Thoughts as Data, Thoughts as Code: Natural Goodness and a Model of the Will*

Peter Eichman

April 2007

In *Natural Goodness*, Philippa Foot presents a conception of human moral goodness that places it on a continuum with other sorts of goodness of other sorts of living beings. She holds that human moral difference is no different in kind from the goodness of ants or oak trees, it is simply the goodness of a faculty unique to the human mind: the rational will. In this paper, I adopt Foot's position regarding species-specific goodness. My main project will be to suggest a computational model of the human will that could underpin Foot's conception of human goodness. The most important part of my model of the will is a distinction between mental content that is read and mental content that is executed. I conclude with some further speculative thoughts about consciousness that this (partial) model of the human mind suggests.

1 Natural Goodness

Foot's central thesis is that there is a natural or autonomous goodness that is specific to each species[1]. Anything that a creature does will either be in accord with the natural goodness of its species or not in accord with that goodness. This goodness is related to not just an individual, but to the general life cycle or natural history of the individual seen as a member of some species.

Foot begins discussing this idea of natural goodness by discussing attributes that philosophers might dismiss as mere biology. However, she holds that the same sort of natural goodness that she is talking about when addressing the natural histories of plants and animals also applies to human

*Special thanks to Robert Shanklin, who reviewed a draft of this paper and made many helpful suggestions

[1]It should be understood that I am, as Foot is, talking about a rather general notion of species, not any specific technical notion used by biologists. A better gloss might be "life form", as is used by Michael Thompson, but I follow Foot in using species throughout. See also Foot (2001, p.15 note 14)

morality. Foot sees human beings as not so much removed from the other animals as other theories of moral philosophy may seem to paint us. Just as there is a natural goodness of humans in terms of biology, she thinks, there is also a natural goodness of humans that applies in the moral domain. By moral domain here, I simply mean the set of things that can be said to be morally good or bad.

Foot's claim is that goodness in the moral domain is tied to the natural goodness of the human will. This is rather vague, and I am at least somewhat sympathetic to those who see its vagueness as a reason to doubt that there is real continuity between the natural goodness of a tree or a wolf and the "natural goodness" of the human will. What I propose to do in this paper, however, is to suggest an account of how Foot's idea of natural goodness and defect in the human will might be spelled out in more concrete and empirical terms. While this is itself not an empirical project, the model I will be describing is the sort of thing that I think may stand or fall by empirical lights. I take this to be a positive feature of my account.

1.1 What This Is Not

First a word about what Foot's project (and my extension, my project) is not. Foot is not concerned with giving a complete conceptual account of the English word "good"[2]. It is true that one of the motivations behind the idea of natural goodness is her observation that we use the same word to talk about goodness in plants animals, (mere) physical goodness in human beings, *and* moral goodness in human beings. But beyond this motivation for the project, Foot is not concerned with what the conceptual analysis of the word "good" is going to come down to. Instead, her project has a rather more empirical flavor. What makes an action or a being good is the degree to which it is in accord with its species' natural good[3] . This, in turn, will depend on some empirical examination of species in order to determine what its natural good is. We cannot know *a priori* what specific things are going to be important to a particular species' life cycle or natural history, we have to observe the species and its natural history to find these things out.

In addition to not being about semantics, I am also not concerned about the problem of analyzing the boundaries of the moral domain. That is, I will take it as a given that there are some

[2]Nor, presumably, the German word "gut", or the French word "bon", and so on.
[3]Specifically, I will claim, good actions are the sort that come from good beings.

things that are in the moral domain and that there are some things that are not, and that we may have an intuitive sense of what distinguishes these things, but I am not going to try and give an analysis of these conditions. In doing so, I am following the lead of Aristotle, who also took it for granted that there were things in the moral domain, but did not concern himself with explicitly describing the boundaries of that domain (cf. *Nicomachean Ethics*, Bk. 2 Ch. 2).

1.2 The Role of Species

It is the goodness of beings that comes to the forefront of under this sort of moral evaluation, not the goodness of actions. The goodness of actions does come up, but the goodness of an action must always be evaluated in the context of the species of the being that is performing that action[4].

At this point, one objection that could be raised is that we must have at least some kernel of a conceptual notion about good and goodness if we plan on trying to find it in the world around us. This much I grant: we will need some conceptual starting point for our empirical investigation. Broadly put, the notion of goodness that we can use to get some account off the ground is going to involve something which speaks to the interests of the creatures involved. By interests, Foot proposes the Aristotelian notions of nutritive and reproductive function. But we should remember that although these are our principles at the outset, they will be subject to revision based on feedback from our investigation of actual species.

Like Foot, I am baffled that others might object that what species a being belongs to should play no role in determining goodness, even moral goodness (Foot, 2001, pp.37,51). I think what has often been overlooked—probably because it seems so obvious that it does not seem worth mentioning—is that when we discuss morality, we are often implicitly restricting our discussion to just the human species. That is fine, a critic might say. Humans are the only candidates for moral agency, so when we talk about morality, we should be talking about just humans. But then it seems my critic has just admitted that species must play some role in determining goodness. If a being is human, then it is subject to evaluation in terms of moral goodness or defect. That is because

[4]I will leave open the possibility that it may turn out that there are some actions that will be universally wrong; i.e., wrong for a being of *any* species to do. However, since these would appear to be the special case, I am leaving them to one side for the purposes of the present discussion

there is something about human beings that make us apt for this sort of evaluation, and excludes all other species. But I see no difference between this instance of goodness depending on species, and goodness depending on species in any other case. Imagine evaluating the goodness of a cobra based on its venom (i.e., how well the venom benefits the cobra by helping it subdue its prey). Since we as human beings do not produce venom, it makes no sense to judge us on the goodness of our venom. But it is quite appropriate to judge a cobra on just that criterion. Likewise, as a human being, I have a rational will, so it makes sense to judge my goodness (in this case, my moral goodness) based on the goodness of my rational will. But for other animals, who by hypothesis do not have rational wills[5], it makes no sense to judge their goodness based on the goodness of their (non-existent) rational will.

One source of the discomfort that a lot of moral philosophers have about putting moral goodness on a continuum with other sorts of natural goodness may have to do with a certain perceived inequality between the two sorts of goodness. Natural goodness in the non-moral domain seems much easier to accept than in the moral domain. It is easy to point to the physical features of a creature and appeal to its biology and natural history when making an evaluation of goodness or defect.

But when it comes to persons—not the human being, but the *person*, the moral agent—it seems hard to find something to point to, and hard to find some empirical way to evaluate that thing for goodness and defect, even if you could point to it. Foot wants to point to the rational will as the thing which can be evaluated for goodness and defect, but her account leaves it is somewhat hard to see how that could be done.

I think Foot has taken a step in the right direction, but there is a further step that needs to be taken to address this objection. Foot puts forward the thesis that the moral goodness of human beings is to be equated with the goodness of the human rational will. However, she does not go on to give a good account of the will[6], at least in terms that would bring it closer to things like arms, legs, or venom. I would like to try and push Foot's idea of goodness of the human will that one further step, and suggest an account of the natural structure of the will, or at least, the structure

[5] At least, not the same sort as humans.

[6] And she is not alone. A great many philosophers appeal to "the will", and then leave that term mostly unanalyzed.

of its relation to our faculties of moral reasoning and judgment.

2 The Rational Will

Foot is in particular concerned with the rational will, which is to say the will that is normally controllable by reasons. I take this to be one of the primary features by which she distinguishes human beings from other animals, which will be important when it comes to determining what natural goodness for our species is. We are the only beings with a rational will, and it is the goodness or defect of this that determines the facts about whether we are morally good or not.

2.1 Weakness of the Will

If we are going to be addressing the goodness or defect of the human will, an obvious topic is that of *akrasia*, or "weakness of the will". There are two glosses of this phenomenon that often come up. One is that we knowingly choose the bad, and the other is that we know the good but fail to choose it. The first formulation could be more formally put as "S knows A is bad but does A anyway", whereas the second could be put as "S knows B is good but fails to do B". An important distinction to notice is that the second formulation seems agnostic with respect to whether S actually does anything. S may just fail to do B (and not do anything else), or S may do C instead of B. When considering this second formulation, I will be adopting the first reading in what follows; that is, the one where the agent S does not act at all.

From this it seems we can describe *akrasia* as picking out the two extremes or vices[7] that the human will may suffer from. In the first case of *akrasia*, it seems that the will allows too many things to be done. We see that A is bad, thus we have a reason for not doing A, but we do it anyway. The second case appears as the reverse; we see B is good, thus we have reason to do it, but our will again fails us, as we do not do B. If we consider these two extremes as the vice ends of an Aristotelian virtue (i.e., the mean between too extremes; see also Book II Ch. 6 of the *Nicomachean Ethics*), then the "good" or virtuous will is the one that lets neither too many nor too few actions get done.

[7] In the Aristotelian sense of the word.

What remains, then, is to figure out what it means for a particular action to be the sort that the will should allow or the sort that it should deny. That is, if we adopt the idea that a healthy human rational will should permit all and only good actions to be done, we still need some criteria for determining whether a given action is good or not. For Foot, this will be determined not by *a priori* theorizing about morality, but by investigation into the nature of a particular species (in this case, human beings). This is where we connect a discussion of the goodness of a being with the goodness of actions. The former is where Foot begins her discussion of natural goodness, while the latter is where many ethical and metaethical theories are focused. By suggesting a possible account of the human will, I am attempting to solidify the connection between these two areas.

2.2 Whence Normative Grip

A related issue to this question of fitness or defect in the rational will is the question of normative grip; where does morality get the hold on us that many[8] feel it has? At first glance, we might say it has something to do with the rational will. I think this is true, but something further must needs be said. For merely having our will (and thus, I take it, our actions) controllable by reasons does not seem like quite enough.

There seems to be something morally more admirable about someone who would be *able* to will the bad actions but still only wills and does the good actions than someone who does good actions because they can *only* will good actions. In other words, the first person is perfectly capable of performing bad actions, but they do not choose to do so, whereas the second person has no choice; they are "hardwired", as it were, to only do the good actions. We might characterize the second person as innocent, unworldly, or naïve, and those might be seen as positive attributes in some lights, but they still do not confer the sort of moral admiriablity that the first person has. Furthermore, the first person seems to have some sort of moral responsibility that the second one lacks. The first person, the one with a choice, is the one who has full moral agency, and it is because we think that we as human beings are more like the first person than the second that morality matters to us. That is, morality matters because we can choose to be immoral.

[8] And not just philosophers, but people in general.

I am not sure that notion of the rational will alone can capture this aspect of choice. If we take rationality as simply deciding between (two or more) options according to some rules[9], then it seems that we could build a computer program that would fulfill these requirements. In fact, it seems likely that such a computer program would always choose exactly in the manner that the rules dictated. So the computer program would be the same in this respect as our naïve person; that is, it would lack something that allows us to describe it as a full-blown moral agent.

Here I need to make a distinction between what I broadly introduced as the ability to choose (as in, the ability to choose to be immoral) with the notion of a decision that I introduced in the computer example. The difference here is not in the nature of choosing versus deciding, but rather what sort of things these are choices between[10]. When I say "decide", as in the computer program deciding between two options based on some rule set, what I mean is a first-order choice; that is to say, a choice between things that are themselves not the result of some prior choice. In the case of the will, it is the first-order choice of whether to act or not. But when I speak of "choosing" in the context of choosing to be moral or immoral, what I mean is a second-order choice; namely, the choice of whether to accept the first-order choice of the will. Thus it should be obvious why I think there must be something more than the will at work in the full course of our moral deliberations.

G.E. Moore made famous the claim that it makes no sense to ask of something, if it is good, "ought I to do it?" However, I think that there is something that needs to be examined here. Moore is claiming that once we have reached the conclusion that something is good, there is no room for any further reflection about whether we ought to do it. However, if we take seriously (as Foot and I both do) the possibility of the amoralist, then some further explanation of this relation between the good and the ought is needed.

Let us assume we have a properly functioning[11] rational will, so that we can see the good actions as good and the bad actions as bad. That is, all of our first-order choices come out correct. But now, if we assume that there is some second-order choice that can be made, we have room for the

[9]This is a simplification of a more fully-formed theory of rationality, to be sure, but the concept of choice based up reasons or rules is what I take to be at the heart of such a theory.

[10]The "domain" of choice might be another way to think of this.

[11]By properly functioning, I am not intending to introduce any sort of functionalism or proper functionalism. What I have in mind is something more along the lines of a "fit" or "healthy" rational will.

amoralist again. In most people and in most cases, the second-order choice will be to accept the first-order choice; thus, the feeling that Moore presumably had of good passing directly to ought. But I think the mere fact that we can phrase the question "ought I to do the good?", and especially that we can respond "no",[12] shows that something more must be at work here.

Now consider the case where someone has an improperly functioning rational will. Let us say specifically that it allows too much; some things which are in fact bad are conceptualized as good. Unchecked, this would lead to bad acts. And if Moore is right that there is no room for anyone to ask the question of "ought I to do the good?"[13], then this will lead this person to believe that they ought to commit what are, in fact, bad actions. But if we allow this person the ability to ask "ought I to do the good?", then other factors (prudential reasoning and so forth) may be able to prevent them from doing even what they see as good. This is hard to imagine, as I assume most of us have a (reasonably) functional rational will. But for instance, imagine a pedophile who truly, *truly* sees sexual conduct with children as good, but has learned to control his actions and not pursue sexual contact with minors for prudential reasons. That is, he has a pathological problem with his rational will, not the mere intermittent failures that most of us have from time to time when we experience momentary weakness of the will. Without the ability of some other faculty to vet the choices of the rational will, it should be impossible for such a person to not act on what he sees as good. But since people are able to do just this, there must be something in the gap between good and ought.

3 A Model of the Will

I will now suggest a model of a portion of the human mind. This model connects our faculty of moral judgment to our will. But first, there are a few assumptions that I am making that I should explain here. I take the human mind to be a computational system, modular, and representational. Each of these three assumptions requires further explication.

[12] Just because we can deny that we ought to do the good does not mean that we are correct in doing so. If we have a properly functioning rational will, we should certainly pay attention to it. What I am interested in is the fact that there seems to be some conceptual space wherein we can at least consider not acting in accordance with our rational will.

[13] That is, the thing that my rational will has presented to me as the good.

There are two things that one could mean by calling the mind computational. On the one hand, it could simply mean that the functioning of the mind may be modeled by explicitly computational processes (e.g., Turing machines or finite-state automata). That thesis makes no commitment about the *actual* structure of the mind. On the other hand, calling the mind computational is to make a claim about the structure of the mind. On this view, not only can the mind be *modeled* as a computational process, but it actually *is* a computational process, one which is (most likely) instantiated by the brain. I hold the latter of these views, the stronger version. While there are aspects of my suggested model of the will might very well survive under just the weaker claim, one of my aims is to suggest something that has empirical connections. Thus, I am suggesting not merely that a portion of the mind could be modeled in the way I will describe. Rather, I am making a claim about the mind's *actual* structure, instantiated by actual brains, and thus potentially verifiable[14] by empirical science. Again, I consider the possibility of verification a positive feature of the theory.

By a modular mind, I am adopting Carruthers' notion of mental modules[15]. That is, I am assuming that the mind is composed of a number of "isolable function-specific processing systems, all or almost all of which are domain specific (in the content sense), whose operations aren't subject to the will, [...] and whose internal operations may be inaccessible to the rest of cognition" (Carruthers, 2006, p.12).

When I say that the mind is representational, what I mean is that the objects that the mind operates on are mental representations. They can stand for concrete things in the world, abstract concepts, and so forth. Of particular importance to my account of the will is the assumption that there are parts of the mind that act on these representations in one way while there are parts that act on the very same representations in a different way. That is, the representations themselves are neutral with respect to which operations of the mind are using them.

The specific difference that is crucial here is the distinction between reading and executing. Representations or symbols are in themselves agnostic as to whether they are "mere data" or

[14]Or disconfirmable, of course.

[15]These sorts of modules will not be limited to the peripheral input or output modules in the way that classic Fodorian modules are. Instead, they are the sort of building blocks that Carruthers and others take to make up the entirety of the human mind, spanning both central and peripheral processes. These blocks are also thought to be the result of a long process of evolution and natural selection.

"executable code". In computers, at the very bottom, all of the symbols that make up both data files and running programs are drawn from the same set. What makes them different, and makes one a data file and the other a running program, are the two different operations that the computer is using them for. The distinction is roughly a declarative-imperative distinction. The symbols in a data file are interpreted like declarative statements; call this operation "reading". On the other hand, the symbols in the program are interpreted like imperative instructions; call this operation "executing".

With this in mind, it is easy to look at McCann's thesis that volitions are thoughts whose propositional content is willed from a computational perspective (McCann, 1974). Such an interpretation would be that there are certain mental representations that are sometimes read (and thence are thoughts) and that are sometimes interpreted as instructions, i.e., executed (and thence are volitions).

Before proceeding, I wish to make clear the relationship between the aspect of the mind that Foot calls the rational will, and the mental faculty of the will that I will be discussing in what follows. As we have seen, Foot thinks that moral goodness and defect in human beings is linked to the goodness and defect of their rational wills. So far, I have been dealing with the rational will as the seat of moral decision-making. That is, it is the part of our minds that decides whether a given action is morally good or bad. In what follows, I continue to hold that the rational will contains just such a moral decision-making component. However, I am also expanding my conception of the will to also include a connection to the actual planning and doing of actions. This, I feel, comports well with the notion that "willing an action" is at the root of volitional, intentional actions. What I am suggesting is a possible cognitive architecture for the will. Since the context of the present paper is morality, the actions I am focusing on will be ones usually considered to be in the moral domain.

The part of the will concerned with "willing" (i.e., executing) volitions is presumably connected to some sort of planning or plan-forming mechanism in the brain, which constructs and carries out additional instructions based on the original volition. I do not plan on discussing any of this here, as this belongs more rightly in a discussion of the theory of action. What I do want to discuss

is how this portion of the will may be related to the part of the rational will that makes moral judgments. For clarity, I will call the former portion the active will.

I envision a system with three major components: moral rules, some sort of matching procedure, and the active will. Thoughts come into this system as input[16], and then the matching procedure compares them to some list of moral rules. I am being very loose for now with how I describe this list, or the process of comparison. Ultimately, I believe that the list of rules will be generative in nature, and in many respects similar to the syntactic rules of grammar that allow us to judge the acceptability of novel sentences (see Hauser (2006) and Dwyer (2006) for accounts of what this generative moral faculty might look like). However, I am not going to defend this particular conception of moral rules here.

Once the matching procedure has returned a result, one of two things can happen to the thought. If the matching procedure returns "No", then the thought remains a thought. However, if the matching procedure returns "Yes", then the thought becomes a volition. Nothing in the (propositional) content of the thought changes, however; it merely becomes earmarked as "executable instruction" by the mind.

3.1 Modes of Failure

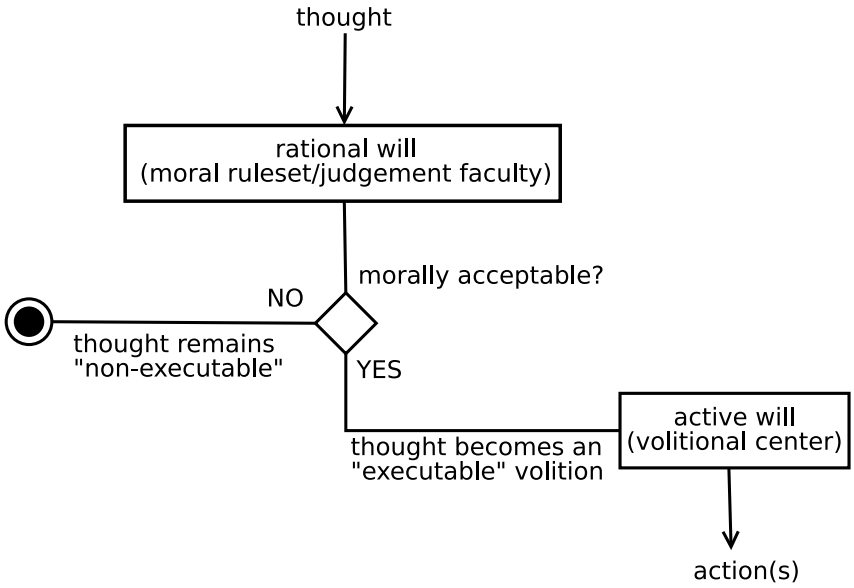
For a system of this sort, there are three immediately obvious ways in which it may fail. The rules may be defective, the matching procedure may incorrectly match thoughts and rules, or the will may not execute the resulting volitions. I will look at each of these possibilities in turn.

The first way that this system may fail is through having defective rules. As I said earlier, I understand these rules to be part of a generative system with important similarities to our linguistic faculty. Thus, one candidate that is immediately eliminated from the set of possible sorts of rule lists is some explicit set of individual moral propositions. If this is what the rule set amounted to, then it is easy to see how it could: it could either be incomplete (i.e., leave out some rules), or be "overcomplete" (i.e., contain bogus rules).

But how then might generative rules fail? The first answer that presents itself is that they

[16]I am unconcerned with exactly *how* they get selected for input to this system, since that is the problem of determining what is or is not in the moral domain.

Figure 1: A computational model of the will



produce the wrong outcome. But this can't be all there is to say, for all this does is merely shift our question to what the wrong outcome of a moral rule is. And indeed, there is something substantive to be said here: a moral rule produces the wrong outcome just in case the volition (i.e., the thought interpreted as an "executable instruction") that it creates causes (or would cause) the agent to act in a manner that is contrary to its species' natural good.

The second way in which this system may fail is if the matching process is defective. If we assume a generative system of rules, it may turn out that the matching process is largely subsumed into the "rule set", which could function either by trying to construct an identical thought to the input based on the generative rules, or else try and deconstruct (i.e., parse) the input thought according to those same rules. And of course, the matching system may fail by either letting through too many things (false positives) or by blocking some things that should be let through (false negatives).

Finally, we have the failure of the active will. There are really two ways that the active will can break down on this model. It can either fail to execute the volitions that are sent to it, or it can execute thoughts that are not supposed to be executed. If the active will fails to execute

the volitions sent to it, we have the classic weakness of the will, assuming that the rules and the matching procedure are functioning correctly. The other way in which the active will can break down would be to execute thoughts that are not actually volitions. The easiest such case to imagine is that if there is a thought that is rejected by the rules and matching procedure (and hence is not marked as a volition), but gets executed by the will anyway[17]. In contrast to weakness of the will, this would be a sort of "promiscuity of the will", where we involuntarily act on our impulses, even if we know them to be wrong.

If any part of this system is defective, then the entire system is defective. This is a principle that is not unique to the current model of the will under discussion. Rather, I take this principle to be a given for any complex system that is made up of a number of (at least partially) autonomous parts. A defect in any of these parts "infects" the whole, in a manner of speaking. And while other parts may compensate for that original defect[18], the fact remains that there is something defective about the system.

In the case of the will, we cannot be "saved" by weakness of the will if our rules are wrong. Even though we did not act on our detrimental volition, it was still there. And in point of fact, merely the fact that our will itself was defective was enough to cause the whole system to be defective. So there is no way for this system to allow two wrongs to make a right, as it were. This seems to me a good result, and I think Foot would agree, since it is in line with Aquinas' principle that seeing a bad action as good does not excuse, but neither does having weakness of the will absolve one from badness (Foot, 2001, p.73).

3.2 Connections to Metaethics

There are at least two ways that I see all this cognitive hypothesizing relating to Foot's metaethical program. The first is that my view makes use of her criterion of a species' specific natural good in order to evaluate the functioning of the system of moral judgment and volition. Good actions are those that are in accord with the natural history and life cycle of the species that a being is

[17] An analog of this in computer programming would be the class of program exploit known as "buffer overruns". In such an exploit, a malicious programmer is able to input data to a program in such a way as to cause the program to interpret some of that data as code, and execute it instead of just reading it.

[18] In other words, "workaround solutions". I think that name is telling.

a member of. And good actions are those that are judged good by a healthy rational will and executed by a healthy active will. If the volitions and thus the actions that the system produces are contrary to the natural good of the species, this is due to a defect in either the rational or the active will. The defect of one part of the mind carries over to the entire individual, as outlined at the end of the prior section, making the individual defective.

The second relation to Foot's program is that this system provides a launching point for an empirical investigation of the human moral faculty. One of the virtues of Foot's program as a whole that it hews to empirical data for its definitions of what goodness with respect to a particular species is. The place where I found a gap in her theory, though, was in the discussion of the human will. Here, her attention to empirical detail seemed to slip, and I would like to suggest a computational model of the will's relation to moral judgments. This model is, of course, not itself an empirical result, but it is the sort of thing which may be implemented by the physical system of the brain.

There are three broad ways in which this model may hold up (or not) to empirical investigation: confirmation, compatible results, or dis-confirmation. Confirmation could be claimed if it turns out that neurological, neuropsychological, or other sorts of brain experiments show that the structure of the brain instantiates something that must be like the system described by this model. On the other hand, the experimental results may show that, for instance, there are parts of the brain implicated in moral decision making and moral actions that are not covered by my model. With results of this sort, there are two possible cases. It may be that my model is still compatible with the observed functioning of the brain. If so, that is great for my model (though of course it may then need revision). However, experimental investigation may also show that what is instantiated by the brain is nothing like this system. If that is the case, then so be it. Like Foot, I am comfortable with letting empirical facts guide my philosophical theories.

4 Coda: On Consciousness

One further thought that is only tangentially related to this problem has to do with the nature and mysteriousness of consciousness. I propose that, like the computational system sketched above, consciousness is also a computational system. Specifically, consciousness functions like a debugger

program that gets hooked onto another program. The debugger can examine the data that that program has access to, as well as observe the instructions that that program is about to execute. That is to say, to the debugger the code that makes up the representation of a running program is simply another source of data. The debugger does not execute these instructions, it just reads them. Moreover, the debugger can change any of that data and alter any of those instructions.

But it is important to remember that the debugger[19] is itself merely a program, as opaque as the first program was before we hooked the debugger into it. That is, we can see what goes in as input and what comes out as output, but we have no way of seeing what is going on inside the process itself. And by analogy, consciousness itself is just as opaque as the mental process to which it attaches. While consciousness allows us to "look inside" other mental processes, it cannot look inside itself. This would explain some the mystery of why it is so hard for us to introspect our own consciousness. In the case of a computer, we could in principle attach a debugger to the debugger to find out what is going on in it, but in the human mind, this seems to be a (near) impossibility[20].

There is one way in which this idea of consciousness as a debugger of mental processes is more directly related to my above model of the will, and to problems in ethics and moral psychology in general. Recall that I suggested that there are not one, but two moral decision-making processes. The first makes first-order judgments about whether a proposed action is good or bad, while the second makes second-order judgments about whether to accept the first-order decision. In the preceding section, I described how I envisioned the first-order decision-making process works in conjunction with the active will to execute moral actions, but I was mostly silent about second-order judgments.

After the rational will decides that the thought describing some proposed action is okay and this action gets marked as an executable instruction (i.e., is transformed from a thought to a volition), it is passed to the active will. But it is precisely at this point of transmission from the rational

[19]Following Jackendoff (1997), I take language to be of great importance to the issue of consciousness, and in particular, of conscious reflection. Thus I would speculate that our linguistic faculty is our mental "debugger". However, a further discussion of this idea is beyond the scope of this paper.

[20]I hedge this with "near" here in order to not come down one way or the other on whether people who claim to be aware of their own consciousness really are experiencing this sort of second-order consciousness. Perhaps they have the ability to attach a second copy of their debugger to the first debugger. But the fact remains that most of us never experience second-order consciousness, so we seem to be stuck with just the one debugger.

will to the active will that is possible for the consciousness debugger to intercept the new volition. And instead of treating it as an instruction, the way the active will does, the debugger treats it as a descriptive piece of data. This raises the question of how this process is any different from the decision-making process of the rational will.

The difference is that the first-order decision-making process that is performed by the rational will is a non-conscious process. It is a fully automatic process, whereas when the debugger is reading the volition, it is conscious. Of course, the notion I have of consciousness here is not supposed to be very deep or mysterious; it is simply that the debugger is able to present to us the instruction in a conceptualized form, via language or "inner speech". Once we have this conceptualized version of the volition, we can apply conscious reasoning to it, and possibly even override it.

One criticism that is sure to arise is why we should think that there are two decision-making processes in the first place. After all, the reasoning goes, wouldn't it be simpler to just have one conscious, rational decision-making process, rather than have a two systems, where the conscious system has to constantly monitor the non-conscious one? My reply is that yes, from a certain perspective, if we were designing a creature from the ground up to be a conscious, rational, moral agent, then it might be simpler to just provide this being with one conscious system of moral decision-making. But there are two observations that weigh against this simplicity, and in favor of a two-system model.

Both of these observations are broadly empirical. The first is that while if we are building a rational creature from scratch in a lab somewhere, it might make sense for simplicity's sake to stick to one decision-making system. But that is a far cry form the conditions under which human beings evolved. Evolution cannot just create new and completely different systems wholesale; it must modify (and in relatively slight ways) existing parts of the creature. It is, I think, relatively undisputed that we take consciousness (and especially the linguistically-oriented sort of self-reflective consciousness I mention above) to be a rather recently added feature to primate (and maybe exclusively human) minds. But it is plausible to think that the capacity for fairly complex decision making exists without consciousness (Jackendoff, 1997). So it should not surprise us that evolution might give rise to an underlying non-conscious (probably domain-specific) decision-making system

before giving rise to a more general conscious system.

The other fact is that throughout the human mind and brain, we have observed many examples of dual systems, where there is a newer (and slower) conscious system that appears to duplicate some of the functionality of an older (and usually faster) non-conscious system (Carruthers, 2006). Perhaps the most famous example of this is the phenomenon of blindsight. Subjects with blindsight report no conscious visual stimuli, but are able to perform certain tasks requiring visual input at much better than chance levels. Since dual systems are by no means uncommon in human (and indeed, other animals') minds, it does not seem like a stretch the think that our moral decision-making system might have this character.

All of this (admittedly speculative) discussion of consciousness, debuggers, and dual-systems has taken me away form my original point in this paper, so I would like to return to it for some closing remarks. In this paper, I sketched a possible computational model of the will, focusing on how it deals with moral choices. If nothing else about this model survives further scrutiny, the main point I would like to get across is the usefulness of being able to think of representations, be they thoughts or volitions, in a unified way. There is nothing in their content that distinguishes them, but only how they are treated by the rest of the human mind.

References

Aristotle. *Nicomachean Ethics*. Trans. by Terence Irwin. Indianapolis, IN: Hackett, 1999.

Carruthers, Peter. *The Architecture of the Mind*. Oxford: Oxford University Press, 2006.

Dwyer, Susan. "How Good is the Linguistic Analogy." *The Innate Mind, Vol. 2: Culture and Cognition*. Ed. Peter Carruthers, Stephen Laurence, and Stephen Stich. Oxford: Oxford University Press, 2006. 237–256.

Foot, Philippa. *Natural Goodness*. Oxford: Oxford University Press, 2001.

Hauser, Marc D. *Moral Minds*. New York: HarperCollins, 2006.

Jackendoff, Ray. *The Architecture of the Langauge Faculty*. Cambridge, MA: MIT Press, 1997.

McCann, Hugh. "Volition and Basic Action." *The Philosophical Review* 83 (1974): 451–473.