Neurometaethics

David Brax, Department of Philosophy, Lund University

david@brax.nu

…it is not profitable for us at present to do moral philosophy; that should be laid aside at any rate until we have an adequate philosophy of psychology, in which we are conspicuously lacking. (Anscombe,1958)

A meta-ethics for neuroethics

We are a couple of anthologies, a handful of conferences, a journal launch and a couple of quite readable blogs into the discipline known as *neuroethics*. Certain salient data have emerged to evoke interest far beyond the normal range for neuroscience, and for the parts of philosophy for which they claim relevance, making *neuroethics* not only buzzworthy, but also a matter of some concern. There are plenty of reasons, then, to assess some basic questions about this discipline, about its subject matter, nature, and point. Time not only to assess the relevance of the available evidence, its status as such, and the potential of uncovering further evidence, but also to consider the relevance of *philosophy*, its ability and willingness to deal with empirical evidence and contribute to their interpretation.

Reading the proceedings from conferences on the subject, one gets the impression that no one really knows what the term 'neuroethics' is supposed to cover. Presumably, this is because one does not want to constrain the discipline this early on. A short, sufficiently general to be considered accurate, description gives that neuroethics aims to investigate two things: the ethics of neuroscience, and the neuroscience of ethics[1].

[1] Roskies (2007).

Two things can be noted straight away. First: "Ethics of neuroscience" sounds suspiciously like a subdivision of *bioethics*[2]. Nothing *distinctive* about it, in other words, and this aspect will receive little attention in this paper. Secondly: The notion of a "neuroscience of ethics" seems to presuppose that ethics has something to do with how our brains works, and many philosophers will want to stop you right there.

Those convinced that we've already discarded the idea that moral philosophy is part of psychology are unlikely to find a neuroscientific approach any more promising. Even though neuroscience has the potential to reveal new and interesting findings about us[3] this doesn't make it any more *relevant* to the fundamental questions in and about ethics.

Eric Racine recently pointed out that the underdevelopment of this side of the neuroethical project is due to the dominance of meta-ethical *non-naturalism*[4]. This movement, championed by Moore in the early 20-th century, treats ethics as an *autonomous* domain[5], different in kind from (possibly *other*) factual matters.

A different, less general, complaint is that the actual work done in neuroscience is of little relevance: neuroscience is nowhere near the sophistication required to study the relevant faculties yet. An optimistic development of this line of thought is that philosophy should be engaged with this project from the very beginning. Even if we think of philosophy as dealing exclusively with conceptual analysis, concepts are an important part of

[2] Keeping in mind that bioethics itself was originally intended as covering both the ethics of the biosciences and the bioscience of ethics. The latter part of the project now play a marginal role in bioethics (Racine 2008)

[3] Any self-respecting psychologist will tell you that these measurements are still terribly imprecise and their interpretation very much up for grabs (the overpublication of studies with fmri-reports was amusingly revealed by Weisberg et al. (2008)

[4] Racine 2008

[5] Racine believes non-naturalism to be based on the rejection of ethical interpretations of socio-biology in the 19th century the "implications" of which made the notion of empirical relevance unbearable. There might be some truth in this, but it was never the *argument* for non-naturalism. Racine ignores the dominance of non-cognitivist in the 20th century, known to be more sympathetic to empirical investigations.

scientific development. Science is not a purely empirical endeavour, but a *theoretical* one, and developments in science walk hand in hand with conceptual developments. It can, to some extent, be thought to *drive* conceptual development, as well as being driven, and constrained, by concepts. Peter Railton (1989), in a defence of naturalism and in a distinctly Quinean spirit, argued that beliefs once thought to express conceptual necessities can yet come to seem false on the basis of developments in empirical science. Under certain circumstances, we are liable to *revise* our concepts. Adina Roskies recently expressed a similar view

> As we learn more about the neuroscientific basis of ethical reasoning, as well as what underlies self-representation and self-awareness, we may revise our ethical concepts[6].

The general question

The alleged autonomy of ethics raises the more general question whether *anything* in philosophy is addressable by empirical means. Compare with notions like *knowledge* and *thinking*. Epistemology, to some extent is a *normative* issue, much like ethical theory. We ask what we *ought to* believe and think, and we are asked to provide *justification* for beliefs. But we are also interested in what thinking *is*, and *that* could hardly be conceived of outside the realm of the actual. Arguably, the study of thinking today is as much a matter for *cognitive science*, as for epistemology if, indeed, the disciplines should be treated as distinct.

Whatever beliefs are, *we* are certainly capable of having them, and the same seems to go for knowledge. 'Knowledge' defined so that no person has ever been, or are likely to ever be, able to attain it is a concept of little interest. This holds even if this is, in fact, "our" concept of knowledge[7]. This, arguably, was one of the key insights that spurned the naturalistic turn in epistemology.

---

[6] Roskies (2007). The same sentiment is present in Richard Brandts work (See Sturgeon,1982).

[7] A similar reaction to the "error theory" for moral concepts was expressed by David Lewis (1989)

Meta-ethics

Let's accept that we cannot derive an 'ought' from a purely descriptive 'is', and that this inability is due to a *conceptual* impossibility. There are still questions *about* ethics that are matters of fact, even if ethical statements are not. What is, after all, going on in moral thought? While barred from answering directly the normative questions, can neuro-ethics help answer questions *about normativity?* Richard Joyce recently wrote that

> Even if there were an a priori prohibition on deriving evaluative conclusions from factual premises, this need not stand in the way of *meta*ethical implications being drawn from factual premises, for a metaethical claim is not an ethical "ought" claim; it is more likely to be a claim about how we use the word "ought" in ethical discourse, which is a perfectly empirical matter[8].

The sceptic will argue that while factual, meta-ethics is still concerned with *conceptual*, i.e. a priori, facts, and these are not subject to empirical testing. We could reply that this doesn't hold for all concepts: scientific terms are often tested against and redefined in the light of empirical findings.

The problem is that the proposition we tentatively admitted to: that we cannot infer moral facts from natural facts suggests that they are not *reducible* to such facts. Even if we might find that *some* conceptual questions are addressable by empirical means, so the argument goes, moral concepts are not of this kind[9].

Meta-ethics is not only concerned with the analysis of moral concepts, however, but with *all* questions *about* ethics. Questions about moral facts, about moral knowledge and what we are doing when we make moral utterances, all fall within the discipline, broadly conceived. The notion that semantics is somehow theoretically *prior* to these matters is a disposable artefact from the linguistic-turn era. Even if concepts are inaccessible to empirical study: *What* concepts we express seem to be an empirical, a posteriori, matter.

---

[8] Joyce 2008 p 372

[9] (Note that this is a conceptual claim insofar as it is a claim *about* a concept. And now the same dilemma arise for this claim: is it addressable by empirical means or not?)

If meta-ethics is conceived as an interdisciplinary project, we are looking for components that are not merely independently plausible, but mutually supportive. Empirical facts about how moral reasoning in fact works, even if incapable of strictly implying any conceptual facts, might still sit more or less well together with such facts. Seeing how the analysis of normative terms is a highly disputed matter, such indirect indications might be the only source of evidence we've got. Treating meta-ethics as an interdisciplinary theoretical enterprise may very well require putting the notion of conceptual entailment on hold.

Arguably, neuroscience can contribute to *this* project; especially if we are convinced that epistemology should be naturalized.

A moral psychology for meta-ethics

A common conception of meta-ethics is that it consists of a combination of semantics, epistemology and meta-physics. Moral psychology, while arguably fitting the same general description of being *about* normativity, is usually left out. Nevertheless, ethical theories seem to come with favoured moral psychologies, and their empirical accuracy is a factor to be considered. If we believe that moral judgments express proper beliefs, we are committed to the existence and reliability of the cognitive faculties required. Likewise, there are psychological hypotheses that have to be roughly true if normative theories like virtue-theory or utilitarianism are to be plausible. Even though none of those theories are strictly *about* moral thinking, moral psychologies are a reasonable extension of them. Do 'virtues', understood as robust character traits, exist? Do we have the ability to calculate outcomes? Is altruism or amoralism possible? Does proper moral disagreement occur? Are our abilities to detect the good reliable? Can we generalise without bias? These are things quite clearly ripe for empirical assessment.

Empirical moral psychology has a long history, and has been put to use in meta-ethics before the current surge of interest, but its relevance has always

been a controversial issue[10]. The current state of neuroscience offers few new general *arguments* to that discussion; the influence is subtly different.

The data

To be honest, the literature on the neuroscience of ethics is rather limited. There are only a handful of studies actually using *neuroscientific* means and designed experiments, and only a slightly larger set trading on findings neuroscience, even though the studies themselves involve no new data of that kind. The concern so far has mainly been with how the brain works when confronting moral dilemmas. In particular, when faced with versions of the trolley-problem[11]. The moral judgments given by patients with selective brain traumas have also been investigated[12].

To be blunt, what these studies establish is little more than the independently plausible fact that emotions are important to moral judgments, and that moral judgments are regulated by two intimately connected, but anatomically distinct, areas in the brain. What is interesting about these two areas, one roughly correlated with the production of *emotions* and the other with so called "higher cognitive functions", according to the authors of the articles reviewed, is that they can come into conflict[13]. Koenigs et al. (2007) found that damage to emotion specific areas (ventromedial prefrontal cortex) tend to affect how patients react to certain moral dilemmas, causing them to favour the utilitarian option in personalised trolley cases. While both patients and control groups tend to favour this option in *non-personal* trolley cases, where all you have to do to save five people is to pull a lever, thus killing one person, the difference surface when the action required is of a more personal sort: pushing a fat

[10] (Social psychology, Harman, Brandt. The connection between empirical psychology and ethics used to be considerably tighter, see Appiah (2008)
[11] Greene, et al. (2005)
[12] Greene, Damasio et al.
[13] Greene and Haidt (2002) point out that there are no morality *specific* brain areas. Complex cognitive behaviour is usually a distributed matter.

person in front of the trolley[14]. The patients see no difference between the two cases, and judge accordingly, while control groups tend to disallow such actions. In fact, as Jonathan Haidt has demonstrated, non-patients (normal subjects), being part of the control groups, usually have a hard time accounting for the difference as well[15].

The patients are intellectually intact and have a normal baseline mood; their "defect" lies mainly in moral responses and social emotions, like shame and guilt, which tend to be absent, or diverge from the normal. In contrast, their emotional reactions to personal frustrations are not impaired; in fact they tend to react strongly to unfair offers presented to them[16]. The ability to recognise and act in accordance with *norms* seems not to be impaired in these patients, even if the emotional salience of those norms might be. Now, the trolley-case dilemmas are constructed as to invoke *conflicting* norms: Not to kill versus Saving as many lives as you can, and the data seem to imply that the former is more dependent on emotional reactions. It would be interesting to see how patients with a strongly internalised conviction in norms prohibiting the utilitarian option would perform, but this is one of the many experiments that has yet to be devised. Of course, not all patients favour the utilitarian option, the study only show a statistically significant increase in favouring that option.

It is next to impossible to find patients with identical brain damages, not to speak of patients where everything else is equal. The closest thing to such a study would be to examine patients' moral responses before and after such a lesion, alternatively to find a way to *temporarily* knock out the different systems. To my knowledge, this is presently not practically possible and, rather fittingly, there are some moral hurdles to clear as well.

In the cases considered, according to the hypothesis, the negative emotions registering reluctance to comply with the utilitarian option is knocked out,

[14] It is an open question whether this shows that a) those patients manages to disregard the distorting influence of emotion, or b) they are morally challenged, or c) none of the above.

[15] Hauser et al. (2007) argue that some moral judgments (the principle of double effect) are active in moral judgment, but not introspectively *accessible*.

[16] It is note-worthy that such cases target the agent's self-interest and involves *actual* offers, whereas the trolley-cases are other-regarding and involves highly hypothetical scenarios

changing, not cancelling, the outcome in moral judgment. The authors points out that a mere correlation between emotional activity and moral judgments wouldn't settle whether the emotion cause or is caused by moral judgments. But their study does one better: if emotions where a mere consequence of moral judgment, damage to emotional areas would not influence such judgements[17]. The problem with this line of reasoning is that hardly anything in the brain works as simply as that. There are virtually *no* cases where neuronal influences go in only one direction.

Even more striking than the potential conflict between areas in the brain is the remarkable interconnection they exhibit under normal circumstances. Cognitive processes influence emotional responses, and vice versa, and the brain areas involved are remarkably plastic, making it next to impossible to make particular predictions on the basis of these general observations. In addition, the role of emotions in the production of *particular* instances of moral judgments may betray their role in moral reasoning as such. Emotions could be understood as "short-hand" for complex dispositional states, as suggested by Antonio Damasio's "somatic marker" hypothesis[18]. They might play a role in moral judgement similar to perception in judgments about shapes and colours: i.e. as fallible sources of evidence, and just one, possibly the canonical, way of attaining knowledge or collecting evidence about these things. Emotion could be seen as mere evidence, as a factor to be considered in cognitive processes, like practical deliberation. If you loose it, you loose *evidence*, not the ability to judge morally. There is a lot more work to be done here, and it is likely that for quite some time, the data attainable will be insufficient to make the choice between differing philosophical interpretations.

Valdesolo and Destone (2007) demonstrate that contexts can influence the affects on the basis of which we form moral judgments. They found that

[17] Jonathan Haidt (2001) claims that moral reasoning and justification comes *after* the emotional response and do not *cause* or *ground* our moral judgements, which, he thinks, are first and foremost expressions of emotions.

[18] Damasio (1994)

manipulating the affective state of the subject influence moral judgment in reaction to the trolley-case in a way closely resembling the judgments given by patients with emotion-specific damages mentioned above.

> Simply put, environment-induced feelings of positivity at the time of judgment might reduce the perceived negativity, or aversion "signal," of any potential moral violation and, thereby, increase utilitarian responding.

One reflection on this fact, if it is established as such, is that the importance of context puts the experiment itself into question. The context in which we *contemplate* a trolley case is quite different from the context in which we would *encounter* such a dilemma. We usually don't factor in context when we consider thought-experiments; indeed, we are encouraged not to. The same point applies to study of damaged subjects mentioned above. Our moral concepts arguably developed under quite specific conditions, so one might wonder whether any conclusions could be drawn from how we deal with thought experiments, and how intuitive reactions to those can be manipulated. On the other hand, we have good reason to believe that we evolved to be able to make decisions based on abstract hypothetical scenarios. In short, it is not easy to tell what constitutes the ideal conditions for moral judgments, and thus not what judgments we should take to be tracking our best ability to reason about morality.

The fact that emotional manipulation can influence moral attitudes is, of course, nothing new. We tend to attribute changes in emotional state to the most salient change we can find in the environment. Subtle, unnoticed causes of emotional change can therefore easily be misattributed, in particular if there is no established "standard of correctness" for emotional reactions[19]. In a recent experiment, Williams and Bargh[20]managed to manipulate subjects to judge a person as more or less suited for a job, merely by letting them hold a warm or cold cup for a few minutes before the interview. One reaction to these observations is to say that *all* evaluative and moral judgments are such "mis-

[19] This might vary over cultures where emotions are more or less strictly "legislated". The tendency to (mis)attribute emotions is sometimes called the "pathetic fallacy".
[20] Williams and Bargh, science (2008)

attributions", which comes close to the error theory[21], and could be used as an argument for non-cognitivism. While external objects might be part of the cluster of causes that determine the emotion most of the work is done by internal processes. There is need for further study into the nature of unconscious affect influencing moral "reasoning".

Experimental meta-ethics?

Greene and Haidt speak of an *affective revolution* in moral psychology due to findings in evolutionary psychology and primatology, linking moral judgments to emotions and understanding feelings of approval as affect-laden intuitions[22]. *Reasoning* matter but does so mostly in social contexts, where we try to influence each other by appealing to reasons to act. In the few cases where such persuasion is successful, the emotionally inclined theorist might say, it is because the reasons appealed to evoke the emotions that actually cause the change of mind. The reasons themselves play only a mediating role in the explanation. Reasoning without emotion, in a distinctly Humean claim, does nothing.

Now, do any of the findings mentioned above do *anything* to further or undermine any meta-ethical views? Richard Joyce advocates the joining of meta-ethics with a posteriori methods in general, but is critical of the attempts made so far. Joyce (2007) considers whether neuroscience can support moral emotivism or undermine moral rationalism, and finds the support not only lacking, but as misunderstanding the nature of meta-ethics. "Emotivism" as understood in meta-ethics is a *semantic* claim about the meaning of moral terms, not about the cause of moral judgments. Cognitivists can accept that emotions usually figure in the production of moral statements, and emotivists can claim moral statements to express sentiments even when the speaker does not have them, thus being immune to findings in empirical studies like those mentioned. If such studies are to be relevant, we need further argument

[21] Joyce (2006).
[22] Greene and Haidt 2002

to show the cause of moral judgments to be relevant to what they mean, and that is a theoretical, meta-ethical, matter, not an empirical one.

But let's examine this more closely: A systematic correlation between moral judgments and emotions does not on its own imply anything about the meaning of moral terms[23]. But surely, if there is no other undisputed way of settling what moral terms mean, appeal to such facts should be taken into account. Meta-ethics is a discipline absolutely riddled with controversies, so there is considerable interest in finding *some* source of evidence. One quite apparent role for neuroscience and for the cognitive and affective sciences in general, is to complete, and partly to replace, the role played by "phenomenology", and to investigate the role and reliability of *intuitions* with experimental measures.

A metaethical account wanting to make use of findings in neuroscience should find a connection between the *typical causes* of moral statements and the contents of moral concepts. One such account is *naturalism*, in the sense developed by Richard Boyd, Peter Railton and others in the late eighties[24], trading on a causal semantics, and it seems to fit nicely an ambitious neuroethical project. Naturalism in this sense is not only a semantic theory, but latch on to naturalist tendencies in epistemology and reference as well. In particular, it claims an indispensable role for causal processes in the production of true beliefs. Of course, the production of moral *beliefs* is distinct from the production of moral judgments, so a different set of experiments from the ones considered is required, and it is far from clear how those experiments should be designed.

There are *pragmatic* concerns in favour of naturalism. Richard Brandt suggested naturalism as a theory of what we'd *better* mean with moral terms if

[23] Elsewhere (2006) Joyce favours a genealogical debunking explanation of moral language. Joyce thinks the relevance of empirical studies to meta-ethics lies in the *epistemology* for moral beliefs.

[24] Boyd 1988, Railton 1989. Another meta-ethical framework not alien to engaging with empirical science is non-cognitivism, but I will not address that here.

they are to make sense, and be of any use. We might have to revise our moral concepts for naturalism to be true, and we should. The concepts figuring in ordinary talk is only too likely to be confused, and philosophical theories, much like scientific theories, should *improve* on ordinary talk, not merely mirror it[25]. Indeed, the theory offering the best fit with ordinary moral talk might not, in fact, be the best moral theory.

To judge whether the naturalist approach is plausible, we need to know what kind of concepts moral concepts are. A very ambitious neuroethicist would say that *this* question could be studied with empirical means. Rather than investigating the role of emotion in particular moral judgments, its role in *learning* moral concepts might be relevant to this question[26]. The influence of emotions on particular judgments can be taken as evidence for such a theory, but the research need to be deeper and more varied than the studies made so far, if these questions are to be addressed.

As I said, the "neuroscience of ethics" to date consist in a very limited number of studies, and their authors are typically rather modest in their claims. When it comes to neuroscience, we are far from the technology needed for the experiments we would want to do, and far from the conceptual sophistication needed to interpret the results. The future for the neuroethical project depends, I believe, on developing a theoretical framework for such interpretation, and on settling what would serve as a reasonable confirmation of a meta-ethical hypothesis.

Despite my optimism about the direction philosophy has taken in recent years, notably the "experimental philosophy" movement[27], I very much doubt whether neuroscience will be able to tell us which the true meta-ethical theory is. But

[25] See Sturgeon (1982) Also Railton 1989, Lewis (1989)
[26] Of course, this was known to philosophers like David Hume, Adam Smith and John Stuart Mill; all of whom would have been likely to take a keen interest in neuroethics if given the chance.
[27] See Appiah (2008)

this does not mean that neuroscience is *irrelevant* to meta-ethics. In fact none of the domains relevant to meta-ethics, not semantics, not epistemology, not meta-physics strictly *determine* anything in meta-ethics. What they provide are theoretical *possibilities*. A meta-ethical view consists in combinations of positions in these philosophical fields as applied to ethical concepts. Does this mean that semantics, meta-physics, epistemology are *irrelevant* to meta-ethics? Certainly not. We are looking for a coherent theory that fits neatly into our set of plausible beliefs, and an empirically adequate moral psychology should be a part of that theory.

There is nothing special about the "irrelevance" of empirical work, unless you presuppose that philosophy deals with conceptual analysis, conceived as an armchair matter, and this is a controversial, and decreasingly popular, claim.

Neuro-metaethics, i.e. a meta-ethical theory incorporating results from neuroscience, should not be imperialistic, it should not presume to make ethics or meta-ethics, into a subdiscipline to general neuroscience. In fact, it couldn't possibly. But just as neuroethics is not imperialistic, philosophy shouldn't be protectionist: the quest for knowledge is not well served by insisting on disciplinary autonomy. Minimally, the Neuroscience of Ethics is of interest in and of itself, as one of the second-order questions about ethics.

The meta-ethics that *can* make a joint venture with neuroscience and psychology will have a great next couple of hundred years. What will happen to the rest of it, I just don't now.

References:

Appiah, Anthony Kwame: *Experimental ethics* (2008)

Anscombe, G.E.M.: *Modern Moral Philosophy*, Philosophy vol 33, 1-19, (1958)

Boyd, Richard: *How to be a moral realist*, in *Essays on Moral Realism* ed. Sayre-McCord (1988)

Damasio, Antonio: *Descartes Error* (1994)

Greene, Joshua and Haidt, Jonathan: *How (and where) does moral judgment work?* in Trends in Cognitive Sciences vol 6, 517-523, (2002)

Greene, Joshua, et al: *The neural bases of cognitive conflict and control in moral judgment* in Neuron vol 44, 389-400, (2004)

Hauser et al: *A dissociation between moral judgment and rational justifications* in Mind and Language vol 22, 1-21, (2007)

Haidt, Jonathan: *The emotional dog and it's rational tail*, Psychological review vol 108, 814-834, (2001)

Joyce, Richard: *Meta-ethics and the empirical sciences*,Philosophical Explorations vol. 9 133-147 (2006)

Joyce, Richard: *What Neuroscience can (and Cannot) Contribute to Metaethics*, in *Moral Psychology vol. 3*, ed. Sinnott-Armstrong (2008)

Koenigs, et al.: *Damage to the prefrontal cortex increases utilitarian moral judgements*, Nature (2007)

Lewis, David, *Dispositional theories of value* (1989)

Racine, Eric: *Which naturalism for neuroethics*, Bioethics vol 22, 92-100, (2008)

Railton, Peter: *Naturalism and prescriptivity*, Social Philosophy and Prescriptivity vol 95, 151-174, (1989)

Roskies, Adina: *Neuroethics for the new millennium*, in *Defining Right and Wrong in Brain Science* ed. Glannon, Walter, (2007)

Sturgeon, Nicholas: *Brandt's moral empiricism*, the Philosophical Review, vol. 91, 389-422, (1982)

Valdesolo, Piciardo and Desteno, David: *Manipulations of context shape moral judgment*", Psychological Review vol. 17, 476-477, (2006)

Weisberg et al.: *The seductive allure of neuroscience explanations*, Journal of Cognitive Neuroscience, vol. 20, 470-477, (2008)

Williams, Laurence and Barg, John: *Experiencing physical warmth promotes interpersonal warmth*, Science vol .322, 606-609, (2008)