceive of a case where (a) the reflective rules R are satisfied yet (b) our intuitive or pre-reflective processes yield a result different from the reflectively considered rules. This may well lead us to revise R to accommodate the case in question.

It could well be that sometimes we are inferring principles from an examination of cases, and sometimes we are inferring cases from an examination of principles. The model of pre-reflective classification working with reflective processes may provide a way of understanding how both of those practices can coexist. When we learn to navigate social space we master a moral state space that we effortlessly apply pre-reflectively; at any given time, parts of this space can be examined reflectively. However, it is in no way clear that the entire space can be reflectively examined at once. Perhaps moral cognition and its topics (like cognition more generally) are like an iceberg, where only a small part of it is reflectively active on a small subset of its topics at any given time.[13] The

I suspect that there are ways of formulating both objective and subjective views on ethics that are compatible with the view that neither cases nor rules are normatively prior to the other. Someone may argue that there are general theoretical considerations binding on all rational beings that some contributory standard CS1 is correct, and go on to argue that CS1 competes against other objectively correct CSi in the overall assessment of a case, and that we have to refer to cases in working out how to resolve conflicts between the CSi, so both standards and cases are essential. Others may argue that there is no way to argue for the objectivity of standards or cases, claim that whatever normative status they have is a function of how on individual was raised, and then use considerations mentioned in the body of this paper to argue that neither cases nor rules are prior to the other. The sketches offered here are entirely too brief, but they should give a sense of the different ways in which the views mentioned in this paper could be developed.

[13] The idea for the iceberg metaphor comes from Henderson and Horgan (2000), though they are concerned primarily with the epistemology of empirical knowledge in that paper.

https://doi.org/10.1017/CBO9780511978036.023 Published online by Cambridge University Press

rest is beneath the surface, so to speak. If we see ethical reasoning as involving an ongoing interaction between pre-reflective and reflective processes, then it is no surprise that we will have a wide variety of immediate intuitions on the moral status of cases as well as intuitions on level of similarity and difference between cases; nor is it surprising that we use talk of principles and cases to reflectively articulate our views. Computational neural modeling need not be seen as an attempt to supplant traditional linguistic and logical tools; indeed, it may well enrich them. By thinking of the location of cases in a state space, we may be able to develop more precise models of reasoning by similarity. If, as I have argued elsewhere (2010), analogical reasoning involves multidimensional similarity assessments, then understanding the similarity relations that hold between cases may help us better understand analogical reasoning. Algebraic and statistical tools for analyzing high-dimensional state spaces may augment the tools we have for reflective analysis of cases. To speak of contributory standards interacting in complex ways is kind of vague. To speak of a high-dimensional state space with cases clustering in that space opens up a wide variety of rigorous mathematical possibilities. Perhaps Euclidean distance will be a useful measure of the similarity of cases in that space; perhaps taxicab distance will have applications, or perhaps Mahalanobis distance will be an even better measure of similarity, and there are other possibilities still. We may be able to reconceive contributory standards in terms of the contribution they make to structuring a state space or in terms of their impact on one or more distance metrics for cases in a state space. Various forms of cluster plotting or principle components analysis or other analytical tools may be brought to bear on understanding the relationship between cases in state space. It may seem odd to some that we try to capture important patterns in thought using such tools, but it need not be seen as more odd than the introduction of the quantificational predicate calculus or deontic operators or any other set of formal tools currently on offer.

Conclusion

Cummins and Pollock (1991) begin their discussion of how philosophers drift into Artificial Intelligence (AI) by quipping that “Some just like the toys” but stress that “there are good intellectual reasons as well.” The demand for computational realizability requires a level of detail that (a) checks against a lack of rigor, (b) makes testing of one’s theories possible, and (c) requires that one take up the design stance. The first of these checks against philosophers hiding behind vague profundities. The second and third may lead to the discovery of errors and new insights in a way that armchair reflection may not (which is *not* to say that there is no place for armchair reflection). There are at least two different reasons why taking the design stance can lead to new insights on intelligence or rationality. The first is that once a computational system is built or set up to perform some task, it may fail in ways that reveal inadequacies in the theory guiding the construction

https://doi.org/10.1017/CBO9780511978036.023 Published online by Cambridge University Press

of the system. If this were the only benefit of taking up the design stance, then there would be no need to list (b) and (c) as separate points. However, there is another benefit of taking up the design stance. In the process of designing a system in sufficient detail that it could be computationally implemented, one may simply come to consider things that one has not considered before. Although the collection of papers in Cummins and Pollock (1991) does not examine the nature of ethical reasoning, the case can be made that their reasons for philosophers constructing and examining computational models applies to the sort of second-order positions we have been considering in this paper.

In training simple recurrent networks on classifying cases, it became possible to see how a system (a) could be trained on cases without the provision of explicit rules and (b) be subject to testing and analysis that shows its behavior to be in accordance with contributory standards. Moreover, inefficiencies or failures when subcases were not classified led to using the strategies of classifying subcases and staged training. Reflection on staged training led us to see how learning simple generalities could aid in mastering a more complex training set, even if the simple generalities mastered by the network are neither fed in as input nor explicitly represented elsewhere in the network. This is an example of how errors or difficulties in working with a computational system lead to new ways to approach a problem. Taking the design stance also requires us to recognize that we need real-time processes for rapid classification of situations, but we also need to capture the reflective modes of reasoning. Assuming that *ought* implies *can*, studying the constraints under which pre-reflective and reflective processes act and interact might lead to new insights about the constraints that are operative on how we ought to reason about ethical matters – yet another benefit of taking up the design stance. Finally, the admittedly brief discussion of state spaces and similarity in this paper is not cause for despair. There are a variety of mathematical techniques on offer that hold the hope of profoundly improving the rigor with which we explore the nature of similarity.

Acknowledgments

I thank the Shared Hierarchical Academic Computing Network (SHARCNet) for a digital humanities fellowship in support of this work. I also thank Joshua Chauvin for his assistance with the figures and for running neural network simulations.

References

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

Elman, J. 1990. “Finding Structure in Time.” *Cognitive Science* 14, 179–211.

Garfield, J. 2000. “Particularity and Principle: The Structure of Moral Knowledge,” in *Moral Particularism*, B. Hooker and M. Little, eds. Oxford: Oxford University Press.

https://doi.org/10.1017/CBO9780511978036.023 Published online by Cambridge University Press

Guarini, M. 2009. “Computational Theories of Mind, and Fodor’s Analysis of Neural Network Behaviour.” *Journal of Experimental and Theoretical Artificial Intelligence* 21, no.2, 137–153.

Guarini, M. 2010. “Particularism, Analogy, and Moral Cognition.” *Minds and Machines* 20, no. 3, 385–422.

Henderson, D. and Horgan, T. 2000. “Iceberg Epistemology.” *Philosophy and Phenomenological Research* 61, no. 3, 497–535.

Horgan, T. and Timmons, M. 2007. “Morphological Rationalism and the Psychology of Moral Judgement.” *Ethical Theory and Moral Practice* 10, 279–295.

Horgan, T. and Timmons, M. 2009. “What Does the Frame Problem Tell Us about Normativity?” *Ethical Theory and Moral Practice*, 12, 25–51.

Jackson, F., Petit, P. and Smith, M. 2000. “Ethical Particularism and Patterns,” in *Moral Particularism*, B. Hooker and M. Little, eds. Oxford: Oxford University Press.

Little, M. O. 2000. “Moral Generalities Revisited” in *Moral Particularism*, B. Hooker and M. Little, eds. Oxford: Oxford University Press.

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

McNaughton, D. and Rawling, P. 2000. “Unprincipled Ethics” in *Moral Particularism*, B. Hooker and M. Little, eds. Oxford: Oxford University Press.

https://doi.org/10.1017/CBO9780511978036.023 Published online by Cambridge University Press