Esoteric Altruism:
Does effective altruism require its own destruction?

Jeff Sebo

Abstract

A major difference between effective altruism and utilitarianism, on standard interpretations, is that effective altruism is a practical project, not a moral theory. Effective altruists note that this status as a practical project makes effective altruism less vulnerable to some objections to utilitarianism, such as the objection that utilitarianism is too demanding. However, this status might also make effective altruism more vulnerable to other objections to utilitarianism, such as the objection that utilitarianism is an "esoteric" moral theory that implies that nobody should accept it in practice. Plausibly, a moral theory can have this feature and still be correct. Can a practical project have this feature and still be correct? If so, why? If not, then a lot might depend on whether or not effective altruism is, in fact, esoteric in this sense. This essay examines how utilitarians reply to this objection and how effective altruists might be able to reply to it.

1. Introduction

According to common interpretations of utilitarianism and effective altruism (which, to be clear, not everyone endorses), they overlap partly but not entirely. On one hand, they have at least somewhat similar statuses, contents, and communities. They both involve a commitment to do the most good possible by maximizing positive welfare and/or minimizing negative welfare in the world, all else equal. And many utilitarians are effective altruists, and vice versa. But they also have at least somewhat different statuses, contents, and communities: Utilitarianism is a moral theory whereas effective altruism is a practical project. Utilitarianism is a *monist* theory meant to govern all projects, whereas effective altruism is a *pluralist* project meant to exist alongside other projects. And not all utilitarians are effective altruists, and vice versa.

Since utilitarianism and effective altruism are at least somewhat similar (and since many people see them as more similar than they are), they face several of the same objections. For example, utilitarianism and effective altruism both face a demandingness objection, according to which they demand that we *do* too much. They both face a cluelessness objection, according to which they demand that we *know* too much. They both face an injustice objection, according to which they permit or require unjust actions, such as sacrificing the few for the sake of the many. And they both face an esoterica objection, according to which they imply that nobody should accept or promote them. Utilitarians have spent the past two hundred years addressing these objections, and effective altruists are now addressing them as well.

But since utilitarianism and effective altruism are at least somewhat different, each one is more vulnerable to some objections than to others. For instance, effective altruism might be less vulnerable to the demandingness and injustice objections than utilitarianism. Since effective altruism is a pluralist practical project, it requires us to respect rights, cultivate virtuous characters, and cultivate caring relationships, and it permits us to create space in life for other, independent projects, relationships, and moral commitments. But effective altruism might also be more vulnerable to the cluelessness and esoterica objections. Plausibly, a moral theory can be correct even if nobody can apply it, and even if nobody should accept or promote it. Can a practical project be correct when it has these features?

This essay examines how effective altruists can address the esoterica objection. I start by introducing utilitarianism and effective altruism and explaining the similarities and differences between them according to common interpretations. I then survey several objections that apply to both, focusing on the demandingness, cluelessness, injustice, and esoterica objections. I then focus on the esoterica objection, explaining why utilitarianism can rely on its status and content as a monist moral theory when addressing this objection and why effective altruism might *not* be able to rely on its status as a pluralistic practical project in the same kind of way. I then discuss why effective altruism might be esoteric, why it matters whether effective altruism is esoteric, and whether effective altruism is, in fact, esoteric.

Of course, the question whether effective altruism is esoteric is an empirical question, and I will not be able to fully answer it here. Instead, I want to explain and motivate this objection and offer my own hypothesis about how effective altruists can address it. My hypothesis is, roughly, that effective altruism is *partly* esoteric. Doing the most good possible requires thinking indirectly, in the sense of using this ultimate aim to select other, proximate aims, and then focusing on those proximate aims in everyday life. It can also require deception or self-deception in some cases. But many projects are partly esoteric in these senses, and being partly esoteric in these senses is not particularly problematic. However, questions remain about whether effective altruism is *also* esoteric in stronger and more problematic senses.

2. Utilitarianism and effective altruism

In general, assessing whether and to what extent utilitarianism and effective altruism face the same objections requires assessing their similarities and differences. These frameworks have several striking similarities, and so they are vulnerable to several of the same objections. But they also have several striking differences, and so each is more vulnerable to some of these objections than to others. Of course, we might disagree about how to interpret utilitarianism and effective altruism, and so we might also disagree about where these frameworks converge, diverge, and share a common fate as a result. I will here assess these frameworks according to common interpretations that strike me as plausible, while noting that if our interpretations of these frameworks vary, then our assessments of them might vary as well.

On one hand, utilitarianism is, on one common interpretation, a moral theory, that is, a conception of what makes right actions right in principle. In its classical form, it accepts a hedonistic theory of the good, according to which the only intrinsically good thing is positive welfare (pleasure or desire-satisfaction) and the only intrinsically bad thing is negative welfare (pain or desire-frustration). It also accepts an impartially benevolent theory of the right, according to which an action is right if and only if it maximizes positive welfare and minimizes negative welfare for all sentient beings from now until the end of time. Bentham, Mill, Sidgwick, and other utilitarian thinkers have all endorsed these basic ideas (Bentham 1789/2007, Mill 1861/1998, Sidgwick 1907/1981; for general discussion, see Driver 2012).

On the other hand, effective altruism is, at least on one common interpretation, a practical project, that is, an activity that an individual or group can pursue in practice. In its standard form, it involves an aspiration to use evidence and reason to do the most good possible, all else equal. More specifically, effective altruists attempt to maximize positive welfare and/or minimize negative welfare for all sentient beings from now until the end of time, without violating rights or otherwise acting wrongly. And they pursue this goal by researching how important, neglected, and tractable particular cause areas are and how cost-effective particular interventions are within these cause areas. They then pursue these interventions alongside other projects and relationships (MacAskill 2015, Singer 2015, Centre for Effective Altruism 2022).

If we accept these interpretations of utilitarianism and effective altruism for the sake of discussion, then we can note that they have several important shared features. First, they have at least somewhat similar *natures*. Roughly speaking, we can understand them both as *normative frameworks* – that is, sets of beliefs, values, or commitments regarding what to do or how to live – that govern our priority-setting and decision-making in everyday life. And to the extent that we collapse the distinction between theory and practice or, at least, take there to be links between theory and practice (for instance, to the extent that we think that particular moral theories support particular practical projects, or vice versa), then we might also see utilitarianism and effective altruism as *overlapping* or *linked* normative frameworks.

Second, utilitarianism and effective altruism have at least somewhat similar *contents*. In particular, they both involve an aspiration to maximize positive welfare and/or minimize negative welfare for all sentient beings from now until the end of time, all else equal. This shared commitment to welfarism (roughly, that welfare states are what matter), impartiality (roughly, that all welfare states matter equally, all else equal), aggregation (roughly, that outcomes with the most positive welfare and least negative welfare are best, all else equal), and maximization (roughly, that we have the most reason to produce the best outcomes, all else equal) is striking. Given these substantive similarities, it makes sense that people might see utilitarianism and effective altruism as overlapping or linked normative frameworks.

Moreover, many effective altruists are utilitarians and vice versa. This includes many rank and file effective altruists: A 2019 survey by Rethink Priorities shows that a high percentage of effective altruists endorse utilitarianism and other consequentialist theories

(Dullaghan 2019). It also includes especially influential effective altruists. Peter Singer is the most famous living utilitarian, and his articles and books have been highly influential within effective altruism. For example, his article "Famine, Affluence, and Morality" (1972) has been influential within the global health and development branch of the movement, and his book *Animal Liberation* (1975/2009) has been influential within the animal welfare branch (though, to be clear, neither of these works is specifically utilitarian).

At the same time, utilitarianism and effective altruism also have several important differences according to these interpretations. First, and importantly for our purposes here, they have at least somewhat different natures. Utilitarianism is a moral theory: It tells us how we should live at the level of theory. In contrast, effective altruism is a practical project: It does not tell us how we should live at all, but rather simply exists as an opportunity for people who aspire to at least partly live in a particular kind of way at the level of practice. Some effective altruists believe that they have a moral duty (consequentialist or non-consequentialist) to take advantage of this opportunity (to greater or lesser degrees), and others believe that they merely have a right to do so. Either way, the project itself takes no stand on these matters.

Second, and also importantly for our purposes here, utilitarianism and effective altruism have at least somewhat different *contents*. In particular, utilitarianism is a *monist, consequentialist* moral theory: It requires us to do the most good possible, period, at the level of theory. In contrast, effective altruism is a *pluralist, partly consequentialist and partly non-consequentialist* practical project: It aspires to do the most good possible *without*, for instance, engaging in excessive self-sacrifice, harming or wronging others against their will, cultivating or expressing vice, or cultivating or enacting oppression at the level of practice. And since utilitarianism is a moral theory, it means to govern all our decisions, whereas since effective altruism is a practical project, we can limit its scope in our lives if we like.

Finally, while many effective altruists are utilitarians and vice versa, there are many exceptions to this general rule. Not only do many rank-and-file effective altruists reject utilitarianism and consequentialism more generally (Dullaghan 2019), but several influential effective altruists do too, at least in part. For example, William MacAskill and Toby Ord, often credited as founders of effective altruism, claim to be morally uncertain and argue that we should factor moral uncertainty into our decision-making. And their co-authored (with Krister Bykvist) book on moral uncertainty has been impactful within the effective altruism movement in recent years, with many effective altruists now taking it for granted that we should factor moral uncertainty into our decision-making (MacAskill, Bykvist, and Ord 2020).

To be clear, not everyone shares these interpretations of utilitarianism and, especially, of effective altruism. For instance, some people within effective altruism view it as a moral theory *and* a practical project, and some effective altruists also view it as involving a more or less inclusive set of beliefs, values, and commitments than I do here. These are areas of active discussion and negotiation, and depending on whether and how effective altruists resolve these issues, we might or might not need to complicate the analysis that follows. In any case, my aim here is to assess how utilitarianism and effective altruism can reply to standard objections, particularly what I call the esoterica objection, on these common, if not universal, interpretations. We can then complicate this analysis as needed for other interpretations.

3. Demandingness, cluelessness, injustice, esoterica

Over the centuries, critics have developed a wide range of objections to utilitarianism, and utilitarians have replied in each case. Critics are now developing similar objections to effective altruism, and effective altruists are replying in some similar ways and some different ways. Reviewing these exchanges will be helpful for our discussion of what I call esoteric altruism. With that in mind, we can here review four objections that both utilitarianism and effective altruism face: the demandingness objection, the cluelessness objection, the injustice objection, and the esoterica objection. We can then note how, given the interpretations of utilitarianism and effective altruism that we are assuming here, each framework is more vulnerable to some of these objections than the other, given its distinctive status and content.

First, the demandingness objection to utilitarianism and effective altruism holds that these frameworks are too *practically* demanding, in the sense that they demand that we *do* too much (Kagan 1984, Berkey 2020). In order to do the most good possible, we might need to sacrifice our own interests, projects, and relationships for the sake of distant strangers. We might also need to make this sacrifice on a regular basis, and to do so in all domains of life, including in our thinking what to study in school and do for a living, whether to get married and have children, what to do with our nights and weekends, and more. The demandingness objection holds that nobody should be required to achieve or sustain this level of altruism, and that very few of us would be able to do so reliably even if we tried, due to our motivational limitations.

Second, the cluelessness objection to utilitarianism and effective altruism holds that these frameworks are too *epistemically* demanding, in the sense that they demand that we *know* too much (Lenman 2000, Greaves 2016). In order to do the most good possible, we might need to perform impartially benevolent harm-benefit analysis before making all decisions, estimating which action, out of every option available to us, will maximize positive welfare and/or minimize negative welfare for all sentient beings from now until the end of time. But as with the demandingness objection, the cluelessness objection holds that nobody should be required to assess all our decisions this way, and that very few of us – if anyone at all – would be able to do so reliably even if we tried, due to our epistemic limitations.

Third, the injustice objection to utilitarianism and effective altruism holds that these frameworks are unjust, in the sense that they permit and require unjust actions (Rivera-López 2012, Gabriel 2016). In order to do the most good possible, we might need to harm, kill, or otherwise wrong others for the greater good. For instance, we might need to sacrifice the few for the sake of the many, and we might also need to burden the worst-off for the sake of the best-off. Moreover, given that a commitment to doing the most good possible can involve high-stakes decision-making, the kinds of sacrifices that this commitment might seem to permit or require could be very great indeed. The injustice objection holds that many of these actions (particularly actions that harm or oppress others against their will) are morally wrong.

Finally, the esoterica objection to utilitarianism and effective altruism, which will be my focus here, holds that these frameworks are self-contradictory, in the sense that they require their own rejection (Williams 1985, Adams, Crary, and Gruen 2023). In order to do the most good possible, we might need to persuade everyone, including ourselves, to reject the aim of doing the most good possible. After all, attempting to do the most good possible might not be the best way to actually achieve this goal. Instead, we might select and pursue the wrong actions due to our epistemic or motivational limitations, and we might do less good and more harm than we would otherwise do as a result. If so, then *complying* with these frameworks might require neither *accepting* nor *promoting* nor *implementing* them in practice.

When utilitarians reply to these objections, they often rely on the status and content of utilitarianism as a monist, consequentialist moral theory (Sidgwick 1907, Lazari-Radek and Singer, 2010). That is, they make a distinction between theory and practice, where the point of a theory is to be *correct* and the point of a practice is to be *applied*, and they note that utilitarianism as a theory can require the rejection of utilitarianism as a practice. In particular, if we have a duty to do the most good possible, and if attempting to do the most good possible is a bad way to actually do the good possible, then we can have a duty *not* to attempt to do the most good possible. In this scenario, utilitarians can simply bite the bullet accept that we should neither accept nor promote nor implement utilitarianism in practice.

In contrast, when effective altruists reply to these objections, they often rely both on the status and content of effective altruism as a pluralistic practical project (MacAskill 2022, Ord 2020). For example, effective altruists note that if you think that a *total* commitment to this project is too demanding, then you can always make a *partial* commitment instead. And if you value other projects too, then you can always pursue this project alongside these other ones, demanding more or less of yourself as you wish. They also note that effective altruism includes non-consequentialist norms that rule out unjust actions, and that, as with all projects, we can all constrain our participation with moral commitments that we independently accept. So, we can all pursue a version of this project that we take to be ethical by our own lights.

However, our present concern is that while these latter replies might work for some objections, they might not work for others. In particular, effective altruists might not be able to respond to the cluelessness or esoterica objections *either* in the way that utilitarians respond to *these* objections *or* in the way that effective altruists respond to *other* objections. After all, if we have no clue how to do the most good possible, or if we expect that attempting to do the most good possible is a bad way to actually do the most good possible, then how can we be warranted in building a project both *for* and *around* this aim? Unlike a theory, the point of a project is, at least in part, to be applied. So if effective altruism is self-undermining if applied, then that would seem to be a fatal flaw for the project, not merely an interesting implication of it.

Global priorities researchers have acknowledged the cluelessness objection and have examined this problem in detail (see, for example, Greaves 2016). While this literature is still in progress, the prevailing view appears to be that while our epistemic capacities will always be limited, we can still learn enough for impartially benevolent harm-benefit analysis to be at least *sometimes* be at least *somewhat* useful. For example, even if we might not be able to predict or control all our effects on the far future, we might be able to identify at least *some* actions that can increase the chance of a positive future at least *somewhat*. Much of the current focus on reducing existential risks and expanding our moral circle across species, substrates, nations, and generations comes from this kind of reasoning, and I think that this is good.

However, global priorities researchers have not yet given the same amount of attention to the esoterica objection, and this objection is importantly different from the cluelessness objection. Indeed, the esoterica objection *arises* when the cluelessness objection *fails*, and we determine that we can reliably make at least *one* estimate about the impacts of our actions on the far future: that attempting to do the most good possible is destined for failure. That is, the esoterica objection arises when applying impartially benevolent harm-benefit analysis reveals that applying impartially benevolent harm-benefit analysis is, for all other purposes, counterproductive. In that case, effective altruists should use impartially benevolent harm-benefit analysis for one and only one purpose: their own dissolution.

4. Why might effective altruism be esoteric, and why might it matter?

We can now focus on whether effective altruism has an esoterica problem. Again, the concern here is that effective altruism is self-defeating, since it aims to do the most good possible, yet doing the most good possible requires *not* aiming to do the most good possible. We can now consider why effective altruism might have this problem and the different forms that this problem might take. In short, I will suggest that effective altruism faces this problem if it has some or all of the following features. First, it requires *very* indirect reasoning, to the point that the ultimate aim disappears. Second, it requires us to systematically deceive others about our project. Third, it requires us to systematically deceive *ourselves* about our project. And we have at least some reason to think that it might face all three problems.

As a starting point, we can all accept that effective altruism requires indirect reasoning about what to do and how to live. Consider an analogy with a common egoistic aim. Is *attempting* to be happy the best way to *actually* be happy? Plausibly, the answer is no. If you want to be happy, then consciously *attempting* to be happy will likely be self-defeating. A better strategy is to identify projects and relationships that make you happy, and then to focus on pursuing those projects and relationships in practice. In short, you use your ultimate aim to select proximate aims, and you then use these proximate aims to select the particular aims, rules, habits, and other heuristics that can guide your behavior in everyday life. Plausibly, the project of doing the most good possible works in the same kind of way.

Of course, establishing that a project is indirect in this sense is not enough to establish that the project is in any way problematic. After all, many if not all projects are indirect in this sense. But in most cases, even if we need to reason indirectly in order to achieve our ultimate aim, we can still *accept* the ultimate aim, *promote* the ultimate aim, and *use* the ultimate aim to select the proximate aims, rules, habits, and so on that govern our decision-making more directly. For example, it would be a mistake for a soccer player to make every in-game decision by estimating which action (running left? running right?) would be most likely to contribute to winning the World Cup. But they can still accept this aim, promote this aim, and use this aim to identify all the strategies and tactics that govern their play.

However, there three ways in which a project can go beyond being merely indirect. First, a project can be *very* indirect. Suppose that impartially benevolent harm-benefit analysis is to practical reason what string theory is to theoretical reason. It might be foundational, and it might be useful to apply directly *sometimes*. But it is not useful to apply directly for *the vast majority of people the vast majority of the time*. Very few people should think in terms of string theory at all, and even these people should think in these terms only rarely, despite (or perhaps because of) its foundational status. Similarly, perhaps very few people should think in terms of impartially benevolent harm-benefit analysis at all, and even these people should think in these terms only rarely, despite (or perhaps because of) its foundational status.

If effective altruism were esoteric in this sense, then that would be a problem because it would create a disconnect between the stated aims of the project and most of the actual work of the project. It would be like starting a club that looks at planets through telescopes and calling it the string theory club, simply because string theory is at the foundation of astronomy and is, perhaps, *sometimes* useful when thinking about astronomy. Granted, it might not be *self-contradictory* to frame this club this way. But it would still be *weird*. When people frame a project in terms of a particular set of ideas, we expect that these ideas will play a relatively central role in the work. If the ideas instead merely govern the work from a vast, more or less invisible distance, then we might feel, if nothing else, misled.

Another, stronger version of the problem can arise when a project requires *systematic deception*. Suppose that *some* people can be responsible effective altruists, but that most people cannot be. In particular, suppose that most people who aspire to do the most good possible would either apply impartially benevolent harm-benefit analysis too directly or neglect non-consequentialist constraints in the course of applying this decision procedure. Suppose further that, if many people knew about effective altruism, then the harm done by the irresponsible effective altruists would outweigh the good done by the responsible ones. In this case, it might be that those who can participate in this project responsibly should do so, but that they should also keep the project a well-guarded secret, and perhaps even publicly reject it.

If effective altruism were esoteric in this sense, then that would be a problem because systematic deception is risky and costly. Many people regard deception as intrinsically morally wrong. And whether or not we endorse that idea, we can all agree that systematic deception is bad in other ways. It involves the risk of exposure, the cost of maintaining the deception, and the opportunity cost of not being able to grow your community as much as you otherwise might. And of course, conventional wisdom holds that large-scale conspiracies are difficult if not impossible to sustain, since someone, somewhere, will leak the information intentionally or accidentally. So a project that needs to be esoteric in this sense will likely need to either remain small and secretive or become larger in risky and costly ways.

Another, stronger version of the problem can arise when a project requires *systematic self-deception*. Suppose that *nobody* can be a responsible effective altruist. No matter how much research we conduct, no matter how much humility we cultivate, and no matter how much we internalize the value of indirect reasoning and non-consequentialist constraints, we can still expect that our efforts to do the most good possible will backfire, due to our epistemic and motivational limitations. In this case, it might be that anyone who aspires to do the most good possible should try to persuade not only *others* but also *themselves* to reject this aim, sacrificing theoretical rationality (that is, the kind of rationality that governs our beliefs) for the sake of practical rationality (that is, the kind of rationality that governs our actions).

If effective altruism were esoteric in this sense, then that would be a problem because systematic self-deception is, of course, risky and costly too. Many people regard self-deception as, if not morally wrong, then at least theoretically irrational. Self-deception is also difficult to achieve. Our beliefs are beyond our direct volitional control, so persuading ourselves to change our minds requires placing ourselves in situations that cause us to accept as true what we currently take to be false. And even if we achieve self-deception, our doing so can carry further costs, since our beliefs are all interconnected and at the root of many of our decisions, and so having false or irrational beliefs and belief-forming practices in one domain can lead to having false and irrational beliefs and belief-forming practices in other domains too.

The question that we face, then, is whether and to what extent effective altruism has some or all of these three features. Does it require very indirect reasoning, systematic deception, or systematic self-deception? Of course, these are empirical questions, and it would be impossible for me to answer them here. So instead of attempting to do that, I will present a general hypothesis about what the answer will turn out to be. Roughly speaking, my hypothesis is that effective altruism is *partly but not fully* esoteric in all three of these respects. It requires moderate indirect reasoning, moderate deception, and moderate self-deception, which is not particularly problematic. However, I will also suggest that questions remain about whether effective altruism has stronger, more problematic versions of these features, too.

5. *Is* effective altruism esoteric?

I can start by noting one reason why I feel skeptical of the esoterica objection. As noted above, the esoterica objection suggests that we can use impartially benevolent harm-benefit analysis reliably for one and only one purpose: determining that impartially benevolent harm-benefit analysis is, for other purposes, counterproductive. But that seems implausible. Plausibly, either we can *sometimes* use this decision procedure reliably, including for purposes other than its own assessment, or we can *never* use it reliably, including for its own assessment (in which case we return to the cluelessness objection). Either way, the esoterica objection seems to fail. But having expressed that thought, I can now set it aside to consider how much indirect reasoning, deception, and self-deception effective altruism plausibly requires.

First, it seems clear that effective altruism requires indirect reasoning. The only question is how much indirect reasoning it requires, which is a contextual matter. As I discuss elsewhere, effective altruism started out relatively direct, in part because it was an experiment and in part because it was complementing other efforts, and so it could focus on applying impartially benevolent harm-benefit analysis relatively directly. Over time, effective altruism has become more indirect, in part because effective altruists better appreciate the need for indirect reasoning, and in part because effective altruism is more powerful and, so, it needs to diversify internally more (since the more powerful you are, the more you need to be comprehensive rather than merely complementary) (Sebo and Singer 2018, Sebo 2019).

What happens next depends on how effective altruism develops. The more evidence effective altruists collect, the more they can learn how to strike a virtuous balance between direct and indirect reasoning in practice. And the more powerful effective altruism becomes, the more they might need to deploy these indirect decision procedures by default. This might require a different trajectory for different cause areas within effective altruism, since, for instance, effective altruism is on track to be more influential within the animal welfare space than within the global health and development space (given how much more neglected the former cause area is within governments and other foundations), which might require taking a more indirect approach within the former cause area than within the latter, all things considered.

If effective altruism becomes increasingly indirect in this sense, will that make it problematic in the same kind of way that an astronomy club that calls itself a string theory club would be? Not necessarily. As long as the aim of doing the most good possible remains an important part of effective altruism in practice, it will be natural to build this project both *for* and *around* this aim, even if effective altruists pursue other, proximate aims as well, and even if these other, proximate aims are relatively removed from the ultimate aim. And given that the aim of doing the most good possible is part of what informs the antispeciesism, longtermism, and importance-neglectedness-tractability framework in effective altruism, this aim continues to earn a place in the framing of the project even if it fades from view in other respects.

Second, how much deception does effective altruism require? We can start by stating the obvious: Effective altruism addresses domains of life that can involve info-hazards, that is, information that can be dangerous. This includes information about both facts and values. For instance, empirical information about artificial intelligence and biosecurity can be used for good as well as for evil, and this information should be carefully managed. Additionally, even if normative frameworks like antispeciesism and longtermism are correct, they can be used for good as well as for evil too (since it can be easy to rationalize, say, definitely harming current humans in order to *possibly* help *much* larger numbers of future nonhumans), and so information about these frameworks should perhaps be carefully managed as well.

With that said, even if *some* effective altruist ideas are dangerous in these ways, *others* might not be. For instance, the idea that we should prioritize issues like malaria and factory farming are clearly good. Moreover, even when effective altruist ideas are potentially dangerous, effective altruists might be able to manage them without resorting to systematic deception. For example, suppose that effective altruists commit to promoting frameworks like antispeciesism and longtermism only when they can contextualize these frameworks appropriately, by emphasizing the importance of reasoning indirectly, respecting rights, cultivating virtuous characters, and so on. In this case, even if effective altruists need to keep *some* secrets, they might also be able to share many of their most important ideas openly.

My view is that this moderate approach is enough to mitigate the risks associated with promoting effective altruism, and is not particularly problematic. Granted, effective altruists might need to be thoughtful about whether, when, and how they promote potentially dangerous ideas, but this kind of discretion is common and appropriate. And granted, they might sometimes need to fully conceal particular dangerous ideas as well, for instance if global priorities researchers have access to information about artificial intelligence or biosecurity that, if publicly available, would be a clear security threat. But this kind of discretion is common and appropriate as well. In these ways, I expect that effective altruists can manage dangerous ideas responsibly while still cultivating the virtues of honesty and transparency.

Finally, how much self-deception does effective altruism require? We can once again start by stating the obvious: There is no guarantee that theoretical and practical reason will always align. And in this case, there are multiple reasons why *attempting* to do the most good possible might conflict with *actually* doing the most good possible. Info-hazards might apply to us as well. Additionally, investing in our own projects and relationships might require seeing them as primarily *finally* valuable as opposed to primarily *instrumentally* valuable. After all, if you see your education, your career, or your family *primarily* as means to the end of doing the most good possible in everyday life, then you might not be able to show up in these roles in the right kind of way, even if you also see them as ends in themselves.

However, I think that the same caveats about deception apply here as well. Even if *some* effective altruist ideas are dangerous in these ways, *others* might not be. Moreover, even when effective altruist ideas are potentially dangerous, effective altruists might be able to manage them without resorting to systematic self-deception. For example, if effective altruists commit to promoting dangerous ideas only when they can contextualize them, then this commitment might lead not only *others* but also *effective altruists themselves* to internalize this framework in the right kind of way. And we can expect that people who commit to particular projects and relationships will naturally come to see them primarily as ends in themselves *in practice*, even if they continue to see them primarily as means to a further end *in theory*.

As with deception in general, my view is that this moderate approach is enough to mitigate the risks associated with accepting effective altruism, and is not particularly problematic (Sebo 2015). Granted, effective altruists might sometimes experience a tension between how they evaluate particular projects and relationships in theory and in practice. But this kind of tension is common and appropriate. Indeed, non-consequentialists ranging from Immanuel Kant to Thomas Nagel note that this kind of tension arises regularly – implicating our views about meaning, value, the self, free will, and more – and that when it does, it might require us to see some ideas as true in theory but false in practice (Kant 1781/2003, Nagel 1986). So if effective altruism were like that, then it would be in good company. But time will tell.

6. Conclusion

Utilitarians have long understood that the principle of utility might be correct as a criterion of rightness but not as a decision procedure. At the level of theory, we should comply with this principle. An action is right if and only if, or to the extent that, it maximizes positive welfare and minimizes negative welfare for all sentient beings from now until the end of time. But at the level of practice, we should not necessarily always, or even ever, follow this principle. Attempting to maximize utility is not necessarily always, or even ever, a good way to actually maximize utility. So, if and when following a non-utilitarian decision procedure is what would maximize utility, utilitarianism implies that we should follow this non-utilitarian decision procedure, avoiding utilitarian reasoning for the sake of utilitarian outcomes.

Utilitarians have also long understood that the principle of utility might be *esoteric*, in the sense that it might require very indirect reasoning, systematic deception, or even systematic self-deception. Granted, it might be that maximizing utility requires accepting this aim, promoting this aim, and applying this aim at least *sometimes*. But it might also be that maximizing utility requires rejecting this aim and persuading everyone, including ourselves, that this aim is bad. If utilitarianism were esoteric in any of these respects, then utilitarians would need to manage information and arguments about utilitarianism carefully. And if utilitarianism were esoteric in all of these respects, then utilitarians might need to go farther: At the limit, they might need to destroy utilitarianism for the sake of utilitarian outcomes.

To their credit, utilitarians tend to bite the bullet on this issue. If utilitarianism requires burning every book that Bentham, Mill, and Sidgwick ever wrote and convincing everyone, including ourselves, that the Ten Commandments are correct as a moral theory instead, so be it. Hand me a lighter. Utilitarianism can be correct in theory even when it requires its own destruction in practice. Of course, utilitarians might also be skeptical that utilitarianism does, in fact, have this implication; otherwise, if they were truly committed to utilitarianism, then they would presumably be devoting their time and energy to defending the Ten Commandments, not utilitarianism. But at least in principle, utilitarians are prepared to endorse this implication. In my view, this response to the esoterica objection is a good one.

But effective altruism might not be so lucky. If, as I assumed here, effective altruism is a practical project instead of a moral theory, then it makes sense only if doing the most good possible involves accepting, promoting, and implementing this aim at least *somewhat*. Yet whether attempting to do the most good possible actually does the most good possible is an open question. And if it turns out that burning every book that Singer, Ord, and MacAskill ever wrote and persuading everyone, including ourselves, that we should all do what we love and support local charities, then effective altruists would need to accept, promote, and implement this course of action. And effective altruism, as a practical project that involves using evidence and reason to do the most good possible within particular moral limits, would effectively be dead.

My hypothesis, which I developed without fully defending here, is that effective altruism is *partly* esoteric in these respects. Doing the most good possible does, in fact, require applying impartially benevolent harm-benefit analysis selectively and indirectly; exercising discretion about when and how to promote effective altruist ideas; and embracing a tension between the priorities that we accept in theory and in practice. But this is fine. Many projects are partly esoteric in these respects. The real question is whether effective altruism is also esoteric in a stronger, more problematic sense, such that the aim disappears entirely and we need to conceal it from everyone, including ourselves. The viability of effective altruism depends on the answer to that question, and while I expect that the answer is no, the jury is still out.

7. References

Adams, Carol, Alice Crary, and Lori Gruen, ed, 2023, *The Good It Promises, the Harm It Does: Critical Essays on Effective Altruism*, Oxford University Press.

Bentham, Jeremy, 1789/2007, *An Introduction to the Principles of Morals and Legislation*. New York: Dover Publications.

Berkey, Brian, 2020, "Effectiveness and Demandingness," *Utilitas* 32 (3): 368-381.

Centre for Effective Altruism, 2022, "What Is Effective Altruism?" https://www.effectivealtruism.org/articles/introduction-to-effective-altruism

De Lazari-Radek, Katarzyna and Peter Singer, 2010, "Secrecy in Consequentialism: A Defence of Esoteric Morality," *Ratio 23(1)*:34-58.

Driver, Julia, 2012, *Consequentialism*. New York: Routledge.

Dullaghan, Neil, 2019. "EA Survey 2019: Community Demographics & Characteristics": https://rethinkpriorities.org/publications/eas2019-community-demographics-characteristics

Gabriel, Iason, 2016, "Effective Altruism and Its Critics," *Journal of Applied Philosophy 34:4* 457-473.

Greaves, Hilary, 2016, "Cluelessness," *Proceedings of the Aristotelian Society 116(3)*: pp. 311-339.

Kagan, Shelly, 1984, "Does Consequentialism Demand Too Much?" *Philosophy & Public Affairs 13(3)*: 239–254.

Kant, Immanuel, 1781/2003, *Critique of Pure Reason*. London: Penguin Classics.

Lenman, James, 2000. "Consequentialism and Cluelessness," *Philosophy & Public Affairs 29(4*), pp. 342-270.

MacAskill, William, 2015, *Doing Good Better*. New York: Gotham Books.

MacAskill, William, 2022, *What We Owe The Future*. New York: Basic Books.

MacAskill, William, Krister Bykvist, Toby Ord, 2020. *Moral Uncertainty*. Oxford: Oxford University Press.

Mill, John Stuart, 1861/1998, *Utilitarianism*, Oxford: Oxford University Press.

Nagel, Thomas, 1986, *The View From Nowhere*. Oxford: Oxford University Press.

Ord, Toby. 2020. *The Precipice*. New York: Hachette Books.

Rivera-López, Eduardo, 2012. "The Moral Murderer. A (more) effective counterexample to consequentialism," *Ratio*, 25(3): 307–325.

Sebo, Jeff, 2019, "Effective Animal Advocacy," *The Routledge Handbook of Animal Ethics*, ed. Bob Fischer. New York: Routledge.

Sebo, Jeff and Peter Singer, 2018, "Activism," in Lori Gruen, ed., *Critical Terms for Animal Studies*. Chicago: Chicago University Press.

Sebo, Jeff, 2015, "Utilitarianism, Multiplicity, and Liberalism," *Utilitas 27 (3)*, 326-346.

Sidgwick, Henry, 1907/1981, *The Methods of Ethics (7th edition)*. Indianapolis: Hackett Publishing.

Singer, Peter, 1972, "Famine, Affluence, and Morality." *Philosophy & Public Affairs 1:3*: 229-243.

Singer, Peter, 1975/2009, *Animal Liberation: updated edition*. New York: HarperCollins Publishers.

Singer, Peter, 2015, *The Most Good You Can Do*. New Haven: Yale University Press.