Evolution, utilitarianism, and normative uncertainty: the practical significance of debunking arguments

Abstract

It has been argued that evolutionary considerations favour utilitarianism by selectively debunking its competitors. However, evolutionary considerations also seem to undermine the practical significance of utilitarianism, since common-sense beliefs about well-being seem like prime candidates for evolutionary debunking. We argue that the practical significance of utilitarianism is not undermined in this way if we understand the requirements of practical rationality as sensitive to normative uncertainty. We consider the view that rational decision-making under normative uncertainty requires maximizing expected choice-worthiness, as well as the possibility that different theories' choice-worthiness rankings are not all interval-scale measurable nor intertheoretically comparable.

1 Introduction

Many philosophers believe that evolutionary considerations debunk whatever ethical beliefs they explain, drawing on the assumption that natural selection doesn't 'track the truth' when it comes to ethics. If some evaluative disposition has been favoured by selection - so the thought goes - then the truth-value of any associated ethical belief is entirely irrelevant in explaining the fitness-advantages associated with that disposition. Only by a coincidence could it turn out that these beliefs are true, and such a coincidence cannot reasonably be expected.[1]

Some philosophers who regard evolutionary explanations as debunking hold, in addition, that whereas evolutionary considerations provide discrediting explanations for the acceptance of many normative theories, they nonetheless cannot explain why utilitarians accept utilitarianism. Belief in utilitarianism seemingly transcends our evolved biases. Evolutionary considerations are thus thought to tip the balance in favour of utilitarianism,

[1]See Joyce (2006), Ruse (1986), Street (2006). Strictly speaking, Street (2006) argues that natural selection explanations are debunking iff we assume meta-ethical realism.

by selectively debunking its competitors.[2]

The claim that natural selection cannot explain belief in utilitarianism is *prima facie* plausible. Utilitarianism asks us to attach equal value to the well-being of all sentient individuals and act so as to maximally promote the general welfare, so understood. Given its complete impartiality and extreme demandingness, belief in utilitarianism would seem to represent a serious cost to an organism's inclusive fitness. Belief in utilitarianism may therefore be thought to have emerged *in spite of* the selection-pressures shaping human moral psychology.

Our concern in this paper is with the possibility that evolutionary considerations still pose a serious problem for utilitarians, even if we grant that evolutionary considerations favour utilitarianism by selectively debunking its competitors. The problem, highlighted by Kahane (2011, 2014), is as follows. Utilitarianism tells us to do whatever maximizes well-being. This prescription is empty unless we specify the nature of well-being. However, beliefs about well-being are prime candidates for evolutionary debunking. It is easy to see how natural selection would have led us to believe that pleasure is good for us and that pain is bad. It is also easy to see how it could have led us to value desire-satisfaction, or the characteristic ingredients in objective theories of well-being.[3] Since it looks like our beliefs about well-being will be debunked if any evaluative beliefs are, utilitarianism seems to be left without any practical content, even if the utilitarian principle is not itself undermined by evolutionary considerations.

We'll argue that this is not the case. As we'll show, successful debunking arguments targeting our beliefs about well-being do not undermine the practical significance of utilitarianism, provided that we understand the requirements of practical rationality as sensitive to normative uncertainty.

2 Debunking arguments and normative uncertainty

To make our case, we'll begin by clarifying how to conceptualize the damage done by evolutionary debunking arguments.

2.1 What does it mean for a theory to be debunked?

Typically, the notion of debunking is characterized in terms of *categorical belief*: a theory is debunked iff belief in that theory is subject to an (undefeated) defeater.[4] But we could also characterize the notion of debunking in terms of *graded belief*. We would then say that successful debunking arguments require us to (significantly) reduce our credence in various normative theories.

[2]de Lazari-Radek and Singer (2014), Singer (1981, 2005), Greene (2008).
[3]See Crisp (2006: 121-122).
[4]Kahane (2011), Joyce (2006).

Plausibly, a debunking argument never requires us to reduce our confidence in some ethical theory to zero. To assign credence zero to some proposition is to be certain that one could never gain evidence that would raise one's credence above zero. But it would be extreme to suppose that debunking arguments could be so forceful as to render it impossible for any future evidence to support the normative theories we currently believe. Debunking arguments do not salt the earth.

Furthermore, we shouldn't be certain of the soundness of any evolutionary debunking argument. Critics have alleged that these arguments rest on faulty epistemological principles[5] and disputable meta-ethical presuppositions.[6] Thus, even if you're confident that some debunking argument is sound, you ought to assign non-negligible credence to the possibility that it isn't.

2.2 Rational decision-making under normative uncertainty

It's plausible that we should never be completely certain of anything in ethics. Any reasonable person should acknowledge that their values could be mistaken and assign some degree of confidence to a range of ethical views. Since these different views will often diverge in what they tell us to do, we may wonder how we're to decide what to do, given our normative uncertainty. In recent years, a number of philosophers have argued that in cases of normative uncertainty we ought to act so as to *maximize expected choice-worthiness*.[7] This view is analogous to the orthodox decision-theoretic principle of maximizing expected utility.

Here is the basic idea. In a decision-situation, an agent confronts a set of options. The agent's credence function assigns a probability to each member in a finite set of first-order normative theories, corresponding to the agent's confidence in the theory. A theory ranks the agent's options in terms of their choice-worthiness. We assume (for now) that choice-worthiness is interval-scale measurable and intertheoretically comparable. Roughly, this means that each theory tells us how much more (or less) choice-worthy one option is as compared to another and each theory can be represented as ranking the options according to the same scale of choice-worthiness. The expected choice-worthiness of some action is the sum of its choice-worthiness according to each of the theories in the set, weighted according to their probability. The most appropriate option is that which maximizes expected choice-worthiness.

[5] White (2010), Vavova (2014).

[6] Kahane (2011).

[7] Lockhart (2000), MacAskill (2014), Sepielli (2009). For objections see Gustafsson and Torpman (2014), Harman (2014), and Weatherson (2014). Our argument proceeds on the assumption that maximising expected choice-worthiness accounts are at least approximately correct.

Consider a toy example. Suppose S is 70% confident that some form of rights-based deontology is true. According to this theory, it is wrong to intentionally harm one person in order to prevent two others from being harmed in the same way. S assigns the remainder of her confidence to utilitarianism. An evil mastermind offers S the option to electrocute A in order to stop B and C from being electrocuted by the evil mastermind. Alternately, she can refuse and allow$B$ and C to be electrocuted. Her decision-situation might then be represented as follows:

	Deontology 70%	Utilitarianism 30%
Electrocute	5	25
Don't Electrocute	25	5

The numerical values in the cells represent the choice-worthiness scores of the different actions under the two moral theories. The deontological theory ranks *Don't Electrocute* as most choice-worthy. The utilitarian theory ranks *Electrocute* as equally choice-worthy. For simplicity, we assume that utilitarianism ranks *Don't Electrocute* as worse than *Electrocute* to the same extent that the deontological theory ranks *Electrocute* as worse than *Don't Electrocute*. Given these stipulations, the expected choice-worthiness of *Electrocute* is 11 and the expected choice-worthiness of *Don't Electrocute* is 19. Therefore, the most appropriate option in light of S's confidence in the two moral theories is *Don't Electrocute*.

The decision-matrix outlined above assumed that electrocution harms a person, since it causes pain. S might not be totally certain that pain is bad. To take account of this, we might think of S as distributing her credence over four different normative theories, each representing the conjunction of a moral theory and theory of well-being. Assume that S's confidence in utilitarianism remains at 30% and her confidence in deontology at 70%. Suppose, in addition, that she is 99% confident that pain is bad and 1% confident that pain is indifferent. Assuming for simplicity that the probability that pain is bad or indifferent is independent of which moral theory is true, the decision-matrix might then look like this:

	Deontology Pain is bad 69.3%	Utilitarianism Pain is bad 29.7%	Deontology Pain is indifferent 0.7%	Utilitarianism Pain is indifferent 0.3%
Electrocute	5	25	15	15
Don't Electrocute	25	5	15	15

The right-hand side of the decision-matrix looks as it does because we assume that if pain is neutral, then either choice is equally permissible according to either theory. The side-constraint against intentional harm has no force, since A is not harmed by electrocution. And there would be no reason to ensure that a smaller number of people are electrocuted on utilitarianism, since being electrocuted makes no difference to a person's well-being. Whatever S chooses will be equally unobjectionable whichever moral theory happens to be true.

The prescription to maximize expected choice-worthiness still tells S not to electrocute. Its expected choice-worthiness is 18.96, compared to 11.04 for the alternative. Having some slight worry that pain is indifferent makes no difference to what is most appropriate for S to do in this context.

2.3 The significance of debunking arguments

Suppose S becomes aware of a plausible evolutionary debunking argument that considerably reduces her confidence in deontology. Since utilitarianism has always seemed plausible to S apart from the fact that it conflicts with certain entrenched deontological intuitions, she becomes a lot more confident in utilitarianism. Suppose S now assigns 30% confidence to deontology and 70% confidence to utilitarianism. In that case, the expected choice-worthiness of *Electrocute* is 18.96, while the expected choice-worthiness of *Don't Electrocute* is 11.04. In that case, *Electrocute* is the most appropriate choice under normative uncertainty.

What if S is also made aware of a debunking argument targeting her belief that pain is bad? Well, if she loses all confidence in the badness of pain, this would mean that *Electrocute* and *Don't Electrocute* are equal in terms of expected choice-worthiness. In that case, the fact that she is also quite confident that utilitarianism is the correct moral theory would be genuinely irrelevant.

However, we've already ruled out the idea that debunking arguments require us to reduce our confidence to zero. Suppose, more realistically, that S ends up only 30% confident that pain is bad. In that case, the expected choice-worthiness of *Electrocute* is 16.2 and the expected choice-worthiness of *Don't Electrocute* is 13.8. *Electrocute* remains the most appropriate choice.

In fact, it should be straightforward to see that so long as S retains some confidence in the badness of pain, reducing her confidence in this proposition to any arbitrary degree ultimately makes no difference to what would be most appropriate, given her relative confidence in utilitarianism vis-à-vis deontology. If pain is indifferent, then either action is equally choice-worthy no matter which moral theory is true. The normative theories represented in the right hand side of the second decision-matrix in 2.2 make no difference to the relative expected choice-worthiness of the two options. The question of which action is most choice-worthy in expectation is decided entirely by how

S distributes her confidence across those normative theories on which pain is bad, represented in the left-hand side of the decision-matrix. Therefore, so long as her relative confidence in utilitarianism is significantly greater, *Electrocute* remains the most appropriate option.[8]

Therefore, the availability of a debunking argument targeting the belief that pain is bad turns out to be without practical significance. As we recall, the debunking argument targeting S's deontological moral intuitions *did* make a significant difference. In light of that argument, *Electrocute* became the most appropriate choice. And the fact that S is significantly more confident of utilitarianism ensures that this remains so regardless of the extent to which she reduces her confidence that pain is bad, so long as it remains above zero.

3 What follows?

Our discussion has focused on a toy example, constructed using a number of simplifying assumptions. What does this case really tell us about our practical predicament?

3.1 Beyond expected choice-worthiness

The example presumed that the normative theories to which S assigns credence yield choice-worthiness rankings that are interval-scale measurable and intertheoretically comparable. This might seem unrealistic.[9] Where these assumptions don't hold, we cannot act so as to maximize expected choice-worthiness. We have to apply some other rule.

Fortunately, this makes no difference to the key point for which we've argued. On any plausible principle for decision-making under normative uncertainty, the most appropriate option will be determined purely by S's credence in those normative theories that assume the badness of pain. Her credence in those theories that treat pain as indifferent will be irrelevant, since they treat her choice as indifferent. Only those theories that assume pain's badness can tip the balance.

By way of illustration, consider a principle that works for purely ordinal theories: *the Borda rule*.[10] According to the Borda rule, one option is more appropriate than another iff it receives a higher *credence-weighted Borda-score*. An option's Borda-score according to some theory is the number of options to which it is superior, minus the number of options to which it is inferior. Its credence-weighted Borda-score is the sum of its Borda-score under each theory multiplied by one's credence in the theory.

[8] Cf. Ross (2006) on the irrelevance of 'uniform ethical theories' given normative uncertainty.

[9] Gracely (1996), Ross (2006).

[10] MacAskill (2016).

Suppose that deontology and utilitarianism provide only an ordinal ranking of S's options in terms of choice-worthiness. Given the previously stipulated confidence levels assigned by S to deontology, utilitarianism, pain's badness, and pain's indifference, her credence-weighted Borda-score for *Electrocute* is 0.12. For *Don't Electrocute*, it is -0.12. *Electrocute* is still most appropriate.

Furthermore, it's relatively easy to work out that the relative ranking of S's options in terms of their credence-weighted Borda-score is insensitive to her credence in pain's badness vis-à-vis its indifference, in that neither normative theory on which pain is indifferent contributes to the credence-weighted Borda-score of either option. In this respect the Borda rule behaves just like the principle of maximizing expected choice-worthiness. And any other plausible principle should behave similarly.

3.2 Beyond harm

Another respect in which the decision-situation we've considered might be thought unrepresentative is that only the avoidance of harm was assumed to have normative significance.

However, a deontological theory might well posit that a rights-violation occurs when one person electrocutes another without their consent, even if doing so is harmless. In that case, the deontological theory favours *Don't Electrocute* even on the assumption that pain is indifferent. S's choice-situation might then look like this:

	Deontology Pain is bad 9%	Utilitarianism Pain is bad 21%	Deontology Pain is indifferent 21%	Utilitarianism Pain is indifferent 49%
Electrocute	5	25	10	15
Don't Electrocute	25	5	20	15

Here, the expected choice-worthiness of *Electrocute* remains highest. However, this can change if S becomes even more confident that pain is indifferent. Suppose she is 90% confident that pain is indifferent. Then the expected choice-worthiness of *Electrocute* becomes 14.05. The expected choice-worthiness of *Don't Electrocute* becomes 15.95. *Don't Electrocute* would then be most appropriate.

The reason for this should be clear. The utilitarian theory on which pain is indifferent does not tell for or against *Electrocute*. By contrast, the deontological theory on which pain is indifferent tells against. The more confident S becomes that pain is indifferent, the more weight she gives to these theories in deciding what to do. Since the utilitarian theory is indif-

ferent on this point whereas the deontological theory isn't, increasing her confidence that pain is indifferent strengthens her reasons for choosing *Don't Electrocute*.

It doesn't follow that the combined effect of a successful debunking argument targeting S's deontological intuitions and another targeting her belief in the badness of pain will generally leave everything as it was before. This will hold true in some decision-situations, but not in others. Whether things are left unchanged in any given case will be highly sensitive to the confidence S actually assigns to utilitarianism vis-à-vis deontology and to the badness of pain vis-à-vis its indifference. It will also be highly sensitive to the particular choice-worthiness ordering generated by each theory. This is easy to see by tinkering with the credences and rankings we used above. Slight adjustments can easily tip the balance.

It would be an astonishing coincidence if our credences and choice-worthiness rankings were calibrated so that reducing our confidence in deontology and in our beliefs about well-being never made any difference to which option was most appropriate in cases that potentially involve violation of side-constraints. Furthermore, side-constraints are just one point of contention between deontology and utilitarianism. Many of the remaining contrasts are purely a matter of how to weigh harms and benefits befalling different people. For example, deontological theories typically posit *agent-centred permissions*, in light of which each person is entitled to attach added weight to her own well-being. Deontological theories may also posit *irrelevant utilities*: a non-consequentialist might think it is more important to save a single individual from some terrible harm than provide a trivial benefit to each person in an arbitrarily large group of people.[11] The aggregative character of utilitarianism rules out this possibility.

In choice-situations where agent-centred permissions or irrelevant utilities lead deontological theories to issue prescriptions that run against the implications of utilitarianism due to intertheoretic disagreement about the weighting of harms and benefits, reducing one's confidence in deontology will make an important practical difference, whereas reducing one's confidence that one's actions will make any difference to people's well-being will make no difference.

3.3 What about really bizarre views?

A final worry centres on the possibility that debunking arguments require us to increase our credence in bizarre ethical views. For example, we should perhaps increase our credence in the view that pain is intrinsically good for us and pleasure intrinsically bad, as we can be confident that this view would not have been selected for. But we have so far ignored this possibility.

[11] Kamm (1993).

In a similar vein, Kahane (2014) notes that certain highly counter-intuitive beliefs about well-being will resist evolutionary explanation: "These would include the views that the good life consists of ascetic contemplation of deep philosophical truths, or celibate spiritual communion with God, or a kind of Nietzschean perfectionist aestheticism (which might even revel in pain), and so forth." (334) In combination with such theories, he notes, utilitarianism might retain its practical significance. However, its implications would be utterly repugnant: few people would be able to accept these implications. Is our argument vulnerable to this sort of worry? Does the ability of bizarre moral views to escape debunking mean that they are likely to end up playing a substantial role in determining what is most appropriate in light of our normative uncertainty?

That would be the case if evolutionary debunking arguments pushed our confidence in common-sense views about well-being down so far that it was not appreciably higher than our confidence in these wildly counter-intuitive theories. We could end up in this position if debunking arguments required us to reduce our confidence in common-sense intuitions very close to zero. But the effect of encountering these arguments will not be so catastrophic. Debunking arguments may seem convincing, but it's far from certain that they're sound. For this reason, we ought to retain significant credence in common-sense views about well-being of which we were extremely confident prior to encountering these arguments. In the examples we considered earlier, we set S's posterior credence in pain's badness at 30% or 10%. Given S's antecedent confidence and the controversy surrounding the soundness of debunking arguments, even this might be too low.

If she is like the authors, S would have assigned a much, much lower prior probability to the view that pain is good or that celibate spiritual communion with God is the key determinant of well-being. Her posterior confidence in common-sense views could therefore be orders of magnitude greater than her credence in wildly counter-intuitive theories of this kind. The practical significance of these views would therefore be negligible.[12]

Of course, this would *not* be the case if her confidence in these counter-intuitive theories should increase significantly upon encountering debunking arguments. That would be the case if one of these theories of well-being was like utilitarianism in that it seems plausible *apart* from the fact that it conflicts with certain entrenched common-sense intuitions that now get

[12]For the view that pain is good and pleasure bad, there is a further argument discounting its practical significance. When combined with utilitarianism, this view has exactly opposite recommendations to classical utilitarianism. Therefore, under normative uncertainty this theory simply 'cancels out' part of one's credence in classical utilitarianism. For example, with 60% credence in deontology, 38% credence in classical utilitarianism, and 2% credence in pain-is-good utilitarianism, a rational decision-maker will take the same actions as if she had 60% credence in deontology, 36% credence in classical utilitarianism, and 4% credence in a view that was indifferent between all options.

debunked, provided that the plausibility of the theory itself remains intact in the face of debunking arguments.

However, the theories considered here don't seem to fit that description. The view that pain is intrinsically good is not the sort of view that seems somewhat plausible, except for the fact that it conflicts with intuition. As we see it, it has basically zero inherent plausibility. The view that the good life is centred on celibacy, meditation, and prayer strikes us as false principally because it attaches value to things which seem valueless owing to our confidence that God does not exist. Debunking arguments will not change that fact.[13] We are more attracted to the view that contemplation of philosophical truths or the realization of aesthetic value can be intrinsic sources of well-being. Theories that count such goods as the primary or only determinants of well-being seem weird to us principally because they attach too little value to other things, such as pleasure or desire-satisfaction.

Nonetheless, these theories do not fit the criterion we specified above. To the extent that such theories have plausibility in light of the intuitive value of knowledge and aesthetic excellence, they will lose plausibility in the face of debunking arguments. After all, it is easy to see why natural selection should lead human beings to value knowledge: we are informavores by design.[14] There is also good reason to expect that natural selection has played a significant role in shaping our aesthetic responses.[15]

Perhaps there are other theories of well-being that fit the criterion, but we have not been able to think of any. Until suitable candidates are proposed, we feel entitled to presume that there is no significant objection to our argument arising from the possibility that radically counter-intuitive theories of well-being escape evolutionary debunking.

4 Conclusion

Assuming that we ought to take normative uncertainty into account, debunking arguments that selectively undermine non-utilitarian theories have genuine practical significance, even if we're also aware of debunking explanations targeting our beliefs about well-being. The latter do not rob utilitarianism of its practical significance. Given the resulting credence-distribution over different moral theories and theories of well-being, the most appropriate action will in many cases accord with the action required by utilitarianism in combination with common-sense theories about well-being.

[13] Except perhaps to increase our confidence in atheism: see Wilkins and Griffiths (2013).
[14] Dennett (1991: 176-182).
[15] Dutton (2009).

References

Crisp, Roger (2006) *Reasons and the good*. Oxford: Oxford University Press.

de Lazari-Radek, Katarzyna and Singer, Peter (2014) *The point of view of the Universe: Sidgwick and contemporary ethics*. Oxford: Oxford University Press.

Dennett, Daniel (1991) *Consciousness explained*. London: Penguin.

Dutton, Denis (2009) *The art instinct: beauty, pleasure, and human evolution*. Oxford: Oxford University Press.

Gracely, Edward J. (1996) On the noncomparability of judgments made by different ethical theories. *Metaphilosophy* 27, 32732.

Greene, Joshua (2008) The secret joke of Kant's soul. In Walter Sinnott-Armstrong, ed. (2008) *Moral psychology, vol. 3: the neuroscience of morality*, 35-80. Cambridge, MA: MIT Press.

Gustafsson, Johan, and Torpman, Tom (2014) In defence of my favourite theory. *Pacific Philosophical Quarterly* 95, 159-174.

Harman, Elizabeth (2014) The irrelevance of moral uncertainty. *Oxford Studies in Metaethics* 10, 53-79.

Huemer, Michael (2008) Revisionary intuitionism. *Social Philosophy and Policy* 25, 368-392.

Joyce, Richard (2006) *The evolution of morality*. Cambridge, MA: MIT Press.

Kahane, Guy (2011) Evolutionary debunking arguments. *Noûs* 45, 103-125.

Kahane, Guy (2014) (2014) Evolution and impartiality. *Ethics* 124, 327-341.

Kamm, Frances M. (1993) *Morality, mortality, volume 1: death and whom to save from it*. Oxford: Oxford University Press.

Lockhart, Ted (2000) *Moral uncertainty and its consequences*. Oxford: Oxford University Press.

MacAskill, William (2014) *Normative Uncertainty*. D. Phil. thesis. University of Oxford.

MacAskill, William (2016) Normative uncertainty as a voting problem. *Mind* 125, 967-1004.

Ross, Jacob (2006) Rejecting ethical deflationism. *Ethics* 116, 742-768.

Ruse, Michael (1986) *Taking Darwin seriously: a naturalistic approach to philosophy*. Oxford: Blackwell.

Sepielli, Andrew (2009) What to do when you dont know what to do. *Oxford Studies in Metaethics* 4, 5-28.

Singer (1981) *The expanding circle: ethics and sociobiology*. Oxford: Clarendon Press.

Singer, Peter (2005) Ethics and intuitions. *Journal of Ethics* 9, 331 - 352.

Street, Sharon (2006) A Darwinian dilemma for realist theories of value. *Philosophical Studies* 127, 109-166.

Vavova, Katia (2014) Debunking evolutionary debunking. *Oxford Studies in Metaethics* 9, 76-101.

Weatherson, Brian (2014) Running risks morally. *Philosophical Studies* 167, 14163.

White, Roger (2010) You just believe that because... *Philosophical Perspectives* 24, 573-615.

Wilkins, John S. and Griffiths, Paul E. (2013) Evolutionary debunking arguments in three domains: fact, value, and religion. In Greg Dawes and James Maclaurin, eds. *A new science of religion*, 133-146. London: Routledge.