Neuroethics (2011) 4:163–174
DOI 10.1007/s12152-010-9087-z

REVIEW PAPER

Cognitive Neuroscience and Moral Decision-making: Guide or Set Aside?

Derek Leben

Received: 9 July 2010 / Accepted: 16 July 2010 / Published online: 4 August 2010

Abstract It is by now a well-supported hypothesis in cognitive neuroscience that there exists a functional network for the moral appraisal of situations. However, there is a surprising disagreement amongst researchers about the significance of this network for moral actions, decisions, and behavior. Some researchers suggest that we should "uncover those ethics [that are "built into our brains"], identify them, and live more fully by them," while others claim that we should often do the opposite, viewing the cognitive neuroscience of morality more like a science of pathology. To analyze and evaluate the disagreement, this paper will investigate some of its possible sources. These may include theoretical confusions about levels of explanation in cognitive science, or different senses of 'morality' that researchers are looking to explain. Other causes of the debate may come from empirical assumptions about how possible or preferable it is to separate intuitive moral appraisal from moral decisions. Although we will tentatively favor the 'Set Aside' approach, the questions outlined here are open areas of ongoing research, and this paper will be confined to outlining the position space of the debate rather than definitively resolving it.

Keywords Moral psychology · Reasoning · Decision-making · Folk theory

The Problem

The last 10 years have seen an explosion of research in the emerging cognitive neuroscience of morality, revealing what appears to be a functional network for the moral appraisal of situations. Initial evidence for this is similar to evidence for the biological basis of language: people learn to produce complex moral appraisals of novel situations, that show non-random uniformity across societies, despite the lack of explicit teaching in the form of positive or negative evidence [1–3]. Furthermore, researchers have identified a network of systems that are consistently involved in these moral appraisals, including systems for affective appraisal, planning and regulation, and intention-reading and social skills [4–6]. Although there is no 'neural correlate of morality' in the same way that there is a neural correlate of arm movement in the motor cortex of the brain,[1] it is by now a well-supported hypothesis that there exists a procedural or functional network for moral appraisal (FNMA). This network recruits several interconnected systems,

D. Leben (✉)
Johns Hopkins University,
Baltimore, MD, USA
e-mail: dleben1@jhu.edu

[1] As Greene and Haidt [4] state: "… if one attempts to deconfound moral judgment with everything that is not specific to moral judgment (emotion, theory of mind, abstract reasoning, and so on) there will almost certainly be nothing left."

much like the processes of memory and attention, where many cognitive-neural systems are involved in what is pretheoretically called 'memory,' including working memory, episodic memory, declarative memory, motor memory, etc. The idea here is that these memory systems all work together as a network to achieve the function of retrieving information from past experiences.[2] Similarly, the systems in the FNMA all work together to achieve the function of making moral appraisals.

The problem is this: despite much agreement amongst researchers on the results described above, there is a significant disagreement about the *significance* of these results for moral actions, decisions, and behavior. Some of the leading researchers in the field assume that people should (or cannot help but) guide their decisions according to cognitive-neural systems for moral appraisal. Michael Gazzaniga [7] will be chosen as a representative:

> I would like to support the idea that there could be a universal set of biological responses to moral dilemmas, a sort of ethics, built into our brains. My hope is that we soon may be able to uncover those ethics, identify them, and to begin to live more fully by them. I believe we live by them largely unconsciously now, but that a lot of suffering, war, and conflict could be eliminated if we could agree to live by them more consciously [7: preface].

Marc Hauser [1] also claims that the cognitive systems underlying moral judgments should receive a great amount of attention in private and public deliberations, although he is sometimes vague about what kind of attention. At one point, Hauser states that "policy wonks and politicians should listen more closely to our intuitions and write policy that effectively takes into account the moral voice of our species." This is a weaker position than the view supported by Gazzaniga's quote, but still suggests that our moral psychology should play a guiding role in decision-making. For simplicity, we will call this the 'Guide' view.

Taking a completely different approach, researchers like Joshua Greene and Jonathan Haidt suggest that the cognitive-neural systems underlying moral appraisals should be studied so that these can be *set aside* in matters of decision-making [8, 9]. According to Greene, these neural systems evolved in environments we no longer inhabit and they should only constrain decisions in so far as they are useful, which often they are not:

> …[Our moral psychology] may work well enough for life in small, relatively isolated hunter-gatherer bands, but it's absolutely disastrous for billions of people raised in a variety of different cultures and subcultures who must share a world in spite of their incompatible worldviews. Human moral psychology doesn't scale well [8: 232–233].[3]

In contrast to Gazzaniga's suggestions, Greene argues that we should study the neural systems for morality in the same way that pathologists study the biological causes of diseases, so that the effects can be avoided and treated:

> I propose instead that we use our understanding of moral psychology to transcend our ordinary modes of moral discourse rather than to operate more effectively within them [8: 236].

We will call this view about the role of cognitive neuroscience the 'Set Aside' view.

Like many foundational disagreements amongst scientific researchers, this problem is largely implicit in general discussions, and often surfaces only occasionally in remarks found in prefaces and conclusions. However, I believe this is not simply a matter of loose speaking on the part of a few authors; rather, the problem is widespread in many books and articles, often appearing as a range of intermediary positions.[4] For example, in a recent paper on the role

[2] This function is multiply realizable, and could be accomplished differently by, say, the way that a serial computer retrieves information from past experiences. More on levels of explanation will be discussed in the next section.

[3] While not important at this point in the discussion, Greene's quote here refers to a mismatch between the *ultimate* (evolutionary) causes of moral appraisals and actual moral decisions, rather than what we will be focusing on mostly in this paper, which is a mismatch between the *proximate* (cognitive-neural) causes of moral appraisals and moral decisions.

[4] One might object that these intermediate positions are different than the Guide or Set Aside extremes. As we are characterizing these positions, any view that the functional network for moral appraisals should have priority amongst other considerations, or cannot be trumped on other grounds, is a Guide view. Although other positions might differ about how and when to set aside (sometimes? always?), they will here be grouped together.

of affect in attributions of moral responsibility, Roskies and Nichols [10] admit that the bias they discuss might be misleading and harmful, but assume that it cannot be set aside. Why not? In many of these cases there are no explicit reasons for why neural systems should or should not have priority in guiding decision making. Although we will tentatively favor the Set Aside position, the primary purpose of this paper is to identify the possible sources of this disagreement, and the ways in which it might be resolved.

There are many issues in meta-ethics that this problem involves,[5] but the real parallel to our problem lies in the epistemology of ethical claims:

> Can ethical judgments be justified on the basis of intuitions alone (moral intuitionism), or do they require evidential and inferential support?

Most contemporary proponents of intuitionism have in mind a priori intuitions of the kind seen in mathematics [11]. Yet the Guide/Set Aside problem (as we are formulating it) concerns intuitive knowledge that is not established or justified by a priori methods. As such, it is more similar to "moral sense theories" developed in 18th century Scottish philosophy [12, 13]. We might consider our problem a very specific version of this epistemological issue, set in the domain of cognitive neuroscience.

A final preliminary issue is an important question where the researchers do *not* differ: the "is/ought" distinction, sometimes also referred to as the "fact/value dichotomy" or the "naturalistic fallacy." The is/ought distinction argues that it is a fallacy to make normative claims on the basis of (natural) facts. This initially appears plausible; the fact that everyone steals office supplies from work does not entail any normative claims about whether stealing office supplies should or should not be permissible. However, ethical philosophers have long acknowledged that natural facts, at the very least, somehow constrain the space of possible normative claims: to make a normative claim, it must be humanly possible to achieve it ("ought implies can"). All of the authors in our discussion take an even stronger position, that natural facts can actually *inform* rather than just constrain our ethical theorizing.

Flanagan et al. [14] claim that anyone committed to naturalism must be likewise opposed to the is/ought distinction. They point out that David Hume's supposed opposition to this distinction was not that "ought" claims can *never* be drawn from natural facts (the authors note that Hume goes on to do just that himself). Rather, the is/ought distinction is that normative claims cannot be established *deductively* merely from natural facts alone. Flanagan et al. argue that normative claims are reasoned to, like most other claims outside of mathematics, by methods of induction and abduction:

> City-planners, architects, and bridge-builders are naturalists. They engage in discussions within communities of fellow planners (with differential power); the communities then generate ends (from among sets of good ideas) and create structures that achieve these collectively generated ends. So too with morality [14: 22].

If Flanagan et al. are correct, then making normative ethical claims is not any different than decision-making in other domains (there will be discussion later about what might make it specifically 'moral'). For these reasons, we will be referring to normative ethics as "moral decision-making."

It is important that the authors we are considering in the Guide/Set Aside debate agree on this issue. Hauser [1: 4] takes an example of bringing a child to the dentist and being told that she needs a dental procedure, where anesthesia will relieve her pain during this procedure. Hauser basically agrees with Flanagan et al. that we cannot reason deductively from these facts to a normative decision that the child ought to be given anesthesia, but he indicates that these conclusions can be reasoned to inductively or abductively: "Here [the example] it seems reasonable for us to move from fact to value judgment." Greene [15] also agrees that scientific work can have real effects on our normative ethical theorizing.

The is/ought distinction is then clearly not the source of our problem: Greene does not think that the FNMA should be set aside in moral decision-making because it is a fallacy to reason from one to the other. The question is: *how* should we be reasoning from one to the other? Should cognitive neuroscience guide, or be set aside? To preview the rest of the paper, we will discuss two

[5] It has been pointed out to me that many people engaged in the 'emotivism/rationalism' debate (are moral appraisals largely the product of emotion or rational deliberation?) consider it to have immediate consequences for the Guide/Set Aside debate (those who see affective components as primary are more likely to have a Set Aside view). However, our problem is logically consistent with either of these positions: one can hold that we should guide our behavior according to affective appraisals, or set aside appraisals produced by systems for rational inference.

theoretical and two empirical questions that can help resolve this problem. On the theoretical side, it is possible that either our cognitive neuroscience explains human behavior at the level of *decisions*, or that there are no moral properties outside of those explained by cognitive neuroscience. If either of these were the case, then it would be theoretically impossible to have moral decisions that set aside the FNMA. However, it will be argued that neither of these claims hold, and a helpful illustration of this is an analogy to 'folk physics.' On the empirical side, the question then arises as to whether it is, as a matter of fact, possible to make moral decisions without privileging the FNMA. And even if this is possible, this network still may simply do a more effective job of guiding moral decision-making than any other possible method.

Levels of Explanation in Cognitive Systems

One possible origin of the Guide/Set Aside disagreement is a vagueness about what level of explanation our cognitive neuroscience of morality aims at. Almost all of the researchers in moral psychology claim that the phenomena they are looking to explain are "moral judgments," but this does not help us, because "judgment" is often ambiguous and can be used to describe multiple levels of information. For instance, Hauser, Young, and Cushman [16] use 'judgment' to describe at least three different parts of their model:

1. The "unconscious operative principles" that "underlie certain aspects of. . . morality" [16: 109].
2. The outputs or functions of these operative principles, e.g.: "we must describe computations underlying the judgments that we produce" [16: 116], ". . . students of moral behavior might use morality judgments to uncover some of the principles underlying our judgments of what is right and wrong" [16: 118].
3. The entire event, including processing, output, and decision-making, e.g.: "how does the child acquire a particular moral grammar, especially if her experiences are impoverished relative to the moral judgments she makes?" [16: 107]

To some extent, this problem is widespread within psychology and cognitive neuroscience. Psychologists who work on reasoning and inference often use these terms to describe the unconscious processing, the output or use of these systems, and the entire event as well. One important contributing factor to this problem is that the authors of this paper are all proponents of the "linguistic analogy" that takes generative linguistics as a good model for the methodology of moral psychology [2]. It is often assumed by generative linguists that the output of the cognitive system *just is* the function that system performs (producing and comprehending well-formed sentences); therefore the linguistic analogy might suggest that the output of moral appraisals is identical with moral decisions. It is even possible to hold a position that there is actually just one level of explanation in cognitive capacities, be it brain processes [17] or behaviors [18].

Although these positions were popular within philosophy of mind a half-century ago, cognitive scientists have since largely advocated one or another variety of 'functionalism,' which hypothesizes multiple levels of explanation that can be realized in multiple ways. Perhaps the most popular variety amongst cognitive scientists is a framework suggested by David Marr [19] which includes three levels of functional operation:

(1) A neural level of 'hardware,'
(2) An algorithmic/procedural level of 'software,' and
(3) An environmental level of 'function,' or what functions the system actually performs in the local environment.

Proponents of the linguistic analogy should embrace this framework, as Chomsky often acknowledges a similar distinction and comparison to Marr's methodology, especially in his discussions of semantics and phonology. Take the following discussion from Chomsky [20: 17] regarding explanation in the human visual system:

> Study of the visual and motor systems has uncovered mechanisms by which the brain interprets scattered stimuli as a cube and the arm reaches for a book on the table. But these branches of science do not raise the question of how people decide to look at a book on the table or pick it up, and speculations about the use of the visual or motor systems, or others, amount to very little.

Although Hauser and colleagues use "judgment" ambiguously to describe the Marrian levels (2), (3), and (1–3), by their appeal to the standard methods of cognitive

psychology as well as the linguistic analogy, it would appear that they are aiming to provide an explanation of systems at the first two levels. We will call this "moral appraisal." At a theoretical level, then, explanations of moral appraisal have no direct consequences for moral decision-making, just as explanations about the function of the visual system or the functional network for language processing have no direct consequences for their use and applications.[6]

Distinguishing Senses of 'Moral'

A more likely source of disagreement in the Guide/Set Aside debate might be found in exactly what the researchers mean by 'moral appraisal.' If there are multiple senses or properties picked out by moral vocabulary (words like 'right,' 'ought,' 'good,' 'should'), then cognitive neuroscience may apply to some of these properties but not others. This has important consequences for whether it is theoretically possible to have *moral decisions* without the FNMA, and how the FNMA could be evaluated.

To begin with, Greene [8] does claim that the word 'moral' and related vocabulary are ambiguous (or polysemous) in English, and pick out at least two distinct senses, identified by him as: 'morality$_1$' and 'morality$_2$.' Morality$_1$ is defined as concerning "facts about what is right and wrong;" this is what ethicists have in mind when they discuss first-order or *inherent* moral properties such as duties, rights, principles, and obligations. Morality$_2$, in contrast, is defined as: "of or relating to serving…the interests of others," and is what ethicists often describe as second-order or *instrumental* moral properties, where a claim like "X is right" is always taken to be relative to being right for somebody in a specific time, at a specific place, for a specific purpose, and so on.[7]

In contemporary ethics, sometimes ethicists take morality$_1$ to be their subject of inquiry, and other times they take it to be morality$_2$. Although deontological approaches are almost always concerned with moral$_1$ properties [21], sometimes consequentialists are as well [22]. Meta-ethicists are most often concerned with moral$_1$ properties, as these appear to be 'metaphysically strange' entities. In fact, many naturalistic philosophers like Mackie [23] and Joyce [24] have argued that *because* these properties are 'weird,' i.e., inconsistent with the kinds of properties thus far revealed by the natural sciences, we should conclude that they do not exist. While the weird entities argument is weak in itself, Greene [8, 25] argues that cognitive neuroscience actually provides another kind of argument: that moral$_1$ properties track the outputs of the FNMA, and it is therefore a reasonable conclusion that moral$_1$ properties *just are* the products of this functional cognitive network. This much would likely be met with agreement by other cognitive neuroscientists, as well as many naturalistic meta-ethicists. For the purposes of this paper, let us assume that the tracking argument is valid and successful (see Dean [26] and Timmons [27] for replies).

Moral$_2$ properties, as we are defining them here, are something that can be mind-independent; that is, decisions in the best interest of others are things that can be theorized about independently of our intuitions, like the properties studied by physics or engineering. Thus, it is theoretically possible to make moral$_2$ decisions (decisions made in the interests of others) by appealing to psychological appraisals, but also other non-psychological methods such as counter-intuitive empirical studies, domain-general theories of action that benefits others, and formal training. Selim Berker's recent paper "On the normative insignificance of neuroscience" [28] nicely illustrates the theoretical importance of postulating this independent morality$_2$ for the Set Aside position. He writes: "usually when we deem something to be a heuristic [or a folk theory, both described below], we have a good handle on what the right and wrong answers in the relevant domain are…" [28: 20].

A helpful illustration for the proposed distinction between moral$_1$ and moral$_2$ properties is an analogy to 'folk physics,' which contrasts with a scientific theory of physics [29, 30]. Folk physics is a set of cognitive heuristics which developed over evolutionary time to

[6] While vagueness about levels of explanation is one possible cause of the Guide/Set Aside disagreement, it is unlikely that this is the only (or primary) source of contention because these distinctions are so common throughout the cognitive sciences. Therefore, we will move swiftly to the next issue without considering objections.

[7] In personal communication, Greene has expressed some hesitance about this interpretation of his moral$_1$/moral$_2$ distinction. Therefore, the formulation here may be slightly deviant (although not necessarily inconsistent) with Greene's original or current intentions.

navigate around the world, and may be correct in isolated situations, but their primary function is to be useful rather than correct (see Gigerenzer [31] for a review of cognitive heuristics). Scientific physics, in contrast, is a collaborative practice based in theory, observation, and experimentation to reveal correct information about the physical world. These two might be related in many respects: the reason that folk physics is so successful in isolated environments might just be that correct information is often the most useful. However, it is important that the two bear no necessary connection. In the same way, the 'folk morality' that would be the product of our FNMA (also proposed by Premack and Premack [32]) can often overlap with behavior in the interests of others, but there would be no necessary connection between the two.

Let us consider an immediate objection to the proposed distinction between morality$_1$ and morality$_2$: challenging the possibility of an intuition-independent theory of behavior in the interests of others (moral$_2$ properties), or claiming that "the interests of others" is somehow a weird property for a naturalistic framework. Berker [26] questions the postulation of moral$_2$ properties in the following way:

> …in the moral case [unlike in Kahneman and Tversky's cases of logic and reasoning] it is very much up for debate what the right and wrong answers are…How can we proclaim these emotional[8] processes to be quick but sloppy shortcuts for getting at the moral truth unless we already have a handle on what the moral truth is?

In response to Berker's concerns, it should first be pointed out that there is a large difference between questions being "very much up for debate" and there being *no fact of the matter* about the answers. There are many questions within the sciences that are very much up for debate, and yet presumably have a fact of the matter, so there needs to be additional reasons for supposing that questions about the best interests of others are any different (perhaps a pessimistic meta-induction?). A response to Berker's substantive concerns about an objective morality$_2$ is to allude to similar kinds of value posited by other special sciences like economics, political theory, and sociology. If it is possible to develop objective mind-independent theories of social structure, political structure, and economic behavior, then there does not appear to be any principled reason against a mind-independent theory of people's interests taken generally. The 'interests of others' as we are defining them in morality$_2$ are no different than the amalgamation of interests posited by the natural and special sciences. Just as there appears to be an optimal condition that satisfies everyone's economic interests in a given situation (even if that condition is an idealization), there might be an optimal condition for interests taken as a whole. One method for deriving this might look like a kind of cost-benefit analysis, where as many interests and people are factored in as possible.[9]

Another way to formulate this objection is to question the kind of convergence and progress that might be possible in the study of shared interests as opposed to physics and other sciences (also discussed in Flanagan [34]). Mikhail often emphasizes that moral theorists should properly focus on morality$_1$ (the FNMA), comparing morality$_2$ to 'performance' in generative grammar [3], which Chomsky often regards as being theoretically intractable. There are some good reasons to be skeptical about the tractability or commensurability of moral$_2$ properties. Take Sartre's [35] famous example of a young man trying to decide between taking care of his sick mother and fighting for his country. In this case, Sartre plausibly suggests (in different terminology) that there is no possibility of ever successfully comparing the familial interests of taking care of one's sick parent with the social interests of protecting one's community from a real destructive threat. If it is indeed the case that moral$_2$ properties are incommensurable or theoretically tractable, the Set Aside view might become incoherent as Berker suggests (set aside in preference to what?). This is certainly an area where the Set Aside view needs to provide some sort of plausible story about the commensurability of moral$_2$ properties. However, in contrast to Berker and Mikhail's concerns, I do not believe that

[8] As we are describing the FNMA, it consists not only of emotional processes but also typically 'cognitive' ones.

[9] As much as this may seem like consequentialism or Utilitarianism, there are no a priori reasons for why the interests of others must reduce to one kind of utility (e.g., pleasure or pain), or why entities like intentions or character cannot be a part of people's general interests. A non-a priori argument against character in moral$_2$ explanations would look like Doris [33].

such a fully fleshed-out theory of morality$_2$ is required merely to establish the proposed distinction that we are discussing here. It should suffice that in some cases we are able to estimate as best as possible what would be in the best economic, biological, political, etc. interests of a person, and contrast this with the appraisals produced by the FNMA.

Are Moral Decisions Possible Without the FNMA?

Once these conceptual distinctions have been made, progress has been forged towards understanding the source of our disagreement. There may be vagueness about different functional levels of explanation, or between different senses of morality, leading researchers to conclude that appraisals just are decisions, or moral$_1$ properties just are morality. However, we must also consider that those advocating cognitive neuroscience as a guide to moral decisions differ on other more empirical grounds. Assuming it is theoretically possible not to guide moral$_2$ decisions according to appraisals produced by the FNMA, one might object that it is, as a matter of fact, impossible to separate the two. Another argument would be that this is possible, but somehow not empirically preferable. We will deal with each of these claims separately.

The first issue concerns whether it is empirically possible to make decisions in the interests of others without appealing to the FNMA. At first this might appear similar to arguments in the late nineteenth century that evolutionary biology or the new atheism would, as a matter of fact, undermine behavior in the interests of others, which turned out to be empirically false [36]. However, the FNMA is a much more plausible necessary condition for moral behavior, largely based on studies of patients whose cognitive-neural deficits are directly correlated with deficits in their decisions and behavior in the interests of others. The cases of sociopaths, psychopaths, autistic persons, and those with frontal lobe damage are all examples. The Guide-theorist might use these cases to argue that actions in the interest of others are impossible without the functional network for moral appraisals.

Establishing the FNMA as a necessary condition for moral$_2$ behavior does not, however, lead to the conclusion that it is a sufficient condition. This is because, according to the claims of the previous section, there can be other (non-intuitive) contributions to moral$_2$ behavior, and increasing these contributions might 'override' any necessary psychological contributions. Such cognitive overriding can be accomplished by at least two methods: (1) the use of domain-general processing to override modularized and automatic processing, and (2) 'cognitive offloading' techniques, which according to Wilson [37: 628] ". . . make use of the environment itself in strategic ways leaving information out there in the world to be accessed as needed, rather than taking time to fully encode it." This is also known as "distributed cognition," because it involves the use of external tools such as counting, drawing maps, and using reference pointers to overcome cognitive limitations. Both of these are real examples of methods used to override cognitive biases or limitations.

Let us take an example where the function of a cognitive-neural system is corrected based on non-psychological considerations: social stereotypes. Many psychologists argue that normal systems for concept formation produce an effect of reasoning by prototypes, resulting in biases against individuals based on gender or race [38].[10] These psychological systems are obviously not anything we can simply get rid of, because it is extremely important that people be able to categorize things that look similar as belonging to the same type. However, most people attempt to override these biases when evaluating individuals by using certain explicit procedures or criteria regarding what features to direct attention to, which procedures to follow, and how to properly ignore intuitive biases. As in the analogy to folk physics, recommending that we practice scientific physics for progress and information about the physical world does not imply that we eliminate or disregard the psychological systems for spatial orientation and navigation that underlie folk physics. Doing so would probably produce people who couldn't even get out of bed in the morning, much less operate large particle accelerators. Instead, we simply increase the non-psychological contributions to physics to such an extent that the psychological contributions become negligible, and are effectively

[10] This is just one model amongst many in the psychology of race; others explain racial categorizations as the result of perceptual feature groupings [39], or by systems for tracking the boundaries of social groups [40], or even most directly, by a system with a direct function of categorizing ethnicities [41].

drowned out. The Set Aside position suggests that the same might be done in the domain of ethics.

One objection here is that the use of domain-general suppression and distributed cognition can alter *explicit* effects on moral decisions, but not *implicit* ones. Returning to our analogy to racial bias, many studies illustrate that even in individuals who explicitly attempt to override these biases, there is still a large implicit effect evident in measures of skin response, eye movements, and so on [42]. The guide-theorist might respond that because a large amount of our behavior is caused by unconscious psychological mechanisms, these implicit measures demonstrate that it is still impossible to make moral decisions in absence of (or suppression of) moral appraisals.

In response, I will follow a line of argument pursued by Kelley et al. [43], who discuss a version of this Guide/Set Aside debate in racial theory (termed 'conservationism' and 'eliminativism'). They discuss how implicit racial biases pose a prima facie argument for the impossibility of eliminating racial categorization. However, many studies (reviewed in a 2001 special issue of the Journal of Personality and Social Psychology) [44] propose that implicit biases can be not only suppressed but actually altered and modified by employing certain environmental conditions and training. Thus, formal and informal training might be sufficient after all to enable moral decisions not guided by moral appraisals. While this is still at an early stage, some researchers are already investigating whether a similar sort of training can influence appraisals of moral dilemmas and scenarios [45].

Is FNMA the Most Effective Method?

The previous section argued that cases of people with cognitive-neural deficits do not successfully establish the impossibility of overriding or altering the FNMA. This opens the second aspect to our empirical discussion: is it *preferable* to override the FNMA in moral decision-making? It certainly could be the case that the FNMA is more effective than any other method we might adopt for making moral decisions. There are several metrics that can be used to gauge efficiency; we will consider two: accuracy (decisions that are objectively productive some significant amount of the time), and resourcefulness (decision-making given limited time and resources).

The FNMA is More Accurate

The first metric requires examining cases of appraisals produced by the FNMA and comparing them to some objective measure of benefit or harm for others. It may simply turn out, upon investigation, that the FNMA is optimal for making decisions in the interests of others. This is not wildly implausible, given the optimal functioning of many other cognitive systems in relation to their environments: a single photon can be detected by the human visual system, and a vibration of the eardrum even the distance of half the diameter of a hydrogen atom can be detected by the auditory system [46]. If the FNMA is the product of natural selection, it might be the case that it has been selected to perform better at $moral_2$ decision-making than a domain-general or cognitive offloading methodology. On the other hand, we have also been discussing cases of cognitive heuristics like the network for spatial cognition underlying folk physical appraisals, which function extremely well in bounded circumstances, but operate poorly when extended beyond these conditions. Therefore, it is simply an open question how effective the FNMA is in decision-making compared to other non-psychological methods.

To answer this open question, the Set Aside proponent points to situations in which the systems involved in the FNMA can often skew identical moral decisions, as in thought-experiments where it is obligatory to help a wounded pedestrian you pass for a high price, yet permissible not to help wounded people in other countries for the same price. Walter Sinnott-Armstrong [47] compares these to the framing effects found in the heuristics and biases literature, where irrelevant information can skew objectively identical appraisals, thus undermining the reliability and justification of a mechanism. As Sinnott-Armstrong states: "If moral intuitions are subject to framing effects, then they are not reliable in those circumstances." Indeed, if we are simply evaluating the FNMA as a mechanism, then the fact that this mechanism sometimes responds to irrelevant information is good grounds for doubting its general reliability.

Another approach, adopted by Sunstein [48] and nicely described by Jonathan Haidt [9: 815] is demonstrating how reliance on the FNMA will "often bring about nonoptimal or even disastrous consequences in matters of public policy, public health, and the tort system." Just like scientific physics predicts

that a thrown object will move differently than the systems underlying folk physics suggest, social psychology, economics, and political science tell us that many morally counterintuitive decisions sometimes (though not always) do a better job of benefiting the interests of others than our intuitive moral judgments.

Examples in the domain of public health include abortion and euthanasia situations, where ending life is often in the general interests of all parties involved, and yet the appraisal of an up-close and personal intentional harm still pulls the FNMA against the decision.[11] In public policy, the repulsion of giving assistance to distant people may underlie many reactions to welfare, government-sponsored healthcare, and charity programs [49]. Affect-laden responses produce disgust, anger, and frustration towards helping those who we do not feel morally pulled towards, even if it were to benefit everyone's interests on the whole. In the legal system, it might sometimes be preferable to make decisions without privileging the cognitive systems underlying the FNMA, specifically the systems for the attribution of intentions. Decades of research in social psychology has demonstrated that situational factors have much more of an influence on behavior than intuition suggests (see Doris [33] for a review). People show a shocking tendency to conform to group behavior, obey authority figures, and allow apparently trivial information to influence important actions.[12] Intentions and goals of people in these situations have little role in differences in behavior. Yet the neural systems for evaluating moral actions pull towards attributing intentions to actions that are charged with affect [51]. If paying attention to the findings of social psychology forces us to revise appraisals of intentions, this could have a large impact on attributions of fault and punishment, as in current approaches to addiction.

The Guide-theorist may reply that these cases are cherry-picking, and if we were to take a truly random sample size of moral decisions (not necessarily moral dilemmas), the FNMA would actually be very successful. This kind of argument has been applied in social psychology by Kruger and Funder [52], who claim that supposedly 'irrational' heuristics are actually incredibly effective most of the time at promoting social cohesion within groups. In moral theory, Mikhail [53] (responding to Sunstein) has made a similar claim that these cases of 'misfires' only focus on a small class of performance errors. The vast majority of most people's $moral_2$ decisions are quite mundane, including throwing away your paper towel properly in a public restroom (when nobody is around), tipping at a restaurant you'll never return to, and holding the door for a stranger. When all of these decisions are factored in, it might be the case that the FNMA does a much more accurate job than the above examples suggest.

Yet even if this is the case, it does not establish that the FNMA is *more* accurate than other possible decision procedures for moral scenarios, which is the question at hand. This applies equally to both the framing effect and the 'misfire' arguments. Defenders of the accuracy of the FNMA tend to overlook the fact that we are not simply asking 'is the FNMA accurate?' but rather, comparing it to other possible methods for making moral decisions. Take Sinnott-Armstrong's example of making judgments about height based on intuitive perceptions. His framing argument states that, because one person might look taller while standing next to a Sequoia Tree than a Bonsai Tree, our perceptual systems should not guide decisions about height. Now, if a defender of our perceptual system steps in and replies that actually these systems are quite good at making height judgments in bounded circumstances, this is ignoring the fact that these systems are being *compared* to other possible methods, such as using a tape measure. Indeed, it is because of this inaccuracy that many fine liquor establishments currently display a convenient height marking on their front door for use in making height judgments. The Set Aside proponent is making this same proposal for moral decisions.

Lacking any further resources at the moment besides examples like the ones above, the burden of proof is placed on the Guide-theorist to show that these are the exceptions rather than the rules. It would be very exciting to see further studies comparing reliance on the FNMA with use of non-intuitive tools, training, and rules in a wide variety of scenarios.

[11] Of course, there exist many sophisticated objections to welfare, abortion, and euthanasia that do not rely solely on affective or intuitive responses.

[12] For instance, if a person finds a small amount of change in a phone booth, that person is significantly more likely to help a stranger than not [50].

The FNMA is Most Effective Given Limited Time/ Resources

We must finally consider the possibility of the Guide-theorist acknowledging that a significant amount of the time the FNMA will lead to decisions that are harmful or counterproductive, but insisting that, given the constraints of limited time or resources in decision-making, the FNMA is simply the best available method. This defense is proposed by Gigerenzer [54], who claims: "when it comes to issues of justice and morals, there are situations in which the use of heuristics, as opposed to an exhaustive analysis of possible actions and consequences, is preferable" [54: 20]. These situations are ones in which "simple heuristics, which ignore part of the available information, are not only faster and cheaper [in a biological and cognitive sense] but also more accurate for environments that can be specified precisely" [54: 19]. Ignoring the accuracy metric that is discussed above, Gigerenzer argues that in many (most?) situations, using explicit criteria, domain-general learning, and formal rules for novel moral decisions would be less effective than employing the FNMA.

Gigerenzer's evidence is largely in non-moral domains such as language learning and estimation judgments (say, about cities with the largest populations). One example discussed by Gigerenzer [55] (originally from Green and Mehr [56]) examined the use of a regression tool for diagnosing coronary symptoms in a rural Michigan hospital compared to the use of intuitive heuristics. The surprising result was that a simple metric based on intuitive cognitive heuristics like "ordered search," a "stopping rule," and "take the best"[13] actually performed better than the use of the regression tool on accuracy, time, and resource use. The fact that the FNMA takes little time or resources to employ might be a serious reply for the Guide-theorist. Many situations that require making decisions in the interests of others have time constraints that do not allow a careful weighing of values—especially given the novelty of most moral scenarios. Hauser and his colleagues often emphasize that the FNMA is valuable because of its ability to handle novel moral scenarios, much in the way that the faculty of language is capable of interpreting novel linguistic constructions. The considerations of limited time and novelty might imply that it is an impractical strategy to weigh probabilities, costs, and benefits in making moral decisions. The Guide-theorist instead suggests that, so long as the FNMA succeeds in the above cases more than random chance, it will be preferable in these specific circumstances to guide decisions according to it. This is a way of arguing that, accuracy aside, the FNMA might still be preferable to any other available method for making moral decisions.

However, scenarios involving constraints on time and resources are not powerful enough to support a Guide view. Time-sensitive decision-making does not limit options to only the FNMA or random chance; this is one of the primary motivations for applied ethics. Consider medical ethics: doctors often find themselves in time-sensitive situations where they must make decisions in the interests of others without being able to take enough time to weigh the costs and benefits to everyone involved. Rather than relying only on their intuitive appraisals or random chance, the doctor instead has at his or her disposal a detailed training in bioethics with the costs and benefits of similar situations. Explicit procedures have been automatized and internalized so that they now become quick and easy. Thus, although the doctor is unable to make a cost-benefit procedure at the moment, the fact that they have done so for similar situations enables them to quickly make a decision based on intuitive considerations rather than the FNMA or random chance.

Again, this is an empirical question open to further research, and it would be very interesting to see studies comparing the quick intuitive judgments of doctors or lawyers who have taken ethics classes as compared to those who have not.[14] To employ the linguistic analogy, these would be like artificial languages and codes (i.e., computing languages) which have certain benefits that natural languages do not have, such as a lack of ambiguity. While ambiguity might be beneficial for natural-language use (see [57: ch.6]), in other domains it can lead to unfortunate consequences.

The solution to limited time and resources is moral education rather than moral intuition-building. The first few pages of Hauser [1] list some varieties of formal education in moral decision-making, such as medical, business, and legal ethics courses, military training, and

[13] See Gigerenzer [55] for details

[14] In the last few years there have been a number of studies examining the appraisals of ethicists and people who have taken ethics courses compared to control groups.

bureaucratic committees, with the claim that these are not the causes of moral "judgments." Hauser is correct that these institutions are not the cause of moral appraisals, but they *are* very often the causes of moral decisions, and if it turns out that these methods are more accurate and can be just as automatic as the FNMA, they *should* be.

Conclusions

In conclusion, a foundational disagreement about the role of cognitive neuroscience in moral decision-making has revealed both theoretical and empirical assumptions. After investigating these assumptions, this paper has suggested that it would be mistaken to look towards the psychological network for moral appraisal as a guide in private or public decision-making. This is an interesting a surprising conclusion, but cognitive neuroscience has been providing similar results about other cognitive domains for years. Eyewitness testimony was once (and still is in many places) the ultimate grounds for decisions in the legal system ("who are you going to believe, me or your own eyes?"), but psychological research has revealed that our attention is selective in ways that fail to see obvious things right in front of us [58], and our memories are not like pictures, but often confabulate false scenarios and draw upon current environments or information [59]. The case of morality is no different; if investigation reveals that it is theoretically possible and empirically preferable to make decisions that set aside our moral intuitions, then this would be yet another application of research in cognitive neuroscience. The consequences for ethics might be a move away from reliance on intuitions as data, and closer towards public policy and clinical psychology. This would be greatly informed by psychological research, but in ways that help us discover how to set aside moral appraisals rather than guiding decisions according to them.

Acknowledgements Earlier versions of this paper were presented at the 12th Annual Pitt/CMU Graduate Philosophy Conference, as well as the 36th Annual Society for Philosophy and Psychology, and I wish to thank the audience for their helpful comments. I also owe a great thanks to Liane Young, Justin Sytsma, Joshua Greene, Fiery Cushman, Colin Klein, Jonathon Hricko, Bryan Miller, Jeff Maynes, and John Waterman for their helpful comments and feedback.

References

1. Hauser, Marc. 2006. *Moral minds: How nature designed our universal sense of right and wrong*. New York: Ecco/ Harper Collins.
2. Dwyer, Susan. 2006. How good is the linguistic analogy? In *The innate mind, volume 2: Culture and cognition*, ed. Peter Carruthers, Stephen Laurence, and Stephen Stich, 169–190. Oxford: Oxford University Press.
3. Mikhail, John. 2008. Moral cognition and computational theory. In *Moral psychology, volume 3: The neuroscience of morality: Emotion, brain disorders, and development*, ed. Walter Sinnott-Armstrong, 81–91. Cambridge: MIT.
4. Greene, Joshua, and Jonathan Haidt. 2002. How (and where) does moral judgment work? *Trends in Cognitive Sciences* 6: 517–523.
5. Casebeer, William, and Patricia Churchland. 2003. The neural mechanisms of moral cognition: A multiple-aspect approach to moral judgment and decision-making. *Biology and Philosophy* 18(1): 169–194.
6. Moll, Jorge, Roland Zahn, Ricardo de Oliveira-Souza, Frank Krueger, and Jordan Grafman. 2005. The neural basis of human moral cognition. *Nature Reviews. Neuroscience* 6(10): 799–809.
7. Gazzaniga, Michael. 2005. *The ethical brain*. New York: Dana Press.
8. Greene, Joshua. 2002. *The terrible, horrible, no good, very bad truth about morality and what to do about it*. Dissertation in the Department of Philosophy, Princeton University.
9. Haidt, Jonathan. 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review* 108(4): 814–834.
10. Roskies, Adina, and Shaun Nichols. 2008. Bringing moral responsibility down to earth. *Journal of Philosophy* 105.
11. Huemer, Michael. 2005. *Ethical intuitionism*. New York: Palgrave Macmillan.
12. Hutcheson, Francis. 1725. *An inquiry into the original of our ideas of beauty and virtue; in two treatises*. London: W. and J. Smith.
13. Hume, David. 1975/1751. *Enquiries concerning human understanding and concerning the principles of morals*, ed. L. A. Selby-Bigge. Oxford: Clarendon Press.
14. Flanagan, Owen, and Robert Williams. 2010. What does the modularity of morals have to do with ethics? Four moral sprouts plus or minus a few. *Trends in Cognitive Science*.
15. Greene, Joshua. 2003. From neural "is" to moral "ought": What are the moral implications of neuroscientific moral psychology? *Nature Reviews. Neuroscience* 4: 847–850.
16. Hauser, Marc, Liane Young, and Fiery Cushman. 2008. Reviving Rawls's linguistic analogy: Operative principles and the causal structure of moral actions. In *Moral psychology, volume 2: The cognitive science of morality: Intuition and diversity*, ed. Walter Sinnott-Armstrong, 107–143. Cambridge: MIT.
17. Smart, Jack. 1959. Sensations and brain processes. *Philosophical Review* 68: 141–156.
18. Watson, James. 1930. *Behaviorism*. New York: Norton.
19. Marr, David. 1982. *Vision: A computational investigation into the human representation and processing of visual information*. New York: Freeman.

20. Chomsky, Noam. 2000. *New horizons in the study of language and mind*. Cambridge: Cambridge University Press.
21. Korsgaard, Christine. 1996. *The sources of normativity*. Cambridge: Cambridge University Press.
22. Singer, Peter. 1995. *How are we to live? Ethics in an age of self interest*. New York: Prometheus Books.
23. Mackie, John. 1977. *Ethics: Inventing right and wrong*. New York: Penguin.
24. Joyce, Richard. 2001. *The myth of morality*. Cambridge: Cambridge University Press.
25. Greene, Joshua. 2007. The secret joke of Kant's soul. In *Moral psychology, volume 3: The neuroscience of morality: Emotion, brain disorders, and development*, ed. Walter Sinnott-Armstrong, 81–91. Cambridge: MIT.
26. Dean, Richard. 2010. Does neuroscience undermine deontological theory? *Neuroethics* 3: 43–60.
27. Timmons, Marc. 2008. Toward a sentimentalist deontology. In *Moral psychology, volume 3: The neuroscience of morality: Emotion, brain disorders, and development*, ed. Walter Sinnott-Armstrong, 93–103. Cambridge: MIT.
28. Becker, Selim. 2010. The normative insignificance of neuroscience.
29. McCloskey, Michael. 1983. Intuitive physics. *Scientific American* 248: 122–130.
30. Hirschfeld, Lawrence, and Susan Gelman. 1994. *Mapping the mind: Domain specificity in cognition and culture*. New York: Cambridge University Press.
31. Gigerenzer, Gerd, Peter Todd, and the ABC Research Group. 1999. *Simple heuristics that make us smart*. New York: Oxford University Press.
32. Premack, David, and Ann Premack. 1994. Moral belief: Form vs content. In *Mapping the mind: Domain specificity in cognition and culture*, ed. Lawrence Hirschfeld and Susan Gelman. New York: Cambridge University Press.
33. Doris, John. 2001. *Lack of character*. Cambridge: Cambridge University Press.
34. Flanagan, Owen. 1996. Ethics naturalized: Ethics and human ecology. In *Mind and morals*, ed. L. May, M. Friedman, and A. Clark. Cambridge: MIT Press.
35. Sartre, Jean-Paul. 2007/1947. *Existentialism is a humanism*, trans. Carol Macomber. New Haven: Yale University Press.
36. Sinnott-Armstrong, Walter. 2009. *Morality without God*. Oxford: Oxford University Press.
37. Wilson, Margaret. 2002. Six views of embodied cognition. *Psychonomic Bulletin & Review* 9(4): 625–636.
38. Oakes, Penelope, and John Turner. 1990. Is limited information-processing capacity the cause of social stereotyping? In *The European review of social psychology*, ed. W. Stroebe and M. Hewstone, 111–135. England: Wiley.
39. Taylor, S., S. Fiske, N. Etcoff, and A. Ruderman. 1978. The categorical and contextual bases of person memory and stereotyping. *Journal of Personality and Social Psychology* 36: 778–793.
40. Hirschfeld, L.W. 2001. On a folk theory of society: Children, evolution, and mental representations of social groups. *Personality and Social Psychology Review* 5(2): 107–117.
41. Gil-White, Francisco. 2001. Are ethnic groups biological 'species' to the human brain? *Current Anthropology* 42(4): 515–554.
42. Greenwald, Anthony, Debbie McGhee, and Jordan Schwartz. 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology* 74: 1464–1480.
43. Kelly, Daniel, Edouard Machery, and Ron Mallon. Forthcoming. Racial cognition and normative racial theory.
44. Special issue of journal of *Journal of Personality and Social Psychology* 81(5)
45. Wright, Jen. Personal communication, June 11, 2010.
46. Hudspeth, A.J. 1983. Mechanoelectrical transduction by hair cells in the acousticolateralis sensory system. *Annual Review of Neuroscience* 6: 187–215.
47. Sinnott-Armstrong, Walter. 2008. Framing moral intuitions. In *Moral psychology, volume 2: The cognitive science of morality: Intuition and diversity*, ed. Walter Sinnott-Armstrong, 47–77. Cambridge: MIT.
48. Sunstein, Cass. 2005. Moral heuristics. *Brain and Behavioral Sciences* 28: 531–573.
49. Inbar, Yoel, David Pizarro, and Paul Bloom. 2008. Conservatives are more easily disgusted than liberals. *Cognition and Emotion* 23: 714–725.
50. Isen, Alice. 1987. Positive affect, cognitive processes, and social behavior. *Advances in Experimental Social Psychology* 20: 203–253.
51. Knobe, Josh. 2005. *Attribution and normativity: A problem in the philosophy of social psychology*. Unpublished manuscript, UNC Chapel Hill.
52. Kruger, Joachim, and David Funder. 2004. Towards a balanced social psychology: Causes, consequences, and cures for the problem-seeking approach to social behavior and cognition. *The Behavioral and Brain Sciences* 27: 1–15.
53. Mikhail, John. 2005. Moral heuristics or moral competence? Reflections on sunstein. *Behavioral and Brain Sciences* 28: 557–558.
54. Gigerenzer, Gerd. 2008. Moral intuition = fast and frugal heuristics? In *Moral psychology, volume 2: The cognitive science of morality: Intuition and diversity*, ed. Walter Sinnott-Armstrong, 1–26. Cambridge: MIT.
55. Gigerenzer, Gerd. 2008. Why heuristics work. *Perspectives on Psychological Science* 3(1): 20–29.
56. Green, Lee, and David Mehr. 1997. What alters physicians' decisions to admit to the coronary care unit? *The Journal of Family Practice* 45(3): 219–226.
57. Pinker, Steven. 2007. *The stuff of thought: Language as a window onto human nature*. New York: Viking Press.
58. Simons, Daniel, and Christopher Chabris. 1999. Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception* 28(9): 1059–1074.
59. Schacter, Dan. 2001. *The seven sins of memory: How the mind forgets and remembers*. New York: Houghton Mifflin.