2013

Case Classification, Similarities, Spaces of Reasons, and Coherences

Marcello Guarini
University of Windsor

CASE CLASSIFICATION, SIMILARITIES, SPACES OF REASONS, AND COHERENCES

Marcello Guarini

mguarini@uwindsor.ca

Abstract

A simple recurrent artificial neural network (ANN) is used to classify situations as permissible or impermissible. The trained ANN can be understood as having set up a similarity space of cases at the level of its internal or hidden units. An analysis of the network's internal representations is undertaken using a new visualization technique for state space approaches to understanding similarity. Insights from the literature on moral philosophy pertaining to contributory standards will be used to interpret the state space set up by the ANN as being structured by implicit reasons. The ANN, on its own, is not capable of explicitly representing or offering reasons to itself or others. That said, the low level similarity space set up by the network could be made available to higher order processes that exploit it for case-based reasoning. It is argued that for normative purposes, similarity could be seen as a contributor to procedural coherence in case-based reasoning and local forms of substantive coherence, but not to global forms of coherence given the computational complexity of managing those more ambitious forms of coherence.

Keywords: analogy; artificial neural network; contributory standard; generalism; global coherence; local coherence; non-local coherence; particularism; procedural coherence; reasoning; reasons; similarity; simple recurrent network; state space analysis; substantive coherence; visualization.

1 INTRODUCTION

How do we understand the kind of similarity at work in analogical reasoning in ethical or legal discourses? State space models of similarity have been around for some time, and a defense (Guarini, forthcoming) against some criticisms of state space approaches (Laakso and Cottrell, 2006; Markman and Gentner, 2005) has been taken up elsewhere. Here, I will present an artificial neural network (ANN) model that classifies cases, and then undertake an analysis of the similarity state space constructed by the trained network. I will argue that the state space is usefully understood in terms of contributory standards, as engaged in the moral philosophy literature on particularism and generalism. A new tool for visualizing high dimensional spaces will be introduced in the process of analyzing the ANN's behavior. The final portions of the paper contain a discussion of similarity and coherence. A role for local forms of procedural and substantive coherence will be defended.

2 BACKGROUND

2.1 TYPES OF SUBSTANTIVE PRINCIPLES

For the purposes of this paper, the expressions “principle,” “rule,” and “standard” will be used interchangeably. In the literature on moral philosophy, there are a number of different conceptions of moral standards (Dancy, 2000 & 2004; McKeever and Ridge, 2006). What has come to be referred to as the contributory or pro tanto standard in that literature has much in common with what Kevin Ashley (1990) has referred to as factors. A contributory standard is *not* the sort of normative standard that, when combined with a statement of fact, licenses a monotonic deduction about some particular case. To a first approximation, a contributory standard asserts that some feature (monadic or relational) makes a contribution towards permissibility (or impermissibility), but the contribution it makes may be outweighed by other features contributing in a different way.

Particularists such as Jonathan Dancy (2000 & 2004) reject all kinds of general standards, including the contributory. When challenged (Jackson et. al., 2000) as to whether a particularist understanding of reasons would allow us to understand how we learn the difference between right and wrong, permissible or impermissible, or the like, Dancy (1999) gestured in the direction of Artificial Neural Networks (ANNs). The hypothesis was that such systems might be able to (a) generalize to new cases based on cases already learned, and (b) do the preceding without making use of general rules, principles, or standards of any kind.

2.2 TRAINING THE MORAL CASE CLASSIFIER

Some work has already been done with respect to testing and critically assessing this hypothesis (Guarini 2006, 2010, 2011, forthcoming). Building on this work we will examine a simple recurrent network designed to classify moral situations into two categories: permissible (output = 1) and impermissible (ouput = -1). The output layer of this ANN has one unit; the input layer has eight units, and the hidden layer has 24 units. There is a context layer with 24 units connected one-to-one with the hidden units. Vectors representing phrases are presented to the network sequentially. Every case presented to the network consists of one of two individuals, Jack or Jill, either killing or allowing someone to die. See Table 1 for a list of sample training or testing cases. All the cases have the following form, where the parentheses indicate an optional component:

Agent, Action, Agent, (Motive(s)), (Consequence(s))

Sometimes the cases have multiple motives, sometimes multiple consequences, sometimes just motives and no consequences, and sometimes just consequences and no motives. Sometimes there is just a single motive or single consequence. All cases in the training set included at least one motive or one consequence. Some of the testing cases had no motives or consequences at all. Since the initial training and testing was done on moral situations, I have referred to the ANN in question as the Moral Case Classifier (or MCC). However, there is nothing that prevents this sort of architecture from classifying legal cases since all that is required is a vectorized description as input and a classification goal for output.

Input	Output
Jill kills Jack to make money	-1
Jack allows to die Jill out of revenge.	-1
Jill kills Jack in self defense, to defend the innocent; the lives of many innocents are saved.	1

Table 1: Sample Cases.

3 LOOKING INSIDE THE "BLACK BOX"

In the early days of computational neural modeling, there was some concern that neural networks were black boxes: they might be able to do pattern classification, but it was not clear how they did what they did. Many techniques have been developed for understanding the internal workings of an ANN. In this section, we will consider (a) a sample classification task and (b) a visualization technique that will allow us to see what is going on in the network. Let us consider an example.

Imagine you are kidnapped, knocked unconscious, and when you wake up, you find yourself connected to another individual. You are informed by the hospital staff that the society of music lovers did this to you to keep their beloved violinist (to whom you are connected) alive. You are free to disconnect yourself and walk away, but this will result in the certain death of the violinist. In discussing the ethics of abortion, Judith Thomson (1971) used this example for a number of reasons. At one point she suggested that the violinist case is similar to the case of pregnancy resulting from rape. The idea appears to be that in both cases, one life has been made dependent on another through force. (Thomson grants that the fetus becomes a person not long after conception, and she argues in the aforementioned paper for the moral permissibility of some abortions even if the fetus is a person.) Some have claimed that in the case of the violinist, unplugging yourself and walking away amounts to allowing the violinist to die, and in cases of abortion, killing is taking place. Thomson claims that there is sufficient similarity between the case of the violinist and the case of rape induced pregnancy that, if it is morally permissible to "walk away" from the violinist (or allow the violinist to die), then it is permissible to have an abortion (or kill the fetus). The moral case classifier (MCC) was trained and tested on cases that are designed to mimic how some see the violinist and rape induced pregnancy cases. Before seeing how the MCC handles these cases, let us consider a new way of visualizing a network's hidden unit activation vector state space.

We can understand what the MCC is doing during training as building up an internal or hidden unit level representation of every case that is being presented to it. The context units are being used as a kind of working memory that allows a representation for the entire case to be built up at the level of hidden units. If we plot the value of every hidden unit on an axis, we get a 24 dimensional moral state space for the network. It is a straight forward matter to plot three dimensions on a two dimensional surface, but three dimensions does not allow us to see very much of what is going on in a 24 dimensional space. Consider the following strategy: instead of representing each 24 dimensional vector for each case with a point, let us represent each case with a cone in 3 dimensional space. The center of the base of the cone in this space gives us 3 dimensions of information. The width of the base gives us a fourth dimension; the height of the cone gives us a fifth dimension; the location where the vertex of the cone is pointing gives us another 3 dimensions; the color of the shell of the cone if coded using RGB color coding gives us another three dimensions, and the color of the base of the cone (again with RGB coding) gives us another three dimensions. In this way we can represent 14 dimensions of information. Using cones in 3 dimensional space, we can project the first 14 principal components of the vectors (or moral cases) from the original 24 dimensional space. This improves our ability to visualize what is going on in this space, and it will come in handy, shortly. Each of the cones in the three figures in this article presents 14 dimensions of information (most of which will go unexplored for the purpose of this short piece).

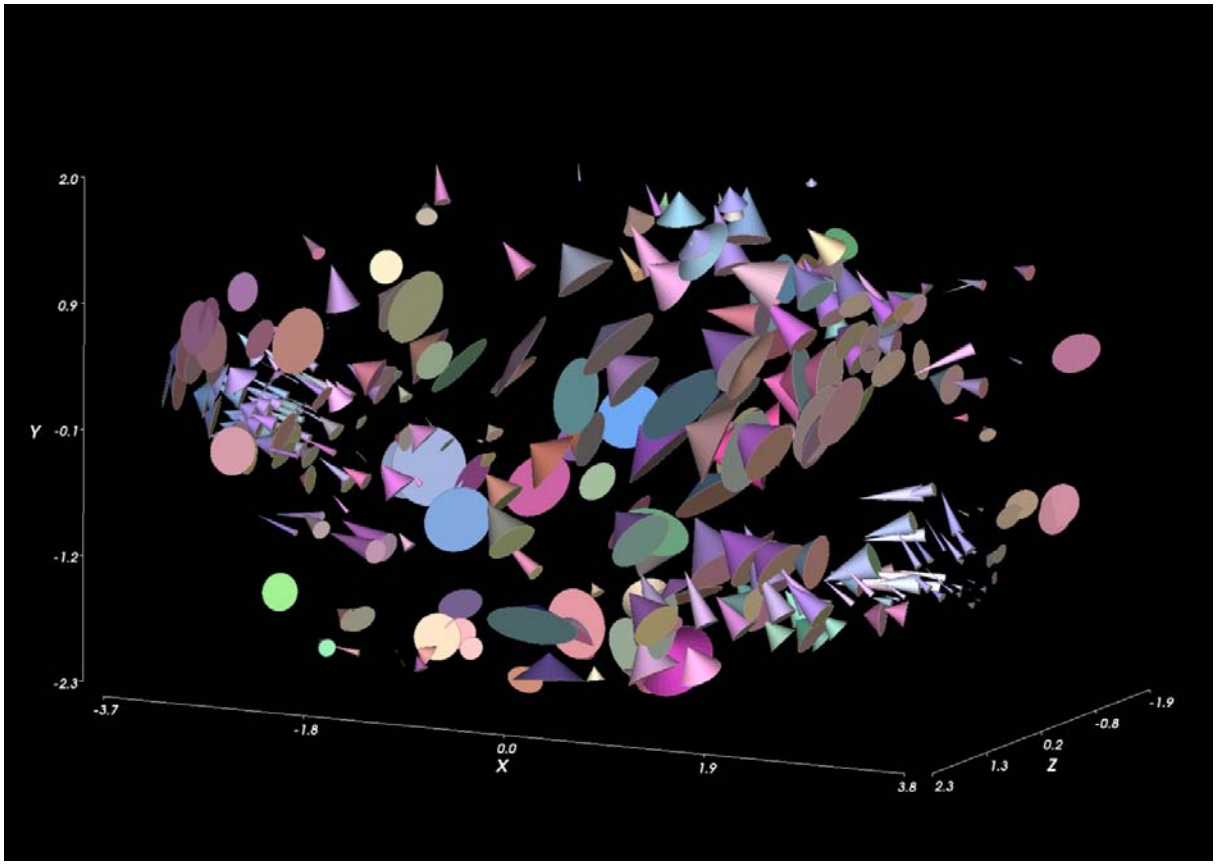

Figure 1. Each cone represents one of 326 cases. (Labels for the cases are omitted.) The x-axis represents the first principal component of the distribution of cases from the MCC's 24 dimension hidden unit state space. All cases to the right of zero on the x-axis are impermissible; all cases to the left of zero are permissible.

The MCC was trained so that cases of the form

x allows to die y, freedom from imposed burden results

were classified as morally acceptable – think of these as violinist type cases. Cases of the form

x kills y, freedom from imposed burden results

were classified as impermissible – think of abortion in cases of rape induced pregnancy. Some who hold positions of this sort have been persuaded by the similarity between the violinist and rape induced pregnancy to change their views. In other words, in spite of the fact that they initially, say at time t_0, classified the cases in different ways, they saw a similarity (of some sort) between the cases when questioned at t_1, and that lead to a change in classification at t_2. How do we understand the nature of this similarity? How can it turn out that cases classified in different ways at t_0 can be seen as, in some sense, similar? Is there an incoherence involved in the preceding? The remainder of the paper explores these questions.

Each of the 326 cones in figure 1 represents one of the training or testing cases for the MCC. The first principal component is plotted on the x-axis. It turns out that cases to the right of zero on the x-axis are impermissible, and those to the left of zero are permissible. Say we take impermissible cases and plot them on their own (figure 2), and permissible cases and plot them on their own (figure 3). Actually, figure 2 contains one permissible case; more on that soon enough. In figures 2 and 3, the x-axis is the Mahalonobis distance from the mean of the cluster being plotted. Mahalonobis distance is a non-Euclidean, statistical distance measure that can be used to check the distance/similarity of a case from/to a cluster of cases. The remaining dimensions of information plot the first 13 principal components. The highlighted cone in figures 2 and 3 represents the following case.

C: Jill kills Jack to obtain freedom from imposed burden; freedom from being bedridden for 9 months results; freedom from invasion of privacy results.

This case could be thought of as an instance of the abortion of a very difficult pregnancy resulting from rape. (“Jack” is the fetus.) The network was trained to classify this case as impermissible, and it did so, and it showed up in the impermissibility subspace of figure 1. C is the only impermissible case included in figure 2. When we check the Mahalonobis distance of C (see the cone with red cube highlighting in figure 2) from the mean of the impermissibility subset and compare it to the Mahalonobis distance of C from the mean of the permissibility subset (see the cone with the red cube highlighting in figure 3), it turns out that C is closer to or more like the permissibility cases than the impermissibility cases (in spite of the fact that it was classified as impermissible). This happened because the network was trained on many cases involving “freedom from invasion of privacy results” and “freedom

from being bedridden for 9 months results" and "to obtain freedom from imposed burden." Moreover, many of those cases were classified as permissible. These other features appear to be contributing to permissibility, but for purposes of the *final output classification*, in case C, killing appeared to outweigh these other considerations. However, those other considerations appear to still "carry weight" in the sense that they have an effect on the location of the case in similarity space. The location of the case in similarity space depends on the first set of synaptic weights between the input layer and the hidden layer. There is still a second set of synaptic weights, between the hidden layer and the output layer, that contributes to the final classification. Even if the features in question have a particular weighting in the similarity space, the output or final classification makes use of another set of synaptic weights that further modifies the contributions of specific features. See Guarini (forthcoming) for further details.

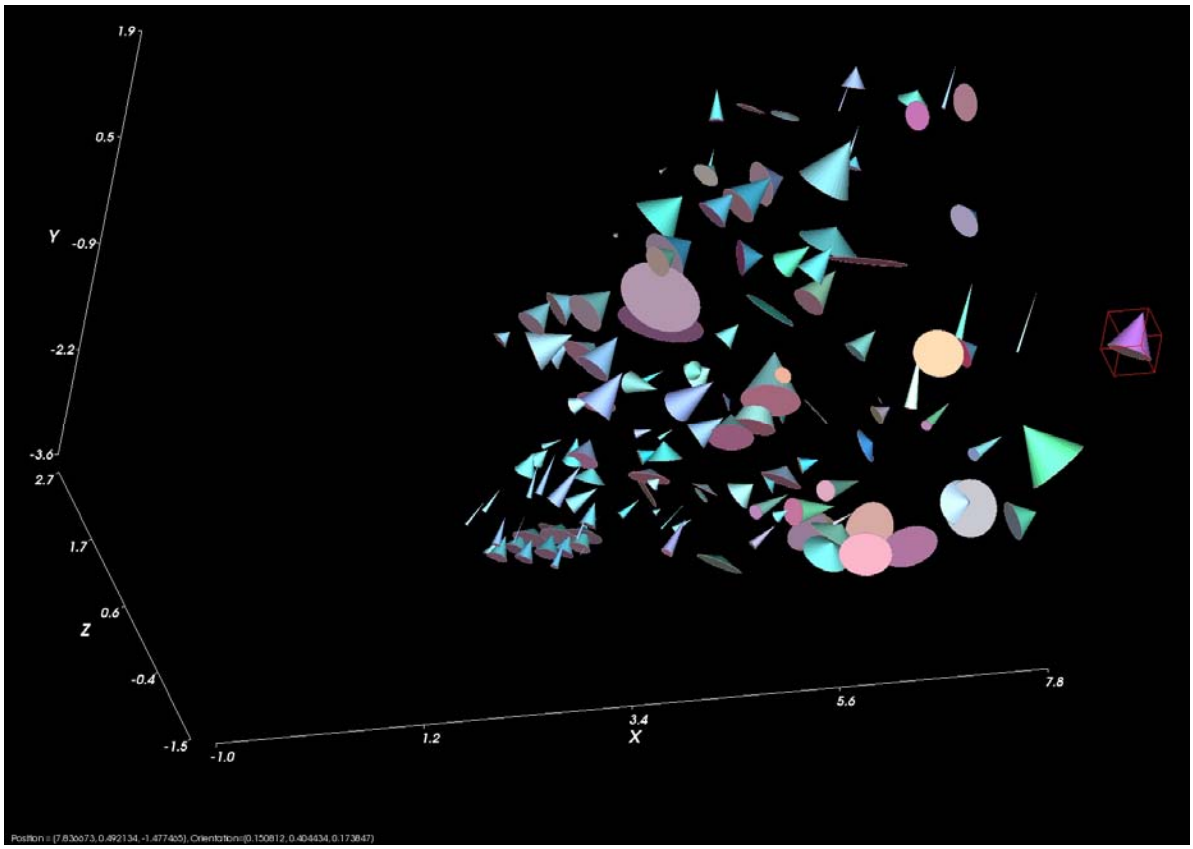

Figure 2. The cases represented here are all classified as impermissible. The x-axis represents the Mahalonobis distance from the mean of impermissible cases. The case highlighted with a red cube (far right of x-axis) is case C discussed in the text. Compare its location here with its location in figure 3.

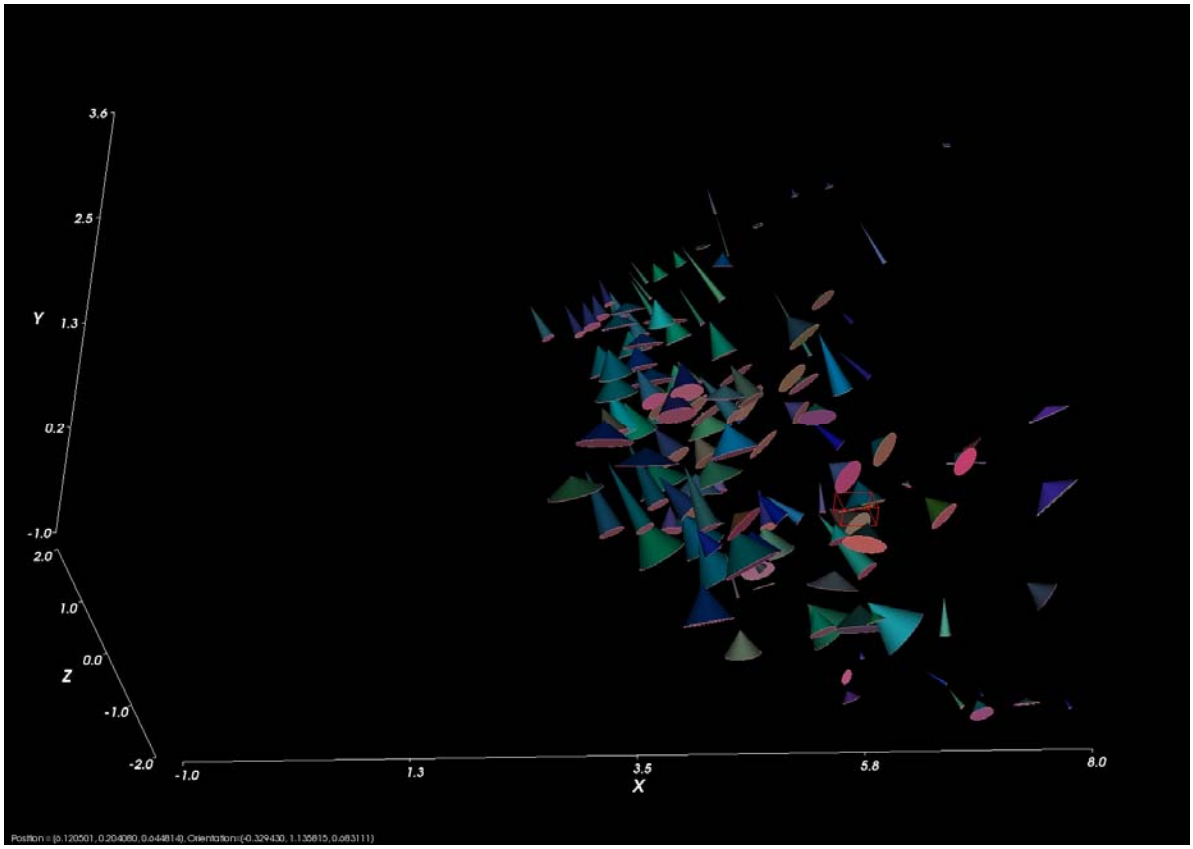

Figure 3. The cases represented here are all classified as permissible, except for case C, highlighted with a red cube (near 6 on the x-axis). The x-axis represents the Mahalonobis distance from the mean of permissible cases. Case C is closer to permissible than to impermissible cases.

4 SPACES OF REASONS

The MCC only performs a low-level classification task. It does not do any higher-level reasoning or reflection. However, it is conceivable that hidden unit representations of the MCC could be fed into another process that does higher-level tasks. The properties of the lower-level representations would then be available to the higher-level processes. Contributions made by specific features or relations would be among these properties since the information is implicit in the representations. Something else that might become available to the higher-level process(es) is information about how one case is related to clusters. It may turn out that one case that is classified as impermissible, but is closer to the permissible cases (using Mahalonobis distance, or perhaps some other metric), may play an important role in understanding which sorts of agents are more open to changing their minds on certain kinds of cases, or at what point in the process of argumentation they may become open to changing their minds.

We can speak of two spaces of reasons. One space of reasons operates at the subreflective level. Let us call this a space of implicit reasons. The MCC is a crude, simplistic, toy-like approximation of how different contributory considerations can be at work for the purpose of unreflective, fast classification of situations. In the space of explicit or reflective reasons, we represent, articulate, and argue about considerations that make normative differences. Getting clear on the relationship between these two spaces is no easy matter. One of the interesting features of that relationship is that genuine surprise or discovery is possible: in

attempting to articulate or make explicit what we take to be contributory considerations that are implicitly at work in classification, we can be presented with examples that show us that our reflective understanding of implicit space is not as well developed as we thought it was. The fact that some who treat abortion as impermissible in cases of rape are genuinely surprised by the force of the violinist argument (or its variants) is just one of many examples of that phenomenon. The working hypothesis for the rest of this paper is that understanding the interplay of these different spaces of reasons can contribute to our understanding of coherence.

5 A Role for Procedural Coherence

Elsewhere (Guarini 2007) I argued against the usefulness of a specific type of coherence model of reasoning, the multiconstraint model of coherence understood as operating in a global manner. My concern in that paper was (a) that we could not consciously compute the required coherence values without aids, and (b) to rely on external aids or unconscious computation does not capture much of what we think is important about normative or reflective reasoning. The coherence being argued against there was a kind of global substantive coherence (applying to all beliefs, goals, actions and the like). It does not follow from any of the arguments in that paper that there is no role for coherence or incoherence to play. Schiavello (this volume) also expresses concerns about global or large scale forms of coherence. It is completely consistent to express certain kinds of skepticism about global coherence and take seriously local forms of coherence. Let us examine this possibility further.

If we say that two or more things cohere, we are suggesting that they, in some sense, "fit together" or "go together" or some such. If two or more things incohere, the reverse is true. Consistency is one contributor to coherence, but pretty much everyone agrees it is not the only one. If we take similarity to be another contributor to coherence, cases that are similar and treated in the same way cohere, but cases that are similar and treated in different ways without an explanation of the different treatments are said to incohere. Note well: there is no inconsistency in saying that we should classify the violinist case (V) in one way and the case of abortion in pregnancy resulting from rape (R) in another way. If you take V and R to be sufficiently similar, then you may feel or claim there to be an incoherence in doing so (but none of this commits you to anything of the form p and not p). Arguments from analogy often turn on making use of an incoherence present in one or more interlocutors. If individuals acknowledge that V and R are normatively similar (without any significant dissimilarities), then they should not classify them in different ways. Sometimes, the argument is not so much from the analogy as it is to the analogy. Individuals may simply not see any similarity at all between V and R, and the burden would then be to try to get them to see the similarity, which leads to their different initial classifications for V and R being incoherent, which may then force a revision. For reasons of space, I will focus on the former scenario, the argument from the analogy.

Imagine that someone is opposed to abortion in cases of pregnancy resulting from rape, and this individual is presented with the violinist argument. He sees the similarity between V and R, concedes the incoherence of admitting similarity but classifying differently, and revises his view. On the state space account developed herein, what happened is something like this. An initial classification of cases lead to R being classified as impermissible even though R

was more like V with respect to the balance of permissibility- and impermissibility-making considerations, so much so that R is more like the permissible cases (in terms of mahalanobis distance) than the impermissible cases. If something like this were going on in a person, we should in no way suggest that the individual in question would have conscious access to all the contributory considerations and how they structure the distribution of cases such that R is closer to one subset of cases rather than another. What the individual has is an intuition that R is similar to V; that intuition may be the result of processing at the subreflective or unconscious level. At the conscious or reflective level, there is a recognition that it would be incoherent to concede that (a) R and V are normatively similar and (b) that they are treated differently, without (c) citing some normative difference-making considerations. We do try to avoid that kind of incoherence, but the reflective and articulated avoidance of that kind of coherence is a local affair. We never consciously consider all the cases in our normative state space or even a significant subset in any given argument. Moreover, the incoherence in question pertains to asserting that two or more cases are normatively similar while treating them differently and not being able to distinguish the cases; the corresponding form of coherence requires that when we assert that cases are normatively similar, that we treat them in the same way, and if we do not, then we recognize an obligation to cite differences that warrant different treatments of the cases (normative similarities notwithstanding). This is a kind of procedural coherence. One person (or system) could take one view on a given case and claim that it is analogous with a given precedent, and a different person (or system) might admit that there are normatively weighty similarities, but argue that these are outweighed by even more weighty dissimilarities. Both would be procedurally coherent.

Our willingness to say, our intuition if you will, that a pair of cases is normatively similar (or dissimilar) may well be informed by much that is going on at the subreflective level, but at the reflective level we can only ever deal with a small number of cases. Even if (for the sake of argument) we assume that computational processes going on at the subreflective level are coherence promoting in some global sense, it does not follow that at the reflective level what makes analogies normatively appropriate is the maximization of global coherence. Reflectively, we consider things like cases, principles, and similarity statements (which sometimes cite principles); we do not appear to do anything like evaluate global coherence, nor is it clear that we could even if we wanted to. To see this, imagine that we have 100 cases. Consider three-wise similarity comparisons of this form, where X_i are cases:

X_1 is more similar to X_2 than it is to X_3.

Given 100 cases, there are 100^3 (or 1,000,000) possible three-wise similarity comparisons. Imagine the conjunction of all those similarity statements; assuming bivalence, checking that conjunction for consistency (which would be a contributor to global coherence) would require $2^{1,000,000}$ steps using an exhaustive truth table method. Even if we checked only 2^{100} similarity statements for consistency, it would be more than we could consciously reflect over in a lifetime. (To put this in perspective, 2^{100} is approximately 10^{30}. Assuming the universe is about 15 billion years old, the total number of seconds in the history of the universe is on the order of 10^{17}.) So even if some sort of computational process could help us to usefully

approximate global coherence at the subreflective level, it does not follow, without further argument, that this plays an important role at the reflective level. The same point can be made for a probably approximate correct – PAC – account of global coherence or other accounts that might be inspired by the machine learning literature. Even if they play a role at the subreflective level in approximating coherence, it does not follow, without further argument, that such global coherence approximations would play a role at the normative, reflective level.

6 A Role for Surveyable, Substantive Coherence

In the previous section, we have only considered the *coherence or incoherence of cases and their classifications*. Normative principles of different sorts, goals of different sorts, and beliefs about empirical matters could also factor into more broadly conceived types of coherence – there is no room to explore all of that here. (Indeed, other papers in this collection are considering types of coherence not considered in this paper, and the arguments herein need not and should not be interpreted as applying to all imaginable types of coherences.) It might be argued, though, that with respect to the coherence of cases, the remarks of the previous section are too pessimistic. Perhaps, it might be suggested, we could design a system that could compute coherence better than we could. I have said a few things about that line of thought elsewhere (Guarini 2007). For now, let us imagine that there is a disputed case X, and we provide it as input to a system that considers two possible verdicts on the case. When we take X, assign it a specific verdict, and add it to a system's case base B_0 (the set of cases for which it already has verdicts) we get set B_1. When we add X with the opposite verdict to B_0, we get set B_2. If the system returned the result that B_1 is more coherent than B_2 and nothing else, I doubt we would find this helpful in normative matters. If the system returned the result that a given case should be treated in a certain way because it bares important similarities to some other case (or several other cases), and the apparent differences are explained away, then this would be quite useful from a normative, reflective perspective. However, it is not clear what simply adding, "B_1 is more globally coherent than B_2," would add from a normative perspective (even if we included numerical descriptions of the coherence levels).

It does not follow from anything said thus far that a coherence engine that seeks after or approximates some sort of substantive coherence would have no role to play. Perhaps it could inform the search for relevant cases to use in analogical reasoning or argument. If that sort of usefulness obtains, it does not follow that what makes the analogy a good or bad one has to do with global coherence. Being able to evaluate the analogy would appear to have more to do with (i) showing that specific similarities are to be afforded normative weight (or not), and (ii) showing that specific similarities outweigh specific differences (or vice versa). Objection: but the considerations involved in doing the preceding could come from anywhere, so that means everything has to hang together, so global coherence is required in the end. Reply: perhaps considerations pertaining to (i) and (ii) could come from almost *anywhere* if we are considering the matter in general, but that does not mean that they come from *everywhere* when we are examining a specific problem Substantive, global, coherence about cases is computationally exacerbating because all cases (in some sense) are used in the computation. There may be no way to specify *in advance of any given dispute* which sorts of considerations (including which cases may be appealed to) may become relevant for carrying out (i) or (ii). It does not follow that we need to make reference to global coherence

in carrying out normative or reflective tasks focused on a specific problem. For example, consider two lawyers arguing about some target case T. Lawyer one appeals to precedent P_1, and lawyer two appeals to P_2, and each makes a strong argument because, for the sake of argument, T is similar to both P_1 and P_2, and the P_i cases were decided in different ways by different courts in the past. This stalemate may be broken if one of the lawyers comes up with an argument that can show that her way of deciding the case fits better with the past practice of the courts in a range of cases. That said, it must be possible to survey this range of cases in an argument in order for the appeal to fit with prior cases to have normative force. This appears to allow a substantive role for coherence because an appeal to better fit is playing a substantive normative role: it may contribute to one argument being better than another in virtue of the fit or coherence between the cases. Still, this is a local form of coherence. It may be true that, to a first approximation, all cases decided in the past and even hypothetical cases are "fair game" with respect to which cases we could *potentially* appeal to in such coherence arguments, but it does not follow that all cases are or need to be (or could be – think of all the logically possible hypothetical cases!) appealed to in such arguments. Once again, to put it roughly, just because considerations could come from anywhere, it does not follow that they come from everywhere. This is especially so at the level of conscious, reflective reasoning.

So far we have considered local forms of coherence – which are restricted to what can be consciously reflected over – and global forms of coherence. A global form of coherence *with respect to cases and verdicts* would have to consider all cases and their verdicts. A global form of coherence without any qualifiers would consider everything that could factor into coherence (principles, goals, beliefs about empirical matters…). To all this, we could add the idea of a *non-local* form of coherence. A non-local form of coherence with respect to cases and verdicts would encompass more cases than we could ever hope to individually, consciously reflect over *for purposes of considering coherence*, but would fall short of considering *all* cases and verdicts. The term "global" is pretty strong, even if we qualify it by saying it refers only to the role of cases and verdicts and their contributions to coherence. The consideration of all cases would include even hypothetical cases. Someone might want to restrict the cases involved in coherence calculations to those which and individual has encountered in his or her lifetime. This is clearly something less than global coherence, but the calculations involved in the coherence assessment would still be more than we could expect someone to do in a lifetime – hence the expression non-local coherence. As we saw above, computing coherence over 100 cases is not manageable at the reflective level, and as a general rule people will encounter more than 100 cases that they will be expected to classify in a lifetime. There may be subreflective processes that could examine more than we could examine consciously, but even if those processes are of a coherence promoting nature, two things need to be kept in mind. First, given the computational complexity involved, there are likely limits even here with respect to the number of cases that can be examined feasibly. Even if we abandon globality and opt for some sort of non-local coherence, a high level of complexity is still in play. Second, even if we subreflectively implement procedures that usefully approximate non-local coherence, it does not automatically follow that that non-local coherence considerations are normatively or reflectively helpful. That point would still have to be defended.

7 Some Methodological Reflections

A computational model was discussed above, and others in this volume discuss such models as well, so some methodological reflections on their use are in order. There are different kinds of criticisms that are aimed at simple or "toy" computational models. One type of critique wonders of what possible relevance descriptive considerations could be for prescriptive or normative theorizing. The response to this is to point out that if in some sense or other we can say that ought implies can, that before claiming someone is obligated to reason in a given way it should (as a matter of fact) be possible for them to reason in that way, then empirical and computational considerations are relevant to normative theorizing. Indeed, some of my own arguments about the computational complexity of certain kinds of coherence is an attempt to use empirical and computational considerations in an attempt to constrain normative theorizing about what kinds of reasoning may be appropriate. A second type of critique starts by pointing out that since existing models tend to be very simple, there is not much point to them either descriptively or prescriptively. Descriptively, they are known to be too simple (even by those who propose them), and if they are descriptively incomplete, then (it is claimed) they could not possibly inform prescriptive or normative theorizing. What is the point of using a descriptively incomplete model to constrain or otherwise inform theorizing about coherence (or reasoning more generally)?

That is a good question. I will start my answer by borrowing a metaphor from Wilfred Sellars (1963), who proposed that our manifest and scientific images of ourselves and our world needed to be fused, and he compared this to the way our two eyes bring together two different and overlapping images of the world. Instead of discussing our manifest and scientific images, I will speak of our prescriptive and descriptive images. If we can say that ought implies can, then there is an area of overlap between these images, since what we ought to do is constrained by what we can do. But that is not the only overlap: what we take to be the case regarding descriptive matters is informed by what we take to be appropriate or inappropriate forms of reasoning or gathering evidence. Consider the example of the multi-site, double blind, clinical trial. Over the years, empirical research turned up results about placebo effects, about bias in individual researchers, and about variations in research cultures in different institutions. These empirical findings were used to develop new methods of gathering evidence, new methods for arriving at conclusions about how we ought to reason about the prospects of using newly developed medicines or treatments. These new methods have lead to the overturning of some past empirical results and to more reliable ways of discovering effective treatments. It is almost surely the case that we have not uncovered all the potential ways in which bias and culture can skew experimental results, but it does not follow that we stop experimenting. Research continues on various psychological and sociological factors that can affect the results of research, and, no doubt, new methods will be developed to control for those effects. No one would suggest that we stop doing research until we are done acquiring all the empirical information about how the aforementioned factors can skew research.

In fusing our descriptive and prescriptive images, I want to suggest that neither is prior to the other. The metaphor of the images being side-by-side (one for each eye as it were) is useful

since neither is taken as more basic than the other. If descriptive considerations are relevant to and can inform prescriptive claims, then it is difficult to argue that the prescriptive is prior to or more basic than the descriptive. If, on the other hand, the prescriptive can inform and constrain how we acquire descriptive information, then the descriptive is not basic either. We start where we are, modifying each image using the other. If this is right, then there is no point in saying that we have to be "done" with descriptive work before we can do the prescriptive work; nor does it help to say that we have to be "done" with our prescriptive theory before we can do empirical work. Each image informs the evolution of the other in an on-going manner. If the idea of equally basic images is on the right track, then we can begin to see how even simple descriptive models can be helpful. A descriptive model does not have to be *done* or *complete* before it is useful. Our descriptive models of how research is done in various fields are not complete, but they have already usefully informed prescriptive claims about how research should be done. We now need to return to models of coherence and show how the fusing of two images is relevant.

One virtue of models, even simple ones known to be incomplete, is that they may make predictions, which can lead to new research, including the need to further refine the model. Thinking about coherence in terms of computational models may lead to new insights about what coherence may or may not be. The main idea behind coherence theories is that, in some sense, the way propositions (or other "things") "hang together" or "fit together" contributes to their justification, warrant, or some other sort of reasonable or normatively appropriate status. Left at that, we have an intolerably vague position. Adding that consistency and explanation are contributors to coherence might help, but it is still pretty vague. Exactly how are ideas supposed to fit together in a way that contributes to positive normative status? It is difficult to specify what it means for things to "fit together" in the relevant respects. Both here and elsewhere, I have expressed some skepticism about what can be accomplished with certain kinds of coherence models. It is a virtue, though, of computational models that they are clear enough to allow for specific criticisms. Assuming some type(s) of coherence play a role in human reasoning, constructing computational models of coherence(s) is a useful way to develop clarity and rigour with respect to what coherence may or may not be. Better descriptive models (which have to start somewhere) could lead to a better understanding of how we can reason, which could be used to place constraints on how we (normatively) expect people to reason. Of course, since coherence is supposed to increase positive normative status, our views on what is normatively appropriate in reasoning will be informing any attempt to build a computational model of coherence. So, our normative views about reasoning inform attempts to build computational models, and the descriptive work of building the model has the potential to feed back on our normative views by showing us that they are too vague, incompletely specified, or otherwise flawed to allow for rigourous modeling, which would force us to revise our normative views. This would lead to better attempts at model development, which will surely run into further problems, and the cycle continues.... The normative and descriptive images inform one another.

To all this it might be replied that many models of coherence are really, really simple, so how could they be useful at this stage in their development? Well, judging by the number of *different* things the authors in this collection are saying about coherence(s) (and what so

many authors have said elsewhere) it may well be that our understanding of coherence(s) is really, really inadequate. If that is right, then it may well be that even working with simple models could lead to important clarifications, more rigour, and new insights for testing and development. It is not just something descriptive that will be tested; it is also something normative. For if the computational model fails to work or is demonstrably incapable of scaling up when we have implemented our best insights on what amounts to normatively appropriate reasoning (coherence-based or otherwise), then maybe our best insights about such reasoning are not good enough.

8 CONCLUSIONS

The preceding has implications for computational modeling, whether we are interested in moving toward a model of how humans reason, or whether we are interested in constructing a system that can aid humans even though it may not work in the way humans work. For example, coherence as constraint satisfaction may have a role to play at the level of reflective or explicit reasoning, but the cases or other considerations being appealed to would have to be surveyable in the course of an argument. There might be a role for coherence to play at the level of subreflective considerations (i.e. considerations that would not be explicitly articulated and offered for normative consideration), but even here we need to be wary of computational complexity, though the constraints operative at this level need not be identical to the constraints operative at the reflective level. Also, we should not assume that machines designed to do reasoning will be subject to human constraints. The access a machine has to a subreflective state space of the kind considered in section four may be different from the kind of access a human has. That said, if some sort of computational system is to interact with humans and be able to provide persuasive reasons to humans about how to classify cases, the constraints on the human cognitive architecture will have to guide the sorts of reasons any such system would communicate to us if they are to be useful in helping us understand why a given case should be treated in some prescribed or suggested manner.

Acknowledgments

I thank the Shared Hierarchical Academic Research Computing Network (SHARCNet) for a digital humanities fellowship that made this research possible. The fellowship included funding for course releases and programming support. Special thanks go to SHARCNet programmer Weiguang Guan for coding the visualization software used to render figures 1 through 3. For comments and suggestions, thanks also go out to Michał Araszkiewicz and Jaromir Savelka, organizers of the ICAIL 2011 workshop on Coherence, as well as Kevin Ashley, Thorne McCarty, and other workshop participants for their comments and suggestions.

REFERENCES

Ashley, K. D. (1990), *Modeling Legal Argument: Reasoning with Cases and Hypotheticals*, MIT Press, Cambridge, MA.

Dancy, J. (1999), 'Can a Particularist Learn The Difference between Right and Wrong?' in K. Brinkmann (ed.), Proceedings from the 20th World Congress of Philosophy, Volume I: Ethics, Philosophy Documentation Center, Bowling Green, Ohio.

Dancy, J. (2000), 'The Particularist's Progress' in B. Hooker and M. Little (eds.), *Moral Particularism*, Clarendon-Oxford press, Oxford, pp. 130-156.

Dancy, J. (2004), *Ethics Without Principles*, Oxford University Press, Oxford.

Guarini, M. (2006), 'Particularism and the Classification and Reclassification of Moral Cases', *IEEE Intelligent Systems Magazine* 21, 22-28.

Guarini, M. (2007), 'Computation, Coherence, and Ethical Reasoning', *Minds and Machines* 17, 27-46.

Guarini, M. (2010), 'Particularism, Analogy, and Moral Cognition', *Minds and Machines* 20, 385-422.

Guarini, M. (2011), 'Computational Neural Modeling and the Philosophy of Ethics' in: *Machine Ethics*, eds. M. Anderson and S. Anderson, Cambridge University Press, Cambridge, UK, 316-334.

Guarini, M. (forthcoming), 'Moral Case Classification and the Nonlocality of Reasons', *Topoi.*

Jackson, F., Petit, P. and M. Smith (2000), 'Ethical Particularism and Patterns' in: *Moral Particularism*, eds. B. Hooker and M. Little, Clarendon-Oxford press, Oxford, pp. 79-99.

Laakso, A. and G. Cottrell (2006), 'Churchland on Connectionism' in B.L. Keeley (ed.), Paul Churchland, Cambridge University Press, Cambridge, UK, 113-153.

McKeever, S. and M. Ridge (2006), *Principled Ethics: Generalism as a Regulative Ideal*, Oxford University Press, Oxford.

Sellars, W. (1963), 'Philosophy and the Scientific Image of Man' in W. Sellars, *Science, Perception, and Reality*. Ridgeview Publishing Company, Atascadero, California, pp. 1-40.

Thomson, J. (1971), 'A Defense of Abortion', Philosophy and Public Affairs 1/1, pp. 47-66.