

Uncertainty and Utilitarianism

Krister Bykvist

Guest essays represent only the views of the author(s).

Table of Contents

1. Introduction
2. Moral Uncertainty Has No Implications For Utilitarian-Friendly Agents
3. Taking Moral Uncertainty Seriously
4. Applications and Trade-Offs
5. Implications For Some Traditional Objections to Utilitarianism
6. Is Normative Ethics Made Obsolete?
7. Resources and Further Reading

Introduction

1. How important is human well-being compared to that of non-human animals?
2. How much should we spend on helping strangers in need?
3. How much should we care about future generations?

Few could honestly say that they are fully certain about the answers to these pressing moral questions, and this holds for utilitarians and non-utilitarians

alike. We feel less than fully certain about the answers partly due to uncertainty about *empirical* facts. We are uncertain about whether shrimps can feel pain, whether we can really help strangers far away, or whether we can make sure people in the far future have good lives.

But sometimes the uncertainty is fundamentally *moral*.¹ Even if we knew all the relevant empirical facts, we could still waver about whether it is right to kill an animal for a small benefit to a human, whether we have strong duties to help strangers in need, and whether future people matter as much as current ones. Fundamental moral uncertainty can also be more general, as when we are uncertain about whether a certain moral theory is correct. Many first-year students express deep uncertainty about which moral theory is correct after taking an introductory course in normative ethics, where all the standard theories are exposed with ‘warts and all’.

In this essay, I am going to assess what implications, if any, moral uncertainty has for utilitarian-friendly agents—agents who have non-negligible confidence in utilitarianism or in some of its central claims.

Moral Uncertainty Has No Implications For Utilitarian-Friendly Agents

The first possible answer is that it will have no implications. There are three main possible reasons for this. First, one could argue that there is no point in asking what one ought to do when one is uncertain about what one ought to do, for the answer is trivial: ‘you ought to do what you ought to do, no matter whether or not you are certain about it’.² So, if utilitarianism is the correct theory and it says that you ought to do something, then this is what you ought to do, no matter whether you have doubts about it or not.

This hard-nosed answer is not very attractive, however. First, suppose that you are in a restaurant that offers both meat and vegetarian options. You are considering only two moral views (which is to simplify considerably): common-sense morality, according to which it is permissible to eat factory-

farmed meat and permissible to eat vegetables, and vegetarianism, according to which it is impermissible to eat factory-farmed meat and permissible to eat vegetables. We can assume here that vegetarianism is based on [utilitarian reasoning about the importance of animal well-being](#). While the option of eating vegetables will not risk doing any moral wrongs, the option of eating meat will do so. No matter which moral theory is true, by eating vegetables you will act permissibly, whereas if you eat meat there is a chance that you are acting wrongly. Given the information available to you, it seems reasonable to hedge here and choose the risk-free option of eating vegetables, and this holds even if vegetarianism turns out to be false in this particular case.

The second possible reason why moral uncertainty will not have any implications for utilitarian-friendly agents, is the idea that the reasonable thing to do under moral uncertainty is to act according to the theory you have most confidence in. This is called the ‘my favourite theory’ account of acting under moral uncertainty. Now, if you are a utilitarian-friendly agent, then utilitarianism is the theory you have most confidence in. Hence, this is the one you should follow.

However, this assumes that utilitarian-friendly agents must have most confidence in utilitarianism. But they need not; they only need to have a non-negligible confidence in utilitarianism. Furthermore, the ‘my favourite theory’ approach is not always helpful for true utilitarians who do have most confidence in utilitarianism, for they can be uncertain about which version of utilitarianism is true. For example, one can wonder whether it is [rule utilitarianism or act utilitarianism](#), or whether it is [hedonist or preferentialist utilitarianism](#). If a true utilitarian is equally uncertain about the different disagreeing versions of utilitarianism, telling them to act on the version that they have most confidence in is not helpful. Finally, suppose that in the restaurant case above you have slightly more confidence in common-sense morality, then the ‘my favourite theory’ approach implies that you should eat

meat despite the fact that you thereby risk doing something wrong when you could have avoided this by eating the vegetarian option.³

There is a final possible reason why moral uncertainty might not have any implications for utilitarian-friendly agents. Even if the question ‘what should I do when I am not sure what I morally ought to do’ does not have a trivial answer, perhaps we can never compare the values different theories assign to our options. If we can never compare them, how are we going to account for them in our deliberations? But this rhetorical question has a partial answer, for incomparability need not always paralyze decision-making. Take for example, the restaurant case and assume that we cannot compare the values that common-sense morality and vegetarianism assign to the meat and vegetarian options, respectively. This does not change the reasonableness of going for the vegetarian option since it dominates the meat option: choosing the vegetarian option is equally as good as the meat option, according to common-sense morality, and better, according to vegetarianism. We do not need to compare values across the two theories to establish this dominance.

Furthermore, to say that we can never compare values across theories seems too extreme. It is true that it seems impossible to compare values across very different theories, like utilitarianism and Kantianism since they use radically different notions of value. But utilitarian-friendly agents are sometimes only uncertain about which [theory of well-being](#) is correct. They are not uncertain about the notion of morally relevant value; it is goodness for individuals. Instead, they are uncertain about what is good for individuals. Is it pleasure or preference satisfaction, for example? Surely, it makes sense to say that hedonism assigns higher value to pleasure than preference utilitarianism does since hedonism says it is good while preference utilitarianism says it is only neutral.

The uncertainty can be even more limited. The utilitarian-friendly agent can be certain that hedonism about well-being is true, but uncertain about which form of hedonism is true. For example, should so-called higher pleasures (e.g.,

pleasures of intellectual endeavours, art, music, and poetry) be assigned more value for individuals than so-called lower pleasures (e.g., sensory pleasures from eating food, drinking, or having a bath), as [John Stuart Mill](#) thought? Here, again, the uncertainty is not about what notion of value is relevant (still goodness for individuals), and it is not even about what has value (pleasant experiences). The uncertainty is just about the relative value of higher and lower pleasures for individuals.⁴

It is worth noting that if we can make comparisons of value differences between options across theories, then the ‘my favourite theory’ account is in deep trouble. You are required to follow your favorite theory, even if you only have slightly more confidence in it than in an alternative theory according to which more is at stake.

Taking Moral Uncertainty Seriously

If we take moral uncertainty seriously and accept that it has implications for utilitarian-friendly agents, we need to clarify what ‘reasonable choice’ means for a morally uncertain agent. It cannot mean the choice that utilitarianism favors, because then we are back to the hard-nosed view according to which ‘we ought to do what we ought to do’ under moral uncertainty.

One option is to invoke rationality as understood in standard decision-theory. On this view, when we talk about what is reasonable to do under moral uncertainty we are talking about what would be rational to do given a morally conscientious agent’s preferences and beliefs. A morally conscientious agent will care about the values of her actions under different plausible moral hypotheses, and her preferences will line up with their possible assignments of value. To make this clear, suppose the conscientious agent can either go to the museum or to the spa. She considers two moral hypotheses, Mill’s qualitative hedonism and [quantitative hedonism](#). The value assignments look like this:

	Mill's qualitative hedonism	Quantitative hedonism
Go to museum and enjoy a higher pleasure	10 (A)	4 (B)
Go to spa and enjoy a slightly more intense lower pleasure	2 (C)	5 (D)

Then the conscientious agent's preferences will mirror the values. She will prefer museum to spa while Mill's qualitative hedonism is correct, i.e., prefer A to C. She will also prefer spa to museum while quantitative hedonism is correct, i.e., prefer D to B.

But the conscientious agent's preferences also line up with the comparisons of value differences across theories, when such comparisons are possible (which we assume is the case in this case). Since the value difference between the options is greater according to qualitative hedonism than it is according to quantitative hedonism, the agent's preference for museum over spa while qualitative hedonism is correct (her preference for A over C) will be stronger than her preference for spa over museum while quantitative hedonism is correct (her preference for D over B).

Now, there is a question about how the 'ought' of rationality relates to the 'ought' of morality. One option is to say that the rational ought coincides with a subjective moral ought that differs from the objective moral ought. On this view, what you objectively ought to do is not sensitive to your beliefs or evidence about your choice situation. This is the ought of *objective utilitarianism*, according to which you ought to bring about the best outcome, no matter your beliefs and evidence. The subjective ought, in contrast, takes into account your beliefs and evidence about your choice situation, including your beliefs and evidence about what you objectively ought to do. When you are

uncertain about what you morally ought to do in the objective sense, there can still be an answer about what you ought to do in a subjective sense. This answer takes into account your degrees of confidence in different hypotheses about what you ought to do and what has value in an objective sense.

Another option is to distinguish morality from rationality and say that when you are uncertain about what you morally ought to do there can still be an answer about what you rationally ought to do, but there is no further question about what you morally ought to do, not even subjectively.⁵

I am not going to take a stand on which option to choose since it does not matter for my points below.⁶

Applications and Trade-Offs

Let us now turn to some applications. We have already seen how hedging can help with the restaurant case above by avoiding the risk of wrong-doing. There are other structurally similar cases: where one action is guaranteed to be permissible, no matter which moral hypothesis is true, while the alternative action risks wrong-doing. Take, for example, charitable donations. Suppose that you have some confidence in the utilitarian-friendly view that we are morally required to [donate to effective charities](#), but also some confidence in the laissez-faire view that allows you to donate but also to abstain. Then hedging is again applicable and the resulting verdict is that you should donate.

Hedging will only take you so far, however. Not all realistic cases have the structure that is necessary for hedging to be applicable. Imagine, for instance, a version of the restaurant case in which you really hate the vegetarian option on the menu and you consider ethical egoism that gives a lot of weight to your own well-being, even in cases where only your gustatory preferences are at stake. Then we no longer have the structure necessary for hedging to apply, for one of the considered moral views now says it is wrong to order the vegetarian option, because of your strong aversion. Similarly, imagine a version of the donation case, in which your donation would come at a significant cost to

yourself. Under the above common-sense view, we no longer have a case suitable for hedging since donating is wrong according to one of the views you are considering.

What shall you do when hedging does not apply? It is clear that you will have to make some trade-offs between the moral views you are considering, based on your degrees of confidence in them and the degrees of value they assign to your options. It is also clear that if your confidence splits evenly across two moral hypotheses, then the hypothesis that sees a greater difference between the options calls the shot. This means that if you are equally certain in the considered hypotheses in the new versions of the restaurant case and the donation case, then the moral hypothesis that sees the greatest difference between the options wins. So, if vegetarianism sees a greater difference between eating vegetarian and eating meat than common-sense does, then you should eat vegetarian. Similarly, if the donation-friendly view sees a greater difference between donating and not donating than common-sense does, then you should donate. Of course, there is no guarantee that these value differences hold between the options, but they seem plausible at least for the donation case, where seemingly much more is at stake for the utilitarian-friendly view than for the common-sense view: saving lives versus bearing a significant but not very severe cost to oneself.

Of course, your confidence does not split evenly in all cases. In most cases, you have some significant confidence in several views, but you have more confidence in some views than in others. Cases where you are more confident in the view that sees the greatest difference between the options pose no difficulty, since here it is obvious that this view rules. For example, if, as a true utilitarian, you are more convinced in the donation-friendly view that also sees the greater value difference between donation and non-donation, then you should donate.

It is more controversial what you should do when the confidence and the value difference do not line up nicely in this way. Here we can follow the lead of what

is reasonable to do under empirical uncertainty.⁷

Suppose Julia considers whether to speed around a blind corner. She thinks it is unlikely that anyone is crossing the road immediately around the corner, but she is not sure. If she speeds and hits someone, she will certainly severely injure them. If she goes slowly, she certainly will not injure anyone, but will get to work slightly later. Here the right option is clearly not to speed, even though it is unlikely that someone will cross the road. The situation has the following structure (the numbers are only illustrative):

	Someone will cross (Probability: 0.2)	No one will cross (Probability: 0.8)
Speed	-20	15
Not speed	10	10

Often decisions in such cases are justified by weighing the probabilities of the possible outcomes against their values by calculating the *expected value*:

Speed: $(-20 \times 0.2) + (15 \times 0.8) = 8$

Not speed: $(10 \times 0.2) + (10 \times 0.8) = 10$

In this case, Julia should not speed because it has a higher expected value than speeding.

To maximize expected value in all situations is controversial. For example, assume that you have minuscule confidence (to the degree of 0.00000000000000000001, say) in some outlandish empirical view: that you cause an enormous disaster if you brush your teeth (perhaps [chaos theory](#) can support this possibility). Then not brushing your teeth may still be what maximizes expected value, even if you are much more confident that if you do brush your teeth nothing bad will happen and that some good will happen—

your dental hygiene will be improved slightly. It is enough that the possible but extremely unlikely disaster is sufficiently bad.⁸

But one need not adopt an unrestricted expected value account to agree that expected value gets the right answer in some important cases, like the speeding case above. It is just that the full justification will have to be more complex.

Now, if you think the trade-off is right in the speeding case and other structurally similar cases of *empirical* uncertainty, then it is also plausibly right in structurally similar cases of *moral* uncertainty, where the hypotheses are moral and the values are moral values assigned by the different hypotheses. This analogy seems plausible since with both empirical and moral uncertainty we are invoking the notion of a rational action given one's preferences and beliefs. It is just that the contents of those preferences and beliefs differ between the empirical and the moral case.

However, this plausible-sounding trade off assumption has quite radical implications for what we should do under moral uncertainty. We have already seen that in the donation cases, the reasonable option is to donate, even if you are much more certain that you are not required to do so. So, even agents who are much more inclined to accept non-utilitarianism will have to follow utilitarianism here.

For another example, take animal ethics. Suppose that you have some confidence in the view that the well-being of animals matters as much as human well-being—a claim that is part and parcel of utilitarianism—but also some confidence that human well-being matters more. Now consider any action that involves a great sacrifice for animals but leads to a slight benefit to humans. An example could be very painful animal testing involving lots of animals to cure some minor illness for a small group of humans. Suppose further that you think it is much more likely that humans matter more, but you still think it somewhat likely that animals matter equally. Then the fact that

much more is at stake according to the animal equality view yields the result that we should avoid sacrificing animals for the sake of slight benefits for humans, if sufficiently many animals are severely harmed by the testing. And this holds even for agents who are much more confident in the non-utilitarian view according to which humans matter more.

The implications of the trade-off principle need not always be clearly utilitarian, however. Consider the ethics of abortion. Suppose that you are twenty weeks into your pregnancy and you are considering having an abortion. You think it is unlikely that twenty-week-old fetuses have a right to life, but you are not sure. If you have an abortion and twenty-week-old fetuses do have a right to life—a distinctly non-utilitarian notion—then your action is seriously bad morally. If you have the child and give it up for adoption, you will avoid this bad action, for the child will have a happy life, though you will bear considerable costs as a result of pregnancy, childbirth, and separation from your child. In this case, the cost to the decision-maker is considerable, but the potential badness of abortion is much greater. So, again, it seems that even if you are fairly confident in the view that fetuses have no right to life, as long as you are not extremely confident, the risk of fetuses having such a right is sufficient to outweigh the possible reason in favor of having the abortion. In which case, the reasonable option for you is to give the child up for adoption, something which not all utilitarians would accept.⁹

Implications For Some Traditional Objections to Utilitarianism

The approach to moral uncertainty sketched above lessens the force of some traditional [objections to utilitarianism](#).

One common objection to utilitarianism is that it is [too impartial](#). You are not allowed to give extra weight to your family and friends just because they are your near and dear. However, according to the approach to moral uncertainty, you should give some extra weight to the interests of your family and friends,

even if you are a confident utilitarian. For even if you are confident that the well-being of your family and friends are equally as important as the well-being of distant strangers, you should not be certain in that view: you should have some confidence in the partial view. However, you should have almost no credence that the well-being of distant strangers is *more* important than the well-being of your family and friends. So you should therefore give the interests of your family and friends some extra weight, though not as much weight as if you were completely convinced of the partial view. If you could benefit your friend or a stranger by the same amount, it is therefore more appropriate to benefit your friend over the stranger.

Another common objection to utilitarianism is that it [does not take into account equality](#). On the present account of moral uncertainty, this is no longer true if you have some confidence in [egalitarianism](#), according to which inequality is in itself bad. For then you need to give equality some weight even if you are almost convinced that utilitarianism is correct. Furthermore, if the badness of inequality is much greater than the difference in overall well-being, you should go for an equal distribution of well-being, even if you are slightly more convinced in utilitarianism.

A final objection to utilitarianism is that it is too permissive because it rejects any [moral constraints](#) on actions, such as ‘Don’t lie!’, ‘Keep your promises!’, ‘Don’t kill innocent persons!’, or ‘Don’t torture!’. You are allowed to violate any of these constraints, if this is the only way for you to promote overall well-being. On the current account of moral uncertainty, even utilitarian agents need to be more cautious when possible constraint-violations are at stake. To justify a violation it is not enough to show that it will lead to more overall well-being. If you have some confidence in moral constraints, which seems reasonable, you need to take into account the possible badness of such a violation, even if you are almost convinced that utilitarianism is correct. If the degree of badness is much greater than the difference in overall well-being, you are not allowed to violate the constraint, even if you are slightly more convinced in utilitarianism. Of course, this does not show that you are *never*

allowed to violate constraints, but such absolutism is a very controversial view even among non-utilitarians.

Is Normative Ethics Made Obsolete?

If the approach to moral uncertainty sketched above works, it might look like normative ethics has been made obsolete in many domains, including vegetarianism, charitable donations, animal testing, and abortion. We seemingly get robust and clear verdicts about what to do even though we have not decided on which moral theory is true. It might appear that much philosophical discussion in normative ethics is for practical (i.e., decision-making) purposes unnecessary.¹⁰

This would be to overstate the arguments' conclusions, however. First of all, normative ethics still plays a crucial role in determining how confident we should be in different moral views. For it is in normative ethics that our moral views are systematically tested in terms of, among other things, their coherence, simplicity, and ability to plausibly explain our considered moral judgments about particular cases. The better a view performs on these tests, the more confidence we should have in it.

Second, some of the verdicts above rely on controversial value comparisons across moral views. For example, can we really compare value between a view that assigns fetuses a right to life and one that does not? If we cannot, then it is no longer clear that the verdict for the abortion case above follows. Here we need a philosophical discussion about when and how to make comparisons of value across moral views, and some ideas about what to do when such comparisons cannot be made.¹¹

Finally, in the cases above only a limited number of moral hypotheses were entertained by the agent. If more hypotheses are added, the verdicts might change. For example, in the restaurant case, if you consider the total utilitarian view—that you ought to maximize total well-being, even if this is achieved by adding lives that are barely worth living—then eating meat *might* be the

reasonable option. Eating meat indirectly supports meat production that will renew livestock populations by creating new animals with lives that could plausibly be barely worth living.¹² Similarly, if the agent in the abortion case also considers views on which creating a happy person can be bad, then adoption might no longer be the reasonable option. Reasons for why creating a happy person might be bad includes the idea that only happy lives above a certain positive level of happiness make things better (the so-called [critical level principle](#)). Again, we need normative ethics to help us decide which moral views to have a non-negligible degree of confidence in.

It is still true, of course, that the approach to moral uncertainty sketched here does provide guidance for agents who are morally uncertain. There is no need to first decide which view is true to make reasonable decisions. But that should come as a relief.¹³

About the Author



[Krister Bykvist](#) is a Professor of Practical Philosophy at the Department of Philosophy, Stockholm University and the Institute for Futures Studies. Before that, he was a Tutorial Fellow at Jesus College, Oxford, and a CUF Lecturer in Philosophy at Oxford University (2001-2013). His research concerns questions about our responsibility to future generations, the foundations of consequentialism, moral and evaluative uncertainty, and the relationship between preferences, value, and welfare.

How to Cite This Page

```
Bykvist, K. (2023). Uncertainty and Utilitarianism. In R.Y. Chappell, D. Meissner, and W. MacAskill (eds.), An Introduction to Utilitarianism, <https://www.utilitarianism.net/guest-essays/uncertainty-and-utilitarianism>, accessed 8/12/2025.
```

Resources and Further Reading

- Bykvist, Krister (2022). Evaluative uncertainty and population ethics. In G. Arrhenius, K. Bykvist, T. Campbell, & E. Finneron-Burns (Eds.), *Handbook on Population Ethics*. Oxford University Press.
- Bykvist, Krister (2017). [Moral uncertainty](#). *Philosophy Compass*.
- Bykvist, Krister (2014). Evaluative uncertainty and consequentialist environmental ethics. In L. Kahn, & A. Hiller (Eds.), *Environmental ethics and consequentialism*. Routledge.
- Graham, Peter (2010). [In defence of objectivism about moral obligation](#). *Ethics*, 121(1): 88–115.
- Gustafsson, Johan & Torpman, Olle (2014). [In defence of my favourite theory](#). *Pacific Philosophical Quarterly*, 95(2): 159–174.
- Hájek, Alan (2003). Waging War on Pascal's Wager. *The Philosophical Review*, 112(1): 27–56.
- Lockhart, Ted (2000). *Moral uncertainty and its consequences*. Oxford University Press.
- MacAskill, W., Bykvist K., Ord, Toby (2020). [Moral uncertainty](#). Oxford University Press.
- Ross, Jacob (2006). [Rejecting ethical deflationism](#). *Ethics*, 116: 742–768.
- Sepielli, Andrew (2014). [What to Do When You Don't Know What to Do When You Don't Know What to Do...](#) *Noûs*, 48(3): 521–44.
- Sepielli, Andrew (2009). What to do when you don't know what to do?. In R. Shafer-Landau (Ed.), *Oxford studies in metaethics: 4*. Oxford University Press.
- Weatherson, Brian (2014). [Running risks morally](#). *Philosophical Studies*, 167(1): 141–163.
- Zimmerman, Michael J. (2014). *Ignorance and moral obligation*. Oxford University Press.

- Zimmerman, Michael J. (2008). *Living with uncertainty: The moral significance of ignorance*. Cambridge University Press.
-

1. For book-length treatments of fundamental moral uncertainty, see Lockhart (2000) and MacAskill et al (2020). ↩
2. This view is defended in Weatherson (2014). ↩
3. Gustafsson & Torpman (2014) defend a sophisticated version of the ‘my favourite theory’ approach that avoids some of the problems mentioned here, but it faces other issues. For more on this, see MacAskill et al (2020), pp. 39–44. ↩
4. Of course, much more needs to be said about how these comparisons are to be made. For a thorough discussion of the pros and cons of different accounts, see MacAskill et al (2020), ch. 5. See also Ross (2006), and Sepielli (2009). ↩
5. This kind of view is defended in Bykvist (2014) and Graham (2010). ↩
6. A third option is to stick to one notion of moral ought but define it in terms of what it would be rational for a conscientious agent to do under uncertainty about the *values* of options. A fully worked out theory of this kind is presented and defended in Zimmerman (2008), (2014). My points below do not depend on whether this is the right approach either. ↩
7. Here and in the following I assume that the agent is not in doubt about the (incomplete) rationality principles in question, at least not for the cases I shall discuss, but that is not always true. What to do when one is uncertain about which action is rational under moral uncertainty opens yet another can of worms. This second-order uncertainty has been discussed in the literature on moral uncertainty. For some proposals

about how to handle it, see, for instance, Sepielli (2014), and MacAskill et al (2020), pp. 30–33. ↩

8. This conclusion does not follow if there is also a minuscule probability that the you will cause some fantastically good outcome, for then the risks will cancel each other out. For more on the problem of fanaticism for the case of empirical uncertainty, see Hájek (2003). For more on the corresponding fanaticism problem for the case of moral uncertainty, see MacAskill et al (2020), ch.6. ↩

9. For more applications, see MacAskill et al (2020), ch.8. See also Bykvist (2022) that discusses evaluative uncertainty in population ethics. ↩

10. This is what Ted Lockhart believes, one of the first philosophers to give a book-length discussion of moral uncertainty. See Lockhardt (2000), p. 52. ↩

11. See MacAskill et al (2020), chs. 3 and 4, for some ideas about what to do when one cannot make inter-theoretical comparisons. ↩

12. Here I assume that the meat is not factory-farmed. Conditions in [factory farms](#) are typically so miserable that most lives of farmed animals are likely not worth living—these animals would have been better off had they not been born. Under such wretched conditions, it is morally bad to create additional lives of misery and suffering. ↩

13. I would like to thank Richard Yetter Chappell and Darius Meißner for incisive and helpful comments on earlier drafts of this essay. ↩