Ethical Decoupling

Stephen A. Butterfill
< s.butterfill@warwick.ac.uk >

Thursday, 27th June 2024

Contents

1 Introduction **2**

2 A Problem with the Leading Theory **2**
2.1 Greene et al's Dual-Process Theory 2
2.2 Mixed Behavioural Evidence for This Theory 3
2.3 Suggestion 4

3 Bitterness and Food Rejection **4**
3.1 Step 1: Change the Question 4
3.2 Step 2: Distinguish Ethical Abilities 5
3.3 Step 3: Find a Non-ethical Model 6
3.4 Bitterness 6
3.5 Disgust 8
3.6 Step 4: Borrow an Idea about Normativity 8

Glossary **9**

1. Introduction

Premise: There are behaviours which are sometimes, but not always, controlled by ethical attitudes.

Terminology: *ethical decoupling* occurs when behaviours with ethical significance are not under the control of ethical attitudes. (The behaviours may, but need not, conflict with ethical attitudes.)

Behaviours with ethical significance potentially include caring for another, cooperating with a group, sanctioning someone for an omission, enslaving an enemy, filial infanticide, cannibalism, sexual activity with another species or a corpse, and eating your own vomit.

Guiding question:

> What is the relation between ethical attitudes and the behaviours which they sometimes but not always control?

Schematic answer: ethically significant behaviours are consequences of two or more processes, some but not all of which involve ethical attitudes.

Challenge: characterise the process and the behaviours.

Idea: the leading account of ethical cognition can be used to this end …

2. A Problem with the Leading Theory

2.1. Greene et al's Dual-Process Theory

Greene et al offer a dual-process theory of ethical cognition:

> 'this theory associates controlled cognition with utilitarian (or consequentialist) moral judgment aimed at promoting the "greater good" (Mill, 1861/1998) while associating automatic emotional responses with competing deontological judgments that are naturally justified in terms of rights or duties (Kant, 1785/1959).' (Greene 2015, p. 203)

The theory was developed in part to explain otherwise apparently anomalous responses to moral dilemmas. In particular, people have substantially different attitudes to killing one person in order to save several others depending on whether the killing involves pressing a switch (as in the Switch dilemma) or whether it involves dropping someone through a trapdoor into the path of great danger (as in the Footbridge dilemma).[1]

[1] See Greene (2015, p. 203): 'We developed this theory in response to a long-standing

What is the explanation Greene et al's theory offers?

> 'this pattern of judgment [Switch—yes; Footbridge—no] reflects the outputs of distinct and (in some cases) competing neural systems [...] The more "personal" harmful action in the footbridge case, pushing the man off the footbridge, triggers a relatively strong negative emotional response, whereas the relatively impersonal harmful action in the switch case does not.' (Greene 2015, pp. 203—4)

2.2. Mixed Behavioural Evidence for This Theory

One prediction of the theory is that increasing time pressure should increase the influence of automatic emotional processes relative to the influence of controlled cognition, which in turn should make responses that are characteristically deontological more likely.

This prediction is supported by (Suter & Hertwig 2011), among others.[2] But Bago & De Neys (2019) consider what happens when subjects first make a moral judgement under time pressure and extraneous cognitive load and then, just after, make another moral judgement (in answer to the same question) with no time pressure and no extraneous cognitive load. They report:

> 'Our critical finding is that although there were some instances in which deliberate correction occurred, these were the exception rather than the rule. Across the studies, results consistently showed that in the vast majority of cases in which people opt for a [consequentialist] response after deliberation, the [consequentialist] response is already given in the initial phase' (Bago & De Neys 2019, p. 1794).

Rosas & Aguilar-Pardo (2020) find, conversely to what Greene et al's theory predicts, that subjects are less likely to give characteristically deontological responses under extreme time pressure.

The converse finding of Rosas & Aguilar-Pardo (2020) is not theoretically unmotivated—there are also some theoretical reasons for holding that automatic emotional processes should support characteristically utilitarian responses (Kurzban et al. 2012).

As there is a substantial body of neuropsychological evidence in favour of Greene et al's theory (reviewed in Greene 2014), its defenders may be little

philosophical puzzle ... Why do people typically say "yes" to hitting the switch, but "no" to pushing?'

[2] See also Trémolière & Bonnefon (2014) and Conway & Gawronski (2013) (who manipulated cognitive load).

moved by the mixed behavioural evidence. But there is a reason, not decisive but substantial, to expect mixed evidence more generally …

2.3. Suggestion

While we have not seen decisive evidence against it, we have seen enough to motivate seeking alternatives.

3. Bitterness and Food Rejection

I do not have an alternative to Greene et al's theory, but there are some steps which may take us towards constructing one.

3.1. Step 1: Change the Question

Greene et al arrive at their theory by asking a question about responses to sacrificial dilemmas:

> 'We developed this theory in response to a long-standing philosophical puzzle known as the trolley problem' (Greene 2015, p. 203; see Greene 2023)

A problem for this starting point is that there appear to be confounds in the dilemmas that give rise to trolley problems. Indeed, the mixed pattern of evidence for and against Greene et al's theory might be explained by their choice of vignettes using trolley cases as stimuli. Waldmann et al. (2012, p. 288) offers a brief summary of some factors which have been considered to influence responses including:

- whether an agent is part of the danger (on the trolley) or a bystander;
- whether an action involves forceful contact with a victim;
- whether an action targets an object or the victim;
- how far the agent is from the victim;[3] and
- how the victim is described.

Other factors include whether there are irrelevant alternatives (Wiegmann et al. 2020); and order of presentation (Schwitzgebel & Cushman 2015).

They comment:

[3] After this review was published, Nagel & Waldmann (2013) provided substantial evidence that distance may not be a factor influencing moral intuitions after all (the impression that it does was based on confounding distance with factors typically associated with distance such as group membership and efficacy of action).

> 'A brief summary of the research of the past years is that it has been shown that almost all these confounding factors influence judgments, along with a number of others [...] it seems hopeless to look for the one and only explanation of moral intuitions in dilemmas. The research suggests that various moral and nonmoral factors interact in the generation of moral judgments about dilemmas' (Waldmann et al. 2012, pp. 288, 290).

For proponents of Greene et al's view, this might be taken as encouragement. Yes, the evidence is a bit mixed. But perhaps what appears to be evidence falsifying predictions of the view will turn out to be merely a consequence of extraneous, nonmoral factors influencing judgements.

Alternatively, Waldmann et al.'s observation could be taken to suggest that few if any of the studies relying on dilemmas presented in vignette form provide reliable evidence about moral factors since they do not adequately control for extraneous, nonmoral factors. As an illustration, Gawronski et al. (2017) note that aversion to killing (which would be characteristically deontological) needs to be separated from a preference for inaction. When considering only aversion to killing, time pressure appears to result in characteristically deontological responses, which would support Greene et al's theory (Conway & Gawronski 2013). But when aversion to killing and a preference for inaction are considered together, Gawronski et al. (2017) found evidence only that time pressure increases preferences for inaction.

While the combination of mixed behavioural evidence and methodological challenges associated with using dilemmas presented in vignettes does not provide a case for rejecting Greene et al's view, it does motivate considering fresh alternatives.

3.2. Step 2: Distinguish Ethical Abilities

Greene et al's theory treats ethical abilities as all explained by a single fast process and a single slow process. Supporting this is a premise about ethical abilities having a single function:

> 'morality is a suite of cognitive mechanisms that enable otherwise selfish individuals to reap the benefits of cooperation.' (Greene 2015, p. 198)

Others suggest that different ethical abilities have different functions (for example, Haidt & Graham 2007). In particular, concerns about purity are probably distinct from concerns about harm (Chakroff et al. 2017).[4]

[4] Not everyone accepts this claim (Schein et al. 2016).

The diversity of ethical abilities indicates that they may not all have one function:

- care
- cooperation (Boyd & Richerson 2022)
- inequality aversion (Brosnan & de Waal 2014)
- balance authority vs autonomy (Wengrow & Graeber 2018)[5]
- discern impurity (Atari et al. 2022)

I will focus on purity.

3.3. Step 3: Find a Non-ethical Model

To guide our thinking, it is useful to have a non-ethical case which is well understood and in which attitudes and behaviours are decoupled.

To this end, consider food-rejection behaviours.

All animals need to balance the need to eat enough of the best available foods against the risk of ingesting toxins. To this end, they exhibit complex pattern of food-rejection behaviours.

Food-rejection behaviours can be driven by attitudes about poison (for example, you may know something about the origins of a food that otherwise appears delicious). But they can also be driven by aversion to bitterness. Importantly, this can occur independently of, and even counter to, your attitudes.

3.4. Bitterness

How does aversion to bitterness work? Poisonous foods are often bitter. Although the correlation between bitterness and toxicity is not super strong and bitter things can be beneficial (such as caffeine, for example), there is broad consensus that avoiding poisons is one of the functions of sensitivity

[5] Wengrow & Graeber (2018) describe two groups, one of which eschewed slavery, the other of which used it in a complex hierarchical society. 'If the ethics of their Californian neighbors bore comparison with mercantile values in early modern Europe, those of the Northwest Coast more closely resembled the aristocratic values of high feudalism. Societies comprised household estates divided into hereditary ranks of nobles, commoners, and slaves. Slaveholding was a defining attribute of nobility , and from Alaska south to W ashington state, intergroup slave raids were endemic. Nobles alone enjoyed the ritual prerogative of engaging with guardian spirits who conferred access to prestigious titles, which defined the legal contents of an estate. Commoners voluntarily provided labor and services to noble kin, who vied for their allegiance by offering spectacular feasts, entertainment, and the pleasure of vicarious participation in heroic exploits.' (Wengrow & Graeber 2018, p. 239)

to bitterness (Nissim et al. 2017). Animals who encounter a greater proportion of poisonous foods in their normal diet (herbivores) show both higher sensitivity to bitterness (Li & Zhang 2014) and a higher tolerance for it (Ji 1994). This makes sense because herbivores need to take more risks: they could not get enough to eat if they rejected everything bitter. Further, two changes in diet which reduce exposure to toxins, namely eating more animals or cooking with fire, may have gradually reduced sensitivity to bitterness (Wang et al. 2004).

Bitterness appears innately aversive. A range of animals including sea anemones become averse to a food type after a single bitter encounter (Garcia & Hankins 1975). In rats, '[t]he modal elicitor of aversive behaviours is a bitter, normally avoided substance like quinine, which evokes chin rubs, gapes, face washes, forelimb flails, and paw treads' (Forestell & LoLordo 2003, p. 141; Grill & Norgren 1978). And in humans mixing a neutral flavour (vanilla, for example) with a bitter substance can reduce liking for that flavour (Baeyens et al. 1996; Dickinson & Brown 2007).

This suggests a minimal model for toxicity. Instead of a special-purpose learning mechanism, what is needed is just for bitter things to be aversive. With aversion to bitterness in place, ordinary learning mechanisms will reduce exposure to bitter things.[6]

Not everything bitter is bad. As well as indicating toxicity, bitterness is also associated with medicinal properties. Nissim et al. (2017) suggest that animals may exploit this association by eating bitter substances when ill. To illustrate, chimpanzees suffering from diarrhoea and other symptoms of parasite infection extract and chew an extremely bitter pith, which appears to improve their health (Huffman 2001, p. 939).

[6] There is a question about special-purpose learning mechanisms for flavour; see (Dickinson & Brown 2007, p. 42): 'The important point about this parallel between human evaluative conditioning and rat aversion conditioning for flavors lies with the issue of whether these forms of conditioning engage nonstandard learning processes. It is generally agreed that flavor aversion conditioning shows all the main phenomena of Pavlovian conditioning (Revusky, 1971), including blocking (Gillan & Domjan, 1977), and, therefore, the fact that two forms of conditioning can be dissociated in terms of their effective conditioned stimuli and response measures does not necessary imply that they recruit learning processes that operate according to differ- ent principles. It may well be that evaluative conditioning and contingency learning are mediated by dissociable systems but that both these systems are governed by similar principles. The critical test of whether pretraining a flavor as a predictor of the sugar or tween would block evaluative conditioning to another flavor during compound flavor training remains to be examined.'

3.5. Disgust

I suggest that food-rejection behaviours are a useful (if limited) model for thinking about purity-related behaviours.

Purity solves a problem: play and exploration bring many benefits, but can also increase the risk of exposure to pathogens. Concerns about purity seem to function to reduce the risk of exposure to pathogens (Atari et al. 2022; van Leeuwen et al. 2012).

Following the food-rejection model, we can compare bitterness to disgust.[7] Disgust has a function linked to disease-avoidance (Oaten et al. 2009). And disgust can drive at least some purity-related behaviours—and even some attitudes (Wang et al. 2019).

The analogy is limited by the complexity of disgust (see, for example, Tybur et al. 2013).

Just here an objection arises: behaviours driven by disgust are not intrinsically ethical at all. There is, therefore, no possibility of using them to explain ethical decoupling.

3.6. Step 4: Borrow an Idea about Normativity

To reply to the above objection, we need to identify a sense in which some disgust driven behaviours are intrinsically ethical.

Normativity is a mark of the ethical:

> 'When it comes to morality, the most basic issue concerns our capacity for normative guidance: our ability to be motivated by norms of behavior …' (FitzPatrick 2021)

But how to get normativity into the picture? Michael and Butterfill (in preparation) offer the notion of a *minimal norm*. This is a pattern of behaviour which exists in part because of others' responses to behaviors which conform to, or violate, the pattern; where these responses have the purpose of upholding conformity to the pattern.

[7] It's just possible that there is a much closer connection between bitterness and ethical behaviours. In adults at least, unfairness can also sensations of bitterness. See Chapman et al. (2009), who establish that (1) responses to bitterness are marked by activation of the levator labii muscle 'which raises the upper lip and wrinkles the nose'; (2) bitter responses are made not just to bitter tastes but also to 'photographs of uncleanliness and contamination-related disgust stimuli, including feces, injuries, insects, etc.'; and (3) in a dictator game, 'objective (facial motor) signs of disgust that were proportional to the degree of unfairness they experienced.' If Chapman et al. (2009) are right that a limited but useful range of moral violations can produce bitter sensations, general-purpose learning mechanisms could produce aversion to actions that generate these moral violations.

Where disgust underpins purity-related minimal norms, we have intrinsically ethical behaviour.

And because digust can underpin purity-related minimal norms independently of normative attitudes, we have ethical decoupling..

Glossary

characteristically deontological According to Greene, a judgement is *characteristically deontological* if it is one in 'favor of characteristically deontological conclusions (eg, "It's wrong despite the benefits")' (Greene 2007, p. 39). According to Gawronski et al. (2017, p. 365), 'a given judgment cannot be categorized as deontological without confirming its property of being sensitive to moral norms.' 3

Drop A dilemma; also known as *Footbridge*. A runaway trolley is about to run over and kill five people. You can hit a switch that will release the bottom of a footbridge and one person will fall onto the track. The trolley will hit this person, slow down, and not hit the five people further down the track. Is it okay to hit the switch? 9

dual-process theory Any theory concerning abilities in a particular domain on which those abilities involve two or more processes which are distinct in this sense: the conditions which influence whether one mindreading process occurs differ from the conditions which influence whether another occurs. 2

Footbridge A dilemma; also known as *Drop*. A runaway trolley is about to run over and kill five people. You can hit a switch that will release the bottom of a footbridge and one person will fall onto the track. The trolley will hit this person, slow down, and not hit the five people further down the track. Is it okay to hit the switch? 2, 3

Switch A dilemma; also known as *Trolley*. A runaway trolley is about to run over and kill five people. You can hit a switch that will divert the trolley onto a different set of tracks where it will kill only one. Is it okay to hit the switch? 2, 3

Transplant A dilemma. Five people are going to die but you can save them all by cutting up one healthy person and distributing her organs. Is it ok to cut her up? 9

Trolley A dilemma; also known as *Switch*. A runaway trolley is about to run over and kill five people. You can hit a switch that will divert the trolley onto a different set of tracks where it will kill only one. Is it okay to hit the switch? 9

trolley cases Scenarios designed to elicit puzzling or informative patterns of judgement about how someone should act. Examples include Trolley, Transplant, and Drop. Their use was pioneered by Foot (1967) and Thomson (1976), who aimed to use them to understand ethical considerations around abortion and euthanasia. 4

References

Atari, M., Reimer, N. K., Graham, J., Hoover, J., Kennedy, B., Davani, A. M., Karimi-Malekabadi, F., Birjandi, S., & Dehghani, M. (2022). Pathogens are linked to human moral systems across time and space. *Current Research in Ecological and Social Psychology*, *3*, 100060.

Baeyens, F., Crombez, G., De Houwer, J., & Eelen, P. (1996). No Evidence for Modulation of Evaluative Flavor–Flavor Associations in Humans. *Learning and Motivation*, *27*(2), 200–241.

Bago, B. & De Neys, W. (2019). The Intuitive Greater Good: Testing the Corrective Dual Process Model of Moral Cognition. *Journal of Experimental Psychology: General*, *148*(10), 1782–1801.

Boyd, R. & Richerson, P. J. (2022). Large-scale cooperation in small-scale foraging societies. *Evolutionary Anthropology: Issues, News, and Reviews*, *31*(4), 175–198.

Brosnan, S. F. & de Waal, F. B. M. (2014). Evolution of responses to (un)fairness. *Science*, *346*(6207), 1251776.

Chakroff, A., Russell, P. S., Piazza, J., & Young, L. (2017). From impure to harmful: Asymmetric expectations about immoral agents. *Journal of Experimental Social Psychology*, *69*, 201–209.

Chapman, H. A., Kim, D. A., Susskind, J. M., & Anderson, A. K. (2009). In Bad Taste: Evidence for the Oral Origins of Moral Disgust. *Science*, *323*(5918), 1222–1226.

Conway, P. & Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision making: A process dissociation approach. *Journal of personality and social psychology*, *104*(2), 216–235.

Dickinson, A. & Brown, K. J. (2007). Flavor-evaluative conditioning is unaffected by contingency knowledge during training with color-flavor compounds. *Animal Learning & Behavior*, *35*(1), 36–42.

FitzPatrick, W. (2021). *Morality and Evolutionary Biology* (Spring 2021 ed.). Metaphysics Research Lab, Stanford University.

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford Review*, *5*, 5–15.

Forestell, C. A. & LoLordo, V. M. (2003). Palatability Shifts in Taste and Flavour Preference Conditioning. *The Quarterly Journal of Experimental Psychology Section B*, *56*(1b), 140–160.

Garcia, J. & Hankins, W. (1975). *The Evolution of Bitter and the Acquisition of Toxiphobia*, (pp. 39–45). Elsevier.

Gawronski, B., Armstrong, J., Conway, P., Friesdorf, R., & Hütter, M. (2017). Consequences, norms, and generalized inaction in moral dilemmas: The CNI model of moral decision-making. *Journal of personality and social psychology*, *113*(3), 343–376.

Greene, J. D. (2007). The Secret Joke of Kant's Soul. In W. Sinnott-Armstrong (Ed.), *Moral Psychology, Vol. 3* (pp. 35–79). MIT Press.

Greene, J. D. (2014). Beyond Point-and-Shoot Morality: Why Cognitive (Neuro)Science Matters for Ethics. *Ethics*, *124*(4), 695–726.

Greene, J. D. (2015). The cognitive neuroscience of moral judgment and decision making. In *The Moral Brain: A Multidisciplinary Perspective* (pp. 197–220). Cambridge, MA, US: MIT Press.

Greene, J. D. (2023). *Trolleyology: What it is, why it matters, what it's taught us, and how it's been misunderstood*, (pp. 158–181). Classic Philosophical Arguments. Cambridge: Cambridge University Press.

Grill, H. J. & Norgren, R. (1978). The taste reactivity test. I. Mimetic responses to gustatory stimuli in neurologically normal rats. *Brain Research*, *143*(2), 263–279.

Haidt, J. & Graham, J. (2007). When Morality Opposes Justice: Conservatives Have Moral Intuitions that Liberals may not Recognize. *Social Justice Research*, *20*(1), 98–116.

Huffman, M. A. (2001). Self-Medicative Behavior in the African Great Apes: An Evolutionary Perspective into the Origins of Human Traditional Medicine: In addition to giving us a deeper understanding of our closest

living relatives, the study of great ape self-medication provides a window into the origins of herbal medicine use by humans and promises to provide new insights into ways of treating parasite infections and other serious diseases. *BioScience*, *51*(8), 651–661.

Ji, G. (1994). Is the bitter rejection response always adaptive? *Physiology & Behavior*, *56*(6).

Kurzban, R., DeScioli, P., & Fein, D. (2012). Hamilton vs. Kant: Pitting adaptations for altruism against adaptations for moral judgment. *Evolution and Human Behavior*, *33*(4), 323–333.

Li, D. & Zhang, J. (2014). Diet Shapes the Evolution of the Vertebrate Bitter Taste Receptor Gene Repertoire. *Molecular Biology and Evolution*, *31*(2), 303–309.

Nagel, J. & Waldmann, M. R. (2013). Deconfounding distance effects in judgments of moral obligation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(1), 237.

Nissim, I., Dagan-Wiener, A., & Niv, M. Y. (2017). The taste of toxicity: A quantitative analysis of bitter and toxic molecules. *IUBMB Life*, *69*(12), 938–946.

Oaten, M., Stevenson, R. J., & Case, T. I. (2009). Disgust as a disease-avoidance mechanism. *Psychological Bulletin*, *135*, 303–321.

Rosas, A. & Aguilar-Pardo, D. (2020). Extreme time-pressure reveals utilitarian intuitions in sacrificial dilemmas. *Thinking & Reasoning*, *26*(4), 534–551.

Schein, C., Ritter, R., & Gray, K. (2016). Harm Mediates the Disgust-Immorality Link. *Emotion*, *16*(6), 862–876.

Schwitzgebel, E. & Cushman, F. (2015). Philosophers' biased judgments persist despite training, expertise and reflection. *Cognition*, *141*, 127–137.

Suter, R. S. & Hertwig, R. (2011). Time and moral judgment. *Cognition*, *119*(3), 454–458.

Thomson, J. J. (1976). Killing, Letting Die, and The Trolley Problem. *The Monist*, *59*(2), 204–217.

Trémolière, B. & Bonnefon, J.-F. (2014). Efficient Kill–Save Ratios Ease Up the Cognitive Demands on Counterintuitive Moral Utilitarianism. *Personality and Social Psychology Bulletin*, *124*(3), 379–384.

Tybur, J. M., Lieberman, D., Kurzban, R., & DeScioli, P. (2013). Disgust: Evolved function and structure. *Psychological Review*, *120*(1), 65–84.

van Leeuwen, F., Park, J. H., Koenig, B. L., & Graham, J. (2012). Regional variation in pathogen prevalence predicts endorsement of group-focused moral concerns. *Evolution and Human Behavior*, *33*(5), 429–437.

Waldmann, M. R., Nagel, J., & Wiegmann, A. (2012). Moral Judgment. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford Handbook of Thinking and Reasoning* (pp. 274–299). Oxford: Oxford University Press.

Wang, R., Yang, Q., Huang, P., Sai, L., & Gong, Y. (2019). The Association Between Disgust Sensitivity and Negative Attitudes Toward Homosexuality: The Mediating Role of Moral Foundations. *Frontiers in Psychology*, *10*.

Wang, X., Thomas, S. D., & Zhang, J. (2004). Relaxation of selective constraint and loss of function in the evolution of human bitter taste receptor genes. *Human Molecular Genetics*, *13*(21), 2671–2678.

Wengrow, D. & Graeber, D. (2018). "Many Seasons Ago": Slavery and Its Rejection among Foragers on the Pacific Coast of North America. *American Anthropologist*, *120*(2), 237–249.

Wiegmann, A., Horvath, J., & Meyer, K. (2020). Intuitive expertise and irrelevant options. *Oxford studies in experimental philosophy*, *3*, 275–310.