On the perceived objectivity of some moral beliefs.

Graham Wood

School of Humanities, University of Tasmania, Launceston, TAS, Australia

This is an 'post-print' version. The Version of Record of this manuscript has been published and is available in the journal *Philosophical Psychology* Volume 33 (2020) Issue 1: https://www.tandfonline.com/doi/full/10.1080/09515089.2019.1696454

Abstract:

This paper presents research in moral psychology and draws on this research to offer an account of the cognitive systems and processes that generate the perceived objectivity of some moral beliefs. The paper presents empirical research on the perceived objectivity of moral beliefs, compares different algorithms employed by human cognition in the context of model-free and model-based reinforcement learning, and uses concepts drawn from dual-system and modular theories of cognition. The central claim of the account is that belief in the objectivity of some moral beliefs results from certain 'modular' features of cognitive systems.

1. Introduction

Empirical evidence indicates that people perceive some of their moral beliefs to be objective (Goodwin & Darley 2008). This paper offers a psychological explanation of the perceived objectivity of moral beliefs. It begins by reviewing research (Goodwin & Darley 2012) that examines how lay people think about the objectivity of their moral beliefs. This empirical research finds that while there is variation in how objective people hold their different moral beliefs to be, some moral beliefs are held to be almost as objective as scientific facts. Just what it means for moral beliefs to be objective is not straightforward. But for the purposes of this paper it will be assumed that objectivity is a categorical notion, meaning that some state of affairs is either in the category 'objectively the case', or it is not in that category. This assumption will lead to some tension within the following analysis because the paper engages

with empirical research that does not seem to assume that objectivity is a categorical notion. But, given the different theoretical perspectives in play, this tension cannot be avoided. Furthermore, there is an important difference to note between the objectivity of a scientific fact and the objectivity of a moral value. The perception of the scientific fact (in and of itself) does not motivate action, while the perception of the moral value does motivate action. This means that the perception of a moral value essentially has a motivational component. This motivational component has been insightfully described as the moral wrong having a "not-to-be-doneness somehow built into it" (Mackie 1977, p. 40).

This prompts the question: what explains the perceived objectivity of particular moral beliefs? A straightforward answer is that the moral domain is a domain of objective truths and people are simply correctly identifying particular objective moral truths. That would explain the (veridical) perception of the objectivity of these moral beliefs. But this paper will not concern itself with the question of objective moral truth. More precisely, this paper will make no philosophical argument either for or against the objectivity of particular moral truths.

This paper will propose that whether or not a particular moral claim is perceived to be objective is dependent upon how the content of the moral claim was actually learnt by the individual (and the paper further proposes that different individuals can learn the content of the same moral claim in different ways, and this explains why individuals can disagree about the objectivity of the same moral claim). Claims learnt using model-free reinforcement learning are perceived as objective while claims learnt using model-based reinforcement learning are not. And the cause of the perception of objectivity relates to certain 'modular'

features of the process of learning. So, again consider the question: what explains the *perceived objectivity* of particular moral beliefs? In order to address this question three themes from psychology and cognitive science will be considered.

Firstly, a series of distinctions, drawn by Cushman (2013), from within computational neuroscience are introduced. Cushman draws a distinction between valuing the intrinsic status of actions, and valuing the expected consequences of actions. He then discusses two algorithms employed by human cognition, one that encodes the value of actions, and the other that encodes the value of outcomes. Cushman then links the distinction between valuing representations of actions and valuing representations of outcomes with two forms of reinforcement learning, model-free, and model-based reinforcement learning, respectively.

Secondly, dual-system theory is introduced. The central idea of dual-system (or dual-process) theory is that there is more than one system (or process) within the mind. The central idea has a long history going back at least to Plato's description of the tripartite mind in his *Republic* (2008). But a contemporary scientific research program has been developing during the past fifty years, or so (for review see Evans & Frankish 2009). I assume that there are two systems, that may have all, or some, of the characteristics as identified by Evans and Frankish (2009).

Stanovich & West (2000) characterize the distinction between System 1 and System 2 as follows. System 1 (intuition) is fast; automatic; undemanding of cognitive capacity; acquired by biology; exposure, and personal experience, and System 2 (reasoning) is slow; controlled; demanding of cognitive capacity; acquired by cultural and formal tuition. Furthermore, Kahneman (2002) describes System 1 as generating intuitive judgments that "occupy a

position – perhaps corresponding to evolutionary history – between the automatic operations of perception and the deliberate operations of reasoning".

Thirdly, two features of Fodorian modules (Fodor 1983) are introduced. The outputs of such modules are 'mandatory', and the manipulations made within these modules are relatively inaccessible to consciousness, this second feature leading to what is characterized as 'informational encapsulation'.

Having introduced these three themes, this paper will present a possible explanation of the perceived objectivity of moral beliefs. The suggestion is that the perceived objectivity of certain moral beliefs is the result of System 2's interpretation of the output of a System 1 module (where the output of the System 1 module is understood as a moral intuition, and the interpretation offered by System 2 is generated by deliberative moral reasoning.) I will suggest that the internal workings of the System 1 module are 'informationally encapsulated', and the modular output is 'mandatory'. Furthermore, unbeknownst to System 2 the process at work within the module has encoded the value of one or more actions via model-free reinforcement learning. To illustrate this process, imagine a representation of an action has a value representation attached to it (perhaps as a result of some form of social referencing in early childhood), then a process of information encapsulation occurs, such that System 2 has no access to the fact that, at some point in the past, a value representation has been attached to the representation of the action. Now imagine some time later. When System 2 receives information from the System 1 module, the representation of the action simply comes with a value representation attached. And further imagine that the information received by System 2 from the System 1 module has a certain 'mandatory' nature about it. This, I will claim, is the explanation of the perceived objectivity of certain moral beliefs.

Finally, the meta-ethical implications of this explanation are considered. If one were inclined to do so, the explanation offered here could be incorporated into a broader argument to the conclusion that the perceived objectivity of moral beliefs is an artefact of the 'mandatory' nature of the output of an 'informationally encapsulated' modular System 1 process and that there is no other reason to believe that the objectivity is 'real' in any sense that carries moral authority. But, strictly speaking, the explanation presented here only explains (assuming it is correct) the psychological origins of the perceived objectivity of moral beliefs, and thus it does not prove that moral beliefs are not objective in some sense that is independent of this psychological account.

2. Empirical Data

Goodwin and Darley (2008; 2010; 2012) have conducted a number of psychological experiments examining the perceived objectivity of the moral beliefs of lay people (where lay people are taken to be people without theoretical training in meta-ethics). Goodwin and Darley (2010, p. 165) claim people "have a set of underlying intuitions about the objectivity or subjectivity of their ethical beliefs, and that these intuitions exist without prior explicit theorizing." Furthermore, their research uncovers a number of interesting findings (2012, p. 252-3). Firstly, not all moral beliefs are perceived as equally objective. Secondly, knowledge of what other people think about certain moral beliefs contributes to how a person themselves thinks about the objectivity of those moral beliefs. This suggests that the process that a person goes through to arrive at a judgment of the objectivity of a moral belief involves consideration of evidence wider than knowledge of the content of the belief itself. In other words, it does not just matter what the person themselves thinks, but it matters what other people think. And finally, if a person has come to the view that a moral belief is objective

they will not be comfortable with the fact that other people disagree with their judgement. This suggests that whatever the process is within their mind that has generated the perception of the objectivity of a moral belief, that process is a powerful one, and one that might trigger strong negative affect if implicitly or explicitly challenged by the contradictory beliefs of others.

3. Advancing a dual-system account

The first empirical observation I highlighted above was that not all moral beliefs are perceived as equally objective by those holding them. There are important aspects of this observation that need to be addressed before proceeding. Goodwin and Darley's experimental design allows for the concept of 'objectivity' to be interpreted in a certain way. However, there are at least two possible interpretations of 'objectivity' that both need to be highlighted. Firstly, objectivity can be taken to be a categorical concept. Meaning that something is either objective, or it is not objective. Or, secondly, objectivity can be taken to identify one end of a spectrum, with subjectivity at the other end. Goodwin and Darley seem to assume that 'objective' is one end of a spectrum, with 'subjective' located at the other end.

In this paper judgements about objectivity are assumed to be categorical judgments. Objectivity does not admit of degrees. However, as I say, Goodwin and Darley seem to be assuming within their experimental design that participants are reporting on the degree of objectivity. The fact that there is another way to conceive of objectivity (i.e., categorically) suggests another approach to empirical study, one that does not ask 'how objective' is a moral belief, but rather 'whether or not' some moral belief is objective. But given that this paper will assume that judgments about objectivity are categorical judgments, how is the spectrum of judgements implicit within Goodwin and Darley's empirical approach to be

accommodated? Here it will be suggested that Goodwin and Darley's supposed spectrum of objectivity judgements is actually made up of two components. One end of the supposed spectrum is constituted by categorical judgments of objectivity. But contrary to the assumption of Goodwin and Darley these categorical judgements are not part of the spectrum. However, every judgment that is not located at the supposedly 'objective' end of the spectrum, is part of a subjective spectrum. Later in this paper it will be suggested that the 'objective' component of the supposed spectrum is generated by model-free reinforcement learning (possibly within a modular process within System 1), while the 'subjective' spectrum is generated by model-based reinforcement learning (possibly within non-modular processes within System 2).

Finally, note that aggregated empirical data could yield 'degrees' of objectivity, but this would be an artefact of the aggregated data and not an aspect of the original judgements of objectivity by individual subjects. As the empirical evidence indicates, some people perceive certain moral beliefs to be almost as objective as scientific facts. Here I will assume that the objectivity of scientific facts is categorical in nature (the 'facts' are either objective or they are not) and in this paper perceptions of the objectivity of moral beliefs will also be assumed to be categorical judgments. The moral beliefs are either objective or they are not. It is these categorically objective perceptions of moral beliefs that are the focus of this paper.

So, again to return to the question: what explains the *perceived objectivity* of particular moral beliefs? One possible explanation is that moral beliefs are being generated by distinct cognitive systems. And that the moral beliefs that are perceived as objective are being generated by only one of these cognitive systems. It is to this possibility that I now turn.

Dual-system or dual-process accounts of cognition are emerging within cognitive science (Evans & Frankish 2009), and both Cushman (2013) and Crockett (2013) claim dual-system accounts of cognition can be usefully applied to our understanding of moral judgment.

Cushman claims that "a central aim of current research in moral psychology is to characterize a workable dual-system framework" (2013, p. 273). He notes that a number of ways to characterize a dual-system framework already exist in the literature, including the set of contrasts listed in Table 1:

Table 1	Pairs of contrasts common in the cognitive science literature
Intuition	Reason
Automaticity	Control
Emotion	Cognition

Table 1: Pairs of contrasts common in the cognitive science literature

Some of the items listed in Table 1 link directly to central themes in metaethics relating to Moral Rationalism (Kant 2002) and Moral Sentimentalism (Hume 1902), particularly the contrast of Emotion and Cognition. But Cushman (2013) and Crockett (2013) focus on normative ethics rather than metaethics. For example, both refer to the distinction between consequentialist and deontological ethics as a way to motivate their own accounts. And both acknowledge how the contrasts in Table 1 play into larger philosophical debates concerning such things as the appropriateness of deontological or consequentialist ways of making normative moral judgments (see, e.g., Greene, 2013).

Cushman notes that there is "pervasive disagreement about the nature of the two systems" (2013, p. 287). Essentially, Cushman is unsatisfied with the particular dualisms that

dominate the current debate as ways of characterizing a workable dual system framework, so he explores a further possibility. His aim is to bring together core concepts from the computational neurosciences with core concepts of the dual process accounts of moral psychology and formulate the "core concepts from each domain in common terms" (2013, p. 273). In his attempt to do this, he highlights the following distinction, drawn from the computational neurosciences: valuing the *intrinsic status of actions* versus valuing *the expected consequences* of actions. And in this context, he draws particular attention to two algorithms employed by human cognition:

> One algorithm encodes the value of actions by associating them with subsequent punishment and reward, leveraging prediction error and temporal difference learning to efficiently represent whether an action is rewarding without representing what makes it so. The other algorithm encodes the value of outcomes and selects actions based on their expected value, relying on a probabilistic causal model that relates actions, outcomes and rewards. (2013, p. 273).

Cushman then links the distinction between valuing representations of actions and outcomes with two forms of reinforcement learning, model-free, and model-based reinforcement learning, and observes that:

> The distinction between model-based and model-free reinforcement learning provides a promising foundation on which to build a dual process model of moral judgment. First it accounts for the distinction between action-based and outcome-based value representations. Second it aligns this action/outcome distinction with mechanisms for

automatic versus controlled processing. Third, it specifies precisely how both cognitive and affective mechanisms contribute to both types of process. (2013, p.282)

So, in short, Cushman is drawing an alignment among a number of systems and we can consider all this in tabulated form as displayed in Table 2:

Table 2	The alignment of value representations with reinforcement learning
Action-based value representations	Outcome-based value representations
Model-free reinforcement learning	Model-based reinforcement learning

Table 2: The aligning value representations with reinforcement learning

So, while Greene (2008) linked non-utilitarian (i.e., deontological) moral judgments with "emotionally-grounded intuitions" (Cushman, 2013, p. 286) Cushman offers an alternative analysis. He claims that linking "deontological theories with action-based value representation and utilitarian moral theories with outcome-based value representation is truer to the core philosophical positions." (2013, p. 286). Cushman continues, by advocating a characterization of two systems within the moral domain that distinguish mechanisms of value representation:

> [O]ne that assigns value directly to actions, and another that selects actions based on the value assigned to their likely outcomes. This same distinction captures an essential difference between the two families of reinforcement learning algorithm. The basic principles of these reinforcement learning algorithms—and specific details of their neural implementation—provide an explanation for several otherwise puzzling

> phenomena of moral judgments, and of human judgment and decision making more broadly. Their explanatory power becomes especially broad when conceiving of certain internal mental representations as constituting "states" and certain processes of thought as constituting "actions". (2013, p. 287).

Essentially Cushman aligns deontological theories with model-free reinforcement learning (e.g., learning the moral rule 'it is morally wrong to inflict pain', a rule that does not require a model) and he aligns consequentialist theories with model-based reinforcement learning (e.g., learning that the morality of inflicting pain depends on the consequences, as represented in some relevant model). And while I have focused on Cushman's (2013) account, Crockett (2013) also draws parallels between consequentialist ethical judgments and model-based systems and deontological ethical judgments and model-free systems.

Building on all this I am interested in using this approach to explore the foundations of the perception of the objectivity of moral belief. In particular, I suggest that the perceived objectivity of some moral beliefs originates in cognitive systems that are based on action-based value representations and model-free reinforcement learning (the left-hand column of Table 2). But in order to explore this idea I want to introduce the concept of modularity. And in particular, I want to ask: Does modularity have a place in this account?

4. Expanding the Dual-System Account to Incorporate Informational Encapsulation and the Mandatory Nature of Certain Modular Outputs

An early account of dual cognitive systems is offered by Stanovich and West (2000) who characterized what they called System 1 and System 2 as follows:

> System 1 (intuition): fast; automatic; undemanding of cognitive capacity; acquired by biology; exposure, and personal experience.
>
> System 2 (reasoning): slow; controlled; demanding of cognitive capacity; acquired by cultural and formal tuition (Kahneman, 2002).

Furthermore, Kahneman (2002) describes System 1 as generating intuitive judgments that "occupy a position – perhaps corresponding to evolutionary history – between the automatic operations of perception and the deliberate operations of reasoning" thus leaving the deliberative reasoning to System 2.

One instantiation of the idea that there are two systems of cognition is the theory that some parts of the mind are modular. Fodor (1983) introduced the idea of peripheral modules that process information coming into the mind from peripheral mental systems, such as the visual or auditory systems. For the purposes of this paper it is worth emphasising three closely related characteristics of Fodorian modules. First, the outputs of such modules are mandatory, and second, the manipulations made within these modules are relatively inaccessible to consciousness. And, with reference to these first two characteristics Fodor notes that:

> It is worth distinguishing the claim that input operations are mandatory (you can't but hear an utterance of a sentence as an utterance of a sentence) from the claim that what might be called "interlevels" of input representation are, typically, relatively inaccessible to consciousness. Not only must you hear an utterance of a sentence as such, but, to first approximation, you can hear it only that way. (Fodor, 1983, p. 55)

Thirdly, and relatedly, modules are informationally encapsulated. Consciousness is not aware of, nor does it have access, to the processes operating within the module.

Could these three features explain why humans perceive that certain moral beliefs to be objective? I suggest that they could. But I should stress that it is just these three features of modules that I am relying on in this account. I am not assuming any more 'modularity' than simply the presence of these three features. I will suggest that a cognitive system (or process) that has these three features is responsible for the fact that humans perceive some moral beliefs to be objective. If, in the mind of any reader, the fact that a cognitive system (or process) incorporates these three features is not sufficient, in and of itself, to justify that system being labelled a 'module' or a 'modular process' then, so be it. This account is not committed to the use of the labels 'module' or 'modular process'. I use these labels simply as way of describing the system (or process) in question that incorporates the three features I have identified. What this account is committed to is simply the presence of these three features.

A standard example used to illustrate the activity of a module is the perception of the Müller-Lyer illusion (Figure 1).

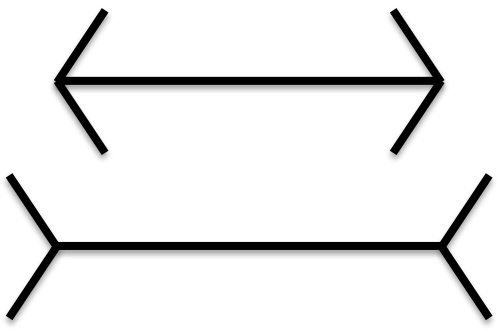

Figure 1: Müller-Lyer illusion

The two horizontal lines in this image are of equal length on the page, but they are often perceived as having different lengths (cf. Segall et al., 1966). When experiencing this illusion, the mental processes leading to the conscious perception of the image presumably must involve the manipulation of the information received on the retina, because one experiences an image of two horizontal lines of *unequal* length when on the page the two lines are of *equal* length. It is assumed that a module is responsible for the manipulation of the information between the retina and conscious experience such that lines of equal length are seen as lines of unequal length.

In the case of this illusion we are able to measure the lengths of the lines on the page with a measuring tape or ruler. So even if the illusion is created by a module in the visual system, even if we are not able to avoid the mandatory nature of this illusion, even if we cannot have conscious access to the processes that create this illusion, and, even if the process is informationally encapsulated, none-the-less, we are able to "get behind" the illusion by measuring the two lines on the page with a measuring tape.

But imagine for a moment that we were not able to measure the lines on the page with a measuring tape? What then? Imagine the only information we had access to was the information provided by the module delivering the illusion to our consciousness. In that circumstance, we might believe strongly that the lines were of different lengths. We might even ask others about their own experience, and (assuming they were subject to the same illusion) they would report a similar experience to ours. And so, with no evidence to the contrary we might be very willing to accept that the difference in the lengths of the lines was an objective feature of reality. Furthermore, if we had asked others (who were subject to the same illusion) and they had reported a similar experience to ours, what would we then think

of some yet other person (not subject to the illusion) who reported no such experience? Perhaps we would be unsettled or uncomfortable knowing that this yet other person did not see things the way we did.

And, now consider all this with reference to the case of value representations. What if a value representation was delivered to central cognition (System 2) by a modular process (in System 1), would that lead us to believe in objective value? I believe it would.

Now before I go any further, I should clarify that the account I offer here refers specifically to value representations of actions not value representations of outcomes. This is because, according to Cushman's account (and assuming he is on the right track), value representations of actions are produced by model free reinforcement learning, and value representations of outcomes are produced by model-based reinforcement learning. The account I offer here does not apply to value representations of outcomes produced by model-based reinforcement learning simply because representations of such values are necessarily embedded in a model and so lack the "mandatoriness" and related "informational encapsulation" central to my account. The account I offer here only applies to value representations of actions produced by model-free reinforcement learning, as it is these value representations that could appear "mandatory" and that could be produced by a modular process that includes the "informational encapsulation" that is at the heart of my account.

So now to the account itself. Remember the features of modules I mentioned above: they are (1) *mandatory*, (2) the workings of the modules are relatively *inaccessible to consciousness* and thus they can be thought of as (3) *informationally encapsulated*. Without the equivalent of a measuring tape or ruler we cannot "get behind" the value representation delivered by a module and so we may take it to be an objective value being represented.

Here is my central idea. Imagine a representation of an action has a value representation attached to it. The value is attached by a process of model free reinforcement

learning perhaps as a result of some form of social referencing in early childhood. Then a process of information encapsulation occurs. This means that central cognition (System 2) has no access to the fact that, at some point in the past, a value representation has been attached to the representation of the action (within System 1). Now imagine a later time. When central cognition receives information from the module (or modular process), the representation of the action simply comes with a value representation attached to it. And further imagine that the information received by central cognition from the module has a certain "mandatory" nature about it. Recall Fodor's point that "you can't but hear an utterance of a sentence as an utterance of a sentence" (1983, p. 55). Or to put all this another way, and to use the insightful phrases offered to us by Mackie, the action simply has a mandatory "to-be-pursuedness", or alternatively a "not-to-be-doneness" attached to it (1977, p. 40).

Importantly, due to the informational encapsulation, central cognition (System 2) has no other information to go on. In particular, it has no historical or causal information about how this value representation was attached to the representation of this action (within System 1). And furthermore, the value representation is presented to central cognition as mandatory. What then is central cognition to think? Due to the informational encapsulation and the mandatory nature of the value representation central cognition rationalizes this observation by attributing objective value to the action that is represented.

5. Considering the 'fit' of this explanation with the empirical research

How do my suggestions fit with the empirical research? I will turn to that question next. But first, here is a brief restatement of my claims in order to make explicit one more central assumption made within the dual-system approach to cognition. I am claiming that there are two systems of cognition in the brain, System 1 and System 2. A central assumption

within the dual system approach to cognition is that System 1 is the default cognitive system and one of the functions of System 2 is to act as an oversight and override system stepping in when necessary. In other words, System 1 has primary responsibility for decision making, and System 2 has a review function. If the outputs of System 1 are congruent with the review function of System 2 then the decisions of System 1 are not overridden by System 2. But if the outputs of System 1 are not congruent with the review function of System 2 then System 2 can override the deliverances of System 1 (Stanovich and West 2000, 662). Importantly, the deliverances of System 1 are not reversed in any way by the fact that System 2 overrides them. The System 1 judgements persist, it is just that System 2 overrules them. Putting this another way, while System 2 might cause a human to 'change' his or her mind with respect to some moral belief, that 'change' is due the human overuling his or her System 1. (There may be an independent way for an individual to change the output of his or her System 1, but this possibility is not considered here.)

I am further suggesting that System 1 systems employ model-free learning and that System 2 systems employ model-based learning. Finally, I am suggesting that there are certain 'modular' features of System 1 processes that mean that System 2 does not have cognitive access to the workings of System 1 processes.

Now let us return to the empirical research. Goodwin & Darley (2012) identified the fact that not all moral beliefs are perceived as equally objective. Cushman (2013) has suggested that model-free reinforcement learning generates some moral beliefs while model-based reinforcement learning generates other moral beliefs. And if the moral beliefs associated with the model free reinforcement learning are subject to the three features of modules I have identified in this paper, then it seems reasonable to hypothesize that a person would perceive those moral beliefs as objective. This hypothesis could be tested by further empirical research examining whether moral beliefs generated by model free reinforcement

learning are perceived to be objective, while moral beliefs generated by model-based reinforcement learning are not.

Goodwin & Darley (2012) also identify the fact that if there is consensus among a person's peers about a certain moral belief then it is more likely that a person will consider a moral belief to be objective. This fits with the dual system analysis of the situation. System 1 delivers the original judgement of moral objectivity. System 2 then reviews that judgement. As part of the review System 2 can access information about the level of consensus among a person's peers. If a person's peers agree, then System 2 endorses the judgement of System 1. If a person perceives a moral belief to be objective, and all their peers agree, then that would reinforce the belief that the moral belief is indeed objective. And finally, Goodwin & Darley (2012) report that (in direct contradiction of their own perception and the consensus of their peers), if some other person does not endorse the objectivity of the person's own moral belief, then the person in question will experience distress regarding the other person's contradictory belief. This is also to be expected. If a person, supported by the consensus of their peers, is faced with some other person who does not endorse the objectivity of a moral belief this will be distressing. To put this in terms of the dual system analysis, the person's System 1 & 2 both agree, but the person is then faced with the further knowledge that some other person does not agree. This does not lead to tension between System 1 & 2 within the individual, but rather distress about what to make of the other person.

A further issue with respect to the fit with the empirical data concerns the question of which learning algorithm (model-free or model-based) is used in any particular learning context. Could it be the case that the certain learning contexts always trigger the use of one or other of the learning algorithms? Or is there much more variability about which learning algorithm is employed in different contexts?

It may be that certain social situations (e.g., highly emotionally valanced situations) trigger model-free learning, while other social situations trigger model-based learning. Or it may be that it is not the social situation that determines which learning algorithm will be used but rather the content of the subject being learnt determines the learning algorithm used.

Thus, further empirical research could also seek to determine if there is a correlation between the content of moral beliefs and the type of reinforcement learning. For example, is particular moral content learnt by model free reinforcement learning while other moral content is learnt by model-based reinforcement learning, or is there no correlation between content and type of learning?

It may be the case that humans are 'hard-wired' to learn using model-free algorithms with respect to certain social contexts or with reference to particular content. Thus, there are a number of questions here that would benefit from further research.

Consider, for example, the "not-to-be-doneness" of incest. The taboo associated with incest is an interesting one because it is often claimed that some form of this taboo is present in all human cultures. So, it is a useful one to consider here as a moral belief perceived (perhaps in all human cultures) to be objective. In relation to the incest taboo, Sripada discusses the Westermarck Mechanism: "In 1891, the Finish sociologist Edward Westermarck proposed that humans have an innate mechanism that generates a powerful aversion to having sex with people with whom one as had extended intimate association during one's childhood years." (2008, p. 334). Sripada goes on to describe research concerning the level of sexual relations between non-biologically related individuals who were raised together in kibbutzim (communal villages) in Israel. Although non-biologically related children who were raised together in such circumstances were not explicitly taught to consider sexual relations between themselves 'not-to-be-done', "the accumulated evidence

suggests that sexual intercourse and marriage between [such individuals] was vanishingly rare" (Sripada, 2008, p. 335). This suggests to me that there was a process that occurred during the development of these children in which the value representation of 'not-to-be-done' was attached to the action of sexual intercourse with individuals with whom they had grown up (this being independent from the biological relatedness of the children). At sexual maturity the causal history of this process is long forgotten and the informational encapsulation of the modular process delivering the value representation to central cognition, together with the mandatory nature of the value representation itself, ensures that central cognition simply interprets the output as "objectively wrong".

Consider the situation illustrated in Figure 2.

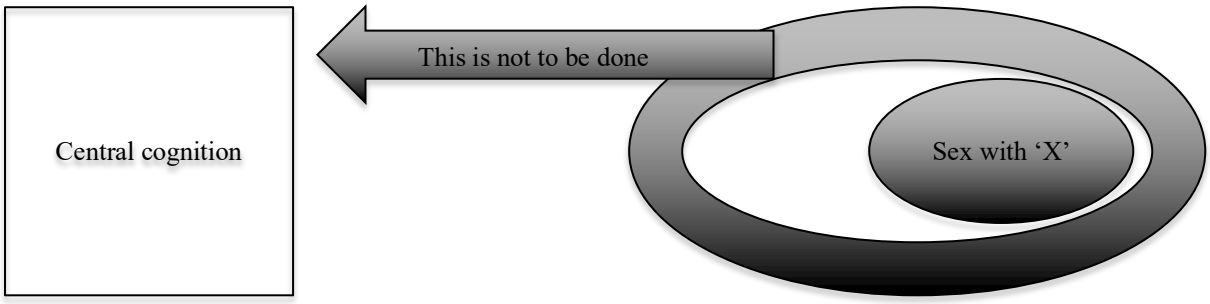

Figure 2: The mandatory "This is not to be done" output provided to central cognition via an informationally encapsulated modular process.

In the case of the operation of the Westermarck Mechanism, the module outputs a value representation, to the effect that it is mandatory that this action (Sex with "X") is not to be done. In the absence of other information (due to the informational encapsulation of the process) central cognition "understands" mandatory not-to-be-doneness as objective wrongness, as illustrated in Figure 3.

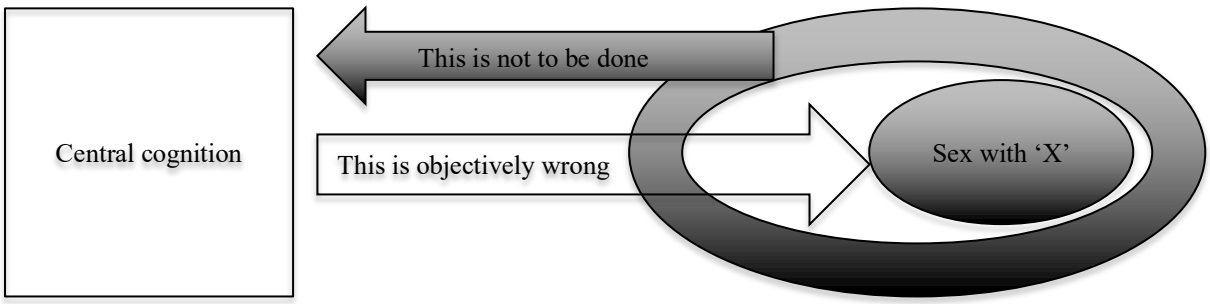

Figure 3: Central cognition's interpretation of the modular output is "Sex with 'X' is objectively wrong".

So, the mandatory "not-to-be-doneness" is interpreted by central cognition as objective wrongness and we report a belief in an objective negative moral position. (When the value representation is positive rather than negative, then the resultant belief is a belief in an objective positive moral position.)

If action based value representations are presented to central cognition as mandatory and the results of model free reinforcement learning are associated with processes that are informationally encapsulated (in the sense in which the process that results in the Müller-Lyer illusion is informationally encapsulated and the experience of the illusion is mandatory in those who experience it), then this could leave central cognition with the only interpretation available to it being the endorsement of objective moral beliefs.

Furthermore, if the thought of sex with other individuals one shared an extended domestic experience with as a child does become associated with mandatory "not-to-be-doneness", and due to informational encapsulation, that mandatory "not-to-be-doneness" is interpreted by central cognition as implying that there is an objectively wrong dimension to such sexual activity, then this provides an account of the cognitive architecture underlying the 'moral dumbfounding' and subsequent 'post-hoc rationalization' described by Haidt (2001) with reference to incest. A person is morally dumbfounded simply because they are receiving the mandatory informationally encapsulated output 'this is not-to-be-done'. In the

absence of any further information, if required to provide 'further information', then that person will provide a post-hoc rationalization.

Now, how does all this fit with Cushman's account? With respect to incest avoidance Cushman is cautious, observing that "Innate and learned action-based aversions may draw on similar psychological mechanisms, or not, at present there is little relevant evidence (but see Delgado, Jou & Phelps 2011)" (2013, p. 287). But balanced against this caution, Cushman observes that "several lines of evidence indicate that model-free value representations can be constructed on the basis of observational learning … or instruction" (2013, p. 284). For example, he mentions "aversive learning, where, for instance, a neutral cue might be paired with a painful shock" (2013, p. 284). Such aversive learning situations may tend to involve the use of model-free value representations. Furthermore, he claims that "[s]ome "evidence for a division between action- and out- come-based valuation in the moral domain comes from a study of people's aversion to pretend harmful actions, such as shooting a person with a fake gun or hitting a plastic baby doll against the table (Cushman, Gray, Gaffey, & Mendes, 2012; see also Hood, Donnelly, Leonards, & Bloom, 2010; King, Burton, Hicks, & Drigotas, 2007; Rozin, Millman, & Nemeroff, 1986)." (2012, p. 275) But the interesting question is whether such action-based value representations are the result of model-free reinforcement learning.

So, building on Cushman's claims, observational learning or instruction, can construct model-free value representations. If the processes that generated the value representations were subject to informational encapsulation such that the result was the generation of 'mandatory' outputs, then the perceived objectivity of moral beliefs could be explained. And if more general social interactions (as are assumed to be relevant in the operation of the Westermarck Mechanism), could generate model-free reinforcement learning (that is then

subject to informational encapsulation) the mandatory 'not-to-be-doneness' of incest could also be explained.

So, I am suggesting the alignments (drawing on the work of Cushman, Fodor, Stanovich and West) listed in Table 3.

Table 3	Alignments drawing on the work of Cushman, Fodor, Stanovich and West.
System 1	System 2
Intuition	Reason
Automatic	Controlled
Mandatory	Non-Mandatory
Informationally encapsulated	Cognitively Accessible
Action-based value representations	Outcome-based value representations
Model-free reinforcement learning	Model-based reinforcement learning

Table 3: Alignments drawing on the work of Cushman, Fodor, Stanovich and West.

Furthermore, the process I have outlined is, I suggest, similar to what Annette Karmiloff-Smith has called "modularization":

> Development and learning, then, seem to take two complementary directions. On the one hand, they involve the gradual process of proceduralization (that is, rendering behaviour more automatic and less accessible). On the other hand, they involve a process of "explicitation" and increasing accessibility (that is, representing explicitly information that is implicit in the procedural representations sustaining the structure of behaviour). (1992, p. 17).

I suggest that the process Karmiloff-Smith refers to as proceduralization could be the process that results in the output of a mandatory value representation from within a module that is informationally encapsulated. Given the mandatory nature of the value representation, and the related informational encapsulation, central cognition explicitly represents that action to itself as objectively 'morally right' (or objectively 'morally wrong'), in the process Karmiloff-Smith refers to as explicitation.

So, I suggest that modules, or to use Karmiloff-Smith's phrase "modularzation", is central to our perception of the objectivity of some of our moral beliefs.

Finally, returning to the empirical work of Goodwin and Darley (2012), they state that this work is primarily "to investigate sources of variance in the perceived objectivity of different moral beliefs." and they address three research questions. "Does the valence of moral beliefs predict their perceived objectivity?", "Does the perceived consensus pertaining to a moral belief predict its perceived objectivity?" and "Are objective moral beliefs associated with 'closed' responses to moral disagreement?" (2012, p. 250-1). To continue this line of investigation, I suggest a fourth research question. Does the type of learning algorithm used (which may be dependent on, or independent of, the content of the belief, or the social context of the learning process) predict the perceived objectivity of the moral beliefs learnt?

6. Meta-ethical implications?

No metaethical implications about the actual existence or non-existence of objective moral truths follow from the content of this paper. Furthermore, this paper presents no philosophical argument for the non-existence of objective moral truths. However, if the

position in this paper is accepted, then it is possible that humans perceive certain moral beliefs to be objective when they are not. But an argument (not presented in this paper), would need to be provided to reach the conclusion that no objective moral truths exist. However, arguments to the conclusion that something does not exist are notoriously difficult.

To make this point using an analogy employed in a different context, Bloom (2009) notes that humans have the ability to think in terms of mathematical objects (including, for example, numbers and sets). So, there should be a psychological explanation of how humans are able to think in terms of mathematics. However, this psychological capacity may be independent from the metaphysical status of mathematical objects themselves. Humans may be able to think in terms of mathematical objects while no mathematical objects actually exist (independently from the psychological representations of those objects), and thus, the existence of the psychological capacity does not settle the question of whether mathematical objects are real or not. By analogy, what I say here does not settle the question of whether or not certain moral beliefs that are perceived to be objective refer to objective moral truths.

7. Conclusion

One of the conclusions of Cushman's paper is that "action-based value representations allow us to explain strong, systematic patterns of non-utilitarian choice in moral judgment and behaviour." (2013, p. 285). I endorse this conclusion and further suggest that if those action-based value representations were mandatory and were delivered to central cognition from within an informationally encapsulated process, then that could lead central cognition to interpret those values as relating to objective moral beliefs. Thus, I suggest that informational encapsulation and the related mandatory value representations are central

aspects of the cognitive processes that ultimately generate non-consequentialist ethical judgments and the objective moral beliefs at the heart of those judgments.

Now I acknowledge that I am not an expert on the neuroscience that I have been presenting here. So, I may have overlooked some crucial aspects of that science. And I further acknowledge that I have not provided a definitive argument that my account is correct. But what I hope I have done is provide a plausible explanation for the perception that certain moral beliefs are objective.

Acknowledgments

I have presented versions of this paper at Duke University, Macquarie University, National University of Singapore, and University of Oxford. I thank those who offered their comments at these events. I have also discussed the ideas in this paper in more detail with a number of individuals. I particularly thank James Chase, Neil Levy, and Fiery Cushman for their help in clarifying the ideas presented in this paper. Finally, I thank the paper's reviewers (including Hanne M. Watkins who was willing to be identified) for helping me significantly improve the paper through the process of revision.

References

Bloom, P. (2009). Religious belief as an evolutionary accident. In J. Schloss, & M. Murray (eds), *The Believing Primate*. Oxford, OUP, Chapter 5.

Crockett, M. (2013). Models of morality. *Trends in Cognitive Science* Vol. 17 No. 8, pp. 363-366.

Cushman, F. (2013). Action, outcome and value: a dual-system framework for morality. *Personality and Social Psychology Review,* 17 (3), 2013, pp. 273-92.

Cushman, F. A., Gray, K., Gaffey, A., & Mendes, W. (2012). Simulating murder: The aversion to harmful action. Emotion, 12, 2-7.

Delgado, M. R., Jou, R. L., & Phelps, E. A. (2011). Neural systems underlying aversive conditioning in humans with primary and secondary reinforcers. *Frontiers in Neuroscience*, 5, 71.

Evans & Frankish. (2009). *In Two Minds: Dual Processes and Beyond*. Oxford: Oxford University Press.

Fodor, J. (1983). *The Modularity of Mind*. Cambridge MA: MIT Press.

Goodwin, G. P. & Darley, J. M. (2008). The psychology of meta-ethics: exploring objectivism. Cognition 106, pp. 1339-1366.

Goodwin, G. P. & Darley, J. M. (2010). The Perceived Objectivity of Ethical Beliefs: Psychological Findings and Implications for Public Policy. *Review of Philosophy and Psychology*. Vol. 1 (2). pp. 161-188.

Goodwin, G. P. & Darley, J. M. (2012). Why are some moral beliefs perceived to be more objective than others? *Journal of Experimental Social Psychology* 48 (2012) 250–256.

Greene, J. (2008). The secret joke of Kant's soul. In W. Sinnott-Armstrong (ed.) *Moral Psychology Volume 3*. Cambridge MA: MIT Press, pp. 35-80.

Greene, J. (2013). *Moral Tribes: Emotion, Reason and the Gap Between Us and Them*. Penguin.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review* 108: 814-34.

Hood, B. M., Donnelly, K., Leonards, U., & Bloom, P. (2010). Implicit voodoo: Electrodermal activity reveals a susceptibility to sympathetic magic. Journal of Cognition and Culture, 10, 391-399.

Hume, D. (1902 [1777]). *Enquiries Concerning the Human Understanding and Concerning the Principles of Morals*. In L. A. Selby-Bigge, M.A. 2nd edition. Oxford: Clarendon Press. Retrieved 5/19/2017 from the World Wide Web: http://oll.libertyfund.org/titles/341

Kant, I. (2002 [1785]). *Groundwork for the Metaphysics of Morals*, translated by Allen Wood, Yale University Press.

Kahneman, D. (2002). *Maps of Bounded Rationality: A Perspective on Intuitive Judgment and Choice*. Nobel Prize Lecture, http://nobelprize.org/nobel_prizes/economics/laureates/2002/kahnemann-lecture.pdf Accessed March 28, 2011.

Karmiloff-Smith, A. (1992). *Beyond Modularity*, Cambridge, MA: MIT Press.

King, L. A., Burton, C. M., Hicks, J. A., & Drigotas, S. M. (2007). Ghosts, UFOs, and magic: Positive affect and the experiential system. Journal of Personality and Social Psychology, 92(5), 905-919.

Mackie, J. L. (1977). *Ethics: inventing right and wrong*. London: Penguin.

Plato. (2008). *The Republic*. Project Gutenburg Ebook. http://www.gutenberg.org/files/1497/1497-h/1497-h.htm Accessed 22 December 2017.

Rozin, P., Millman, L., & Nemeroff, C. (1986). Operation of the laws of sympathetic magic in disgust and other domains. Journal of Personality and Social Psychology, 50, 703-712.

Segall, Marshall, Campbell & Herskovits. (1966). *The Influence of Culture on Visual Perception.* Oxford: Bobbs-Merrill.

Sripada, C. (2008). Nativism and moral psychology: three models of the innate structure that shapes the content of moral norms. In W. Sinnott-Armstrong (ed.) *Moral Psychology Volume 1: The Evolution of Morality: Adaptations and Innateness*. Cambridge MA: MIT Press, pp.319-344.

Stanovich, K. & West, R. (2000). Individual differences in reasoning: implications for the rationality debate. *Behavioural and Brain Sciences* 23, 645–665.