Considering Considerateness

Why communities of do-gooders should be exceptionally considerate

Stefan Schubert[1], Ben Garfinkel[2], and Owen Cotton-Barratt[3]

Summary:[4] *When interacting with others you can be considerate of their preferences, for instance by being friendly or reliable. This normally has small positive direct effects. But, by improving your reputation or strengthening aspects of culture that make a community more cooperative, the positive indirect effects can be large.*

We present the case that these indirect effects are further strengthened when you are acting as part of a community of people doing important work. For instance, being considerate can improve the level of trust and collaborativeness among members of the community. It can also improve the reputation of the community. Conversely, failing to be considerate can harm the community, both internally and in its reputation.

This means that for communities of people striving to do good, such as the effective altruism community, considerateness should be a surprisingly high priority. It could be that, in order to do the most good, they should be considerably more considerate than common sense morality requires.

Introduction

[1] Centre for Effective Altruism, stefan.schubert@centreforeffectivealtruism.org.
[2] Centre for Effective Altruism.
[3] Future of Humanity Institute, University of Oxford.
[4] The authors would wish to thank Raymond Arnold, Harri Besceli, Joseph Carlsmith, Lucius Caviola, Max Dalton, John Halstead, Michelle Hutchinson, Daniel Kokotajlo, Oge Nnadi, James Snowden, Emily Tench, and Julia Wise for their comments on this article.

Common sense morality[5] suggests we should generally be considerate toward others: that we should, for instance, be friendly, honest, and modest. Most ethical views agree.

However, one can only ever be considerate up to a certain point. This is because considerateness has costs, for instance in lost opportunities and divided attention.[6] Furthermore, for people working on large-scale issues like global poverty, these costs ought to be regarded as ethically significant. This raises the question: *how considerate should effective altruists be?*

We will focus on the consequences of considerateness: that is, the costs and benefits of effective altruists being considerate. We argue that in the context of a community of people doing good, outcomes-based reasoning supports a surprisingly high degree of considerateness. For consequentialists, this conclusion is directly action-guiding. For non-consequentialists additional interpretation may be required: they will place less weight on outcomes-based reasoning, but often attribute intrinsic value to considerateness.

Conceptualizing considerateness

What do we mean by considerateness?

> **Considerateness:** One is being *considerate* if the way one treats others, in personal interactions and communications, is notably guided by how they would like to be treated.[7]

For example, friendliness and intellectual honesty are typically considerate behaviors, in that people ordinarily prefer that others are friendly toward them and don't try to mislead them. Considerate behaviors include:

- Friendliness
- Honesty
- Intellectual honesty

[5] We here use the term "common sense morality" to refer to the moral principles that people would embrace without reflection, and that they largely live by. This set of principles may in turn entail further principles which people, inconsistently, do not embrace. These latter principles are, however, not part of common sense morality, as we use that term.

[6] See the section "Possible reasons not to be considerate" for a discussion of types of costs of considerateness.

[7] Two clarificiations. First, although sometimes it will be obvious, or easy to learn, how a person would like to be treated, there will also be many cases in which their preferences are less clear. In such cases, we might say, a bit more precisely, that one is being considerate if one is notably guided by *reasonable expectations* of these preferences.

Second, people often have conflicting preferences. For instance, someone might instinctively dislike to have their career plans criticized, yet upon reflection realize that it is in their best interest to receive such criticism. In such cases being considerate with respect to one preference entails being inconsiderate with respect to another preference. Which course of action to take must be decided on a case-by-case basis, but looking at people's reflected preferences, rather than their instinctive preferences, seems like a good rule of thumb. (Thanks to Oliver Habryka for prompting this second point.)

- Cooperativeness
- Respectfulness
- Modesty
- Integrity
- Reliability
- Rule-following[8]

As this list shows, many considerate behaviors are emphasized by common sense morality. It also appears to be a property of most considerate behaviors that their direct effects are positive but small, at least compared to the direct effects one can have through charitable donations or direct work. For example, the direct effect of friendliness isn't much more than briefly lifting another person's mood.

Inconsiderateness is normally considered selfish. However, there are two main reasons why altruists, and especially altruists striving to have large-scale impacts, might choose to be inconsiderate. First, it may sometimes appear that being inconsiderate can help you do more good. For instance, although potential donors are likely to prefer even-handed portraits of a charity's effectiveness, sharing only positive evidence (and no negative evidence) might help bring in more donations. Second, it may seem that the benefits from being considerate are sometimes so small that you could just as well be, e.g., unfriendly, and spend your attention elsewhere.

If you only look at the direct outcomes, it will not be immediately clear that high levels of considerateness are justified in such cases. If outcome-based reasoning implies that considerateness should be given substantial weight, even by people working on large-scale problems like global poverty, then this must be because some considerate behaviors have significant indirect effects. We argue that this usually is the case, and thus if you want to do good effectively, then you ought to be highly considerate. However, we grant that there are circumstances in which it is warranted to be inconsiderate.

Indirect benefits of considerateness

The distinction between direct and indirect effects has received substantial attention within the effective altruism community.[9] A common criticism of analyses of intervention effectiveness is that they fail to appropriately take significant indirect effects into account.[10]

[8] Some rules may promote inconsiderate behaviour and so not overall be considerate to follow. But when other things are equal, following rules is a form of cooperation with the norms of the system in which one is acting.
[9] The line between direct and indirect effects isn't crisp, but the distinction is nevertheless useful. See the EA concepts page for a discussion of this distinction.
[10] See, e.g., Paul Christiano (2014).

People who fail to take the indirect effects of (in)considerateness into account are sometimes called *naive* or *short-sighted*. A classic example is a utilitarian who lies to further altruistic ends, while failing to pay heed to negative indirect effects. Though such naivety is especially associated with consequentialism, it can affect anyone judging the value of behaviours (such as considerateness) at least partly based on outcomes.[11]

There are at least two types of indirect damage you can incur from being inconsiderate. First, inconsiderateness tends to harm your reputation, which in turn reduces your ability to do good (among other things) in the future. Second, it tends to have wider negative effects on social capital. By this we mean aspects of culture that make a community more cooperative and better able to achieve its shared goals.[12]

To get a more vivid sense of what social capital is, let us look at a key aspect: *trust*. In communities where people don't trust each other, they have to invest significant resources to prevent getting preyed on. They may have to buy sturdy locks for their bikes, and alarm systems for their houses. Perhaps more importantly, they will be wary of entering into any sort of agreement where it is difficult to verify whether their partners will honor the agreement.

If you behave dishonestly, people will increase their credence that others are similarly dishonest. You may also inspire others to be dishonest, through the bandwagon effect. Through both of these mechanisms, you will be undermining trust in your community. Thereby you will destroy social capital.

Just as inconsiderateness normally has negative indirect effects, considerateness tends to have positive indirect effects. It can improve your reputation, and so increase your ability to do good in the future. It can also increase the level of social capital, which can reduce animosity and increase productivity within society.

In fact, when considering different possible definitions of "considerateness," we purposely settled on one that would link the concept closely to these positive effects. If you treat people you interact with in the way they would like to be treated, then they, and others, will (rationally) become more confident that their preferences will be accounted for in future

[11] See, e.g., John Stuart Mill (1838), p. 111, Henry Sidgwick, (1907), pp. xxii–xxiii, Roger Crisp (1992), Toby Ord, (2009), pp. 105-107. For an overview of the philosophical literature on naive applications of consequentialism, see Stanford Encyclopedia of Philosophy. For more recent discussions on naive consequentialism and the value of considerateness within the rationalist and effective altruism communities, see, e.g., Scott Alexander (2014a), (2014b), Raymond Arnold, Amanda Askell, Hannah Blume, Paul Christiano (2016), Patrick LaVictoire, Brian Tomasik (2013/2016), and Eliezer Yudkowsky (2008a), (2008b).

[12] See the definition of "social capital" given by The Saguaro Seminar, Harvard Kennedy School. Note, however, that the term "social capital" is sometimes used in other ways; e.g., to refer to an individual's ability to achieve their self-regarding goals via social means. See Wikipedia.

interactions as well. This confidence in you constitutes a core aspect of positive reputation. The confidence in the community more generally constitutes a core aspect of social capital.

Ethical theories which focus on outcomes, such as utilitarianism and other forms of consequentialism, are sometimes criticized on the grounds that they justify dishonesty and other breaches of common sense norms. The standard defence against such criticism is that once you take indirect effects of your actions into account, there is considerable convergence between them and common sense morality.[13] On this view, even theories which don't assign any intrinsic value to considerateness should value considerateness for instrumental reasons.[14]

Considerateness in a community

We have explained (and it is widely agreed) that the benefits of considerateness are stronger than it might first appear. What is less commonly recognized is that the benefits of considerateness are especially strong for people who are part of a tightly knit community, particularly when that community is doing important work.

First, actions undertaken within the community will affect one's reputation among members of the community, as well as the social capital of the community. Second, actions undertaken as part of the community which are externally visible will affect its reputation and may, in some circumstances, have an outsized impact on the social capital of society at large. Hence considerateness and its absence have four types of indirect effects:

	Effects on reputation	Effects on social capital
Internal effects	Effects on your reputation within the community	Effects on community social capital
External effects	Effects on the community's reputation	Effects on societal social capital

Figure 1: Indirect effects of being (in)considerate as a member of a community

Effects on your reputation within the community

We saw that the indirect effects of considerateness on your own reputation can be a significant reason to be considerate. This reason is amplified when you are acting within a

[13] Cf. footnote 11.

[14] There are also other arguments to that effect. One of them is that following existing common sense norms is a simple and cheap decision procedure.

community—particularly if the community is tightly knit, and if it is important for you to continue operating within that community. If you behave badly in society at large, this may not be remembered. If you are inconsiderate in a community that isn't important to you, you may escape a poor reputation by changing social environment. If you develop a bad reputation in a community which is important to you, particularly one that has increased your ability to do good, there is a risk that this reputation will be costly to operate with or escape. Conversely, if your peers hold you in high esteem in an important environment, your ability to operate effectively can be significantly increased. It is plausible that the effective altruism community may represent such a case for many of its members.

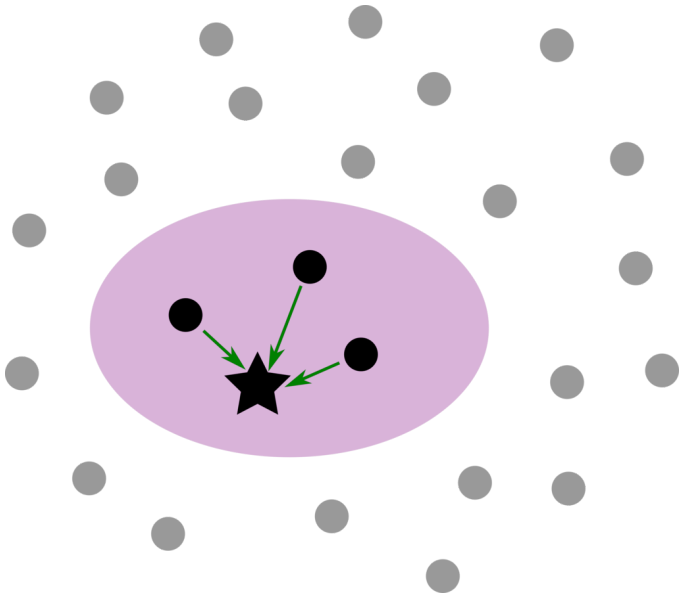

Figure 2: Your reputation within a community

Effects on community social capital

Similarly, the negative effects of failing to be considerate on social capital can be amplified when you are acting within a community whose output you care about. For instance, lowering the level of trust within the effective altruism community would obstruct the work of people who are trying to do good as effectively as possible. If breaches of social norms cause other members of the community to behave in a similar way, the negative effects are strengthened further. On the other hand, if you increase the social capital of an important community, you can, for the same reason, have a large positive impact.

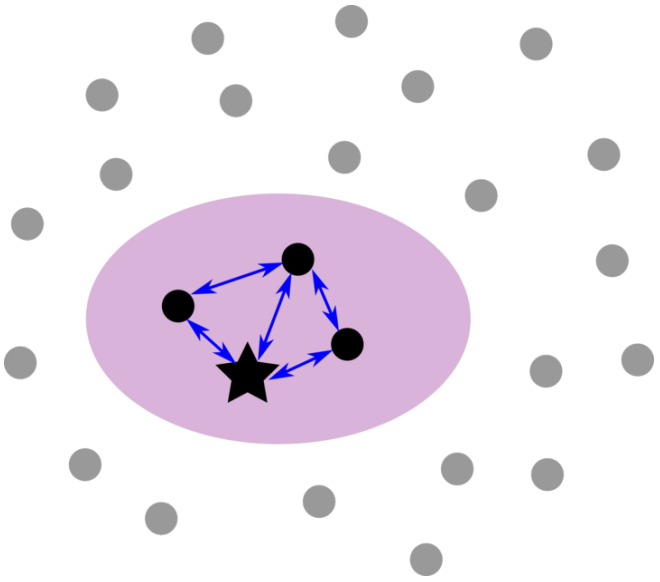

Figure 3: Community social capital

Working to improve the community's social capital may be especially important when the community is likely to grow a lot. Early considerateness helps to set a culture of considerateness among a growing community and thus engender more considerateness in the longer term. That could in turn generate positive effects through each of the mechanisms discussed here (except personal reputation).

Effects on the community's reputation

Members of any community at times act—to varying degrees depending with context—in ways that are externally visible and will reflect on the community. This means that members of the community behaving inconsiderately may threaten its reputation. In some cases the effects could be severe, especially if the behavior is commonly seen as deeply immoral (examples include conspicuously unfriendly behavior, sexism, and racism). Conversely, if members of the community are seen to act in considerate ways, its reputation may be strengthened.

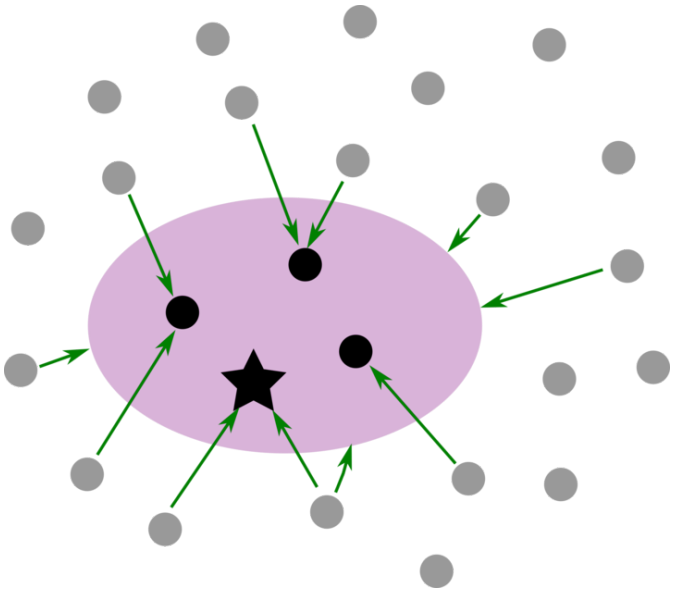

Figure 4: The community's reputation

If the community is doing valuable work, a better reputation may be important by helping to facilitate that. Others are more likely to look favourably upon and help members of a community if they have a generally good impression of that community. People are more likely to want to join communities with good reputations (or leave communities with bad reputations). This is particularly important for communities where this is a major mechanism; for example, social movements depend more on reputation for growth than technical communities do.

Another reason not to be inconsiderate, for instance for short-term gains, is that such a move could be hard to reverse. If the community sacrifices its reputation for short-term gains, it could be difficult and costly to re-build it.

Effects on societal social capital

The considerateness of any action will have some effect on societal social capital. In some circumstances this effect might be amplified when you are acting visibly as a member of a community. This seems most likely when the community is widely viewed favourably and is well-known. For instance, if members of such a community display conspicuous epistemic humility, that may set a positive precedent. It may be that altruistic communities could have a significant impact through such norm-setting effects (we'll return to this in the concluding discussion).

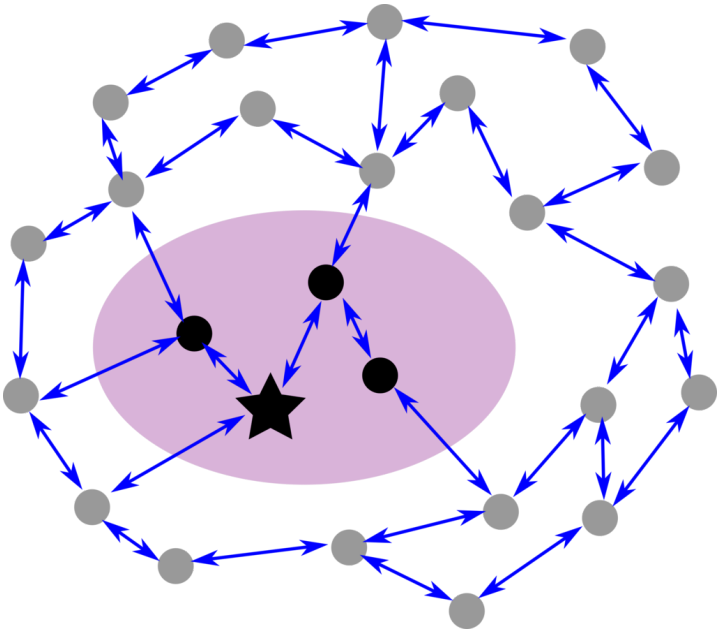

Figure 5: Societal social capital

Reasons for considerateness particular to the effective altruism community

The above reasons for considerateness hold true for members of many different kinds of communities and movements. It is therefore unsurprising that many communities go to great lengths to, e.g., prevent their members from harming their reputation. However, there are also reputational reasons for considerateness which are particular to the effective altruism community.

The effective altruism community is to some extent associated with consequentialism. There is a stereotype of consequentialists and, in particular, utilitarians, as being callous and rule-breaking, as shown, e.g., by Jim Everett, David Pizarro, and Molly Crockett.[15] (Cf. also many characters in novels and movies; e.g, Raskolnikov in *Crime and Punishment* and Grindelwald in *Harry Potter*.) This means that the effective altruism community arguably has to go out of its way to prevent that stereotype from being triggered.[16, 17]

[15] See also Julian Baggini's comment.

[16] Against this, it could be argued that this stereotype is so entrenched that it is effectively impossible to dispel. If that would be true, it would lessen the reasons for considerateness. If the effective altruism community is going to be seen as inconsiderate no matter what it does, investing in pushing against this stereotype by behaving considerately carries little benefit (but can be costly).

However, we find it implausible that it would be impossible to affect this stereotype. Since we are also concerned that this stereotype is one of the larger risks the community faces, we believe it is quite important to push against.

[17] Another reason why it might be particularly important for the effective altruism community to be considerate is that it has a reputation for being elitist. This could create resentment against the community. Being conspicuously considerate might be a way of dispelling that resentment.

In fact, if the community doesn't do that, there is a risk that not only the community, but also the very idea of outcome-based reasoning in altruism, will suffer permanent harm. Rightly or wrongly, it would be argued that such reasoning, when put to practice, lead to objectionable ends-justifies-the-means behavior. And conversely, if the community behaves considerately, the stereotypes could be weakened.[18] (This is of course especially true if the community also does good in other more directly visible ways.) More often than not, ideas are judged by how they fare in practice, rather than by the strength of the arguments for them.[19] The likely effects of consequentialism in practice have been the subject of extensive armchair discussion. The effective altruism community in some ways presents a unique empirical test.[20] It could prove more decisive.

Possible reasons not to be considerate

We have so far explored the positive case for considerateness. However, considerateness also carries costs. First, inconsiderateness can offer a quicker or easier way of getting an effect, which means that considerateness carries opportunity costs. Second, considerateness can in some cases cause indirect harms. This means that inconsiderateness can sometimes be warranted.[21]

In addition, we are often faced with trade-offs between different kinds of considerate behavior. For instance, honesty can sometimes conflict with friendliness and politeness (although it is usually possible to combine these virtues). Though this isn't a reason to lower our general level of considerateness, it may be a reason against being considerate in some specific way.[22]

The costs of considerateness will no doubt vary across different types of considerate behaviors. For instance, the costs of behaving with impeccable integrity may be considerable. By contrast, refraining from online bullying is normally less costly.

---

[18] Note, however, that if consequentialism in fact does has negative effects when applied, then consequentialists should, ironically, want consequentialism's reputation to be harmed. Consequentialists should arguably want people's beliefs of how consequentialism would work if applied to match the true facts of the matter.

[19] For instance, communism probably grew unpopular partly because of how it fared in practice; e.g., the Soviet Union's loss in the Cold War.

[20] Recent years have seen an increased interest in the use of empirical methods in philosophy. See, e.g., experimental philosophy, experimental moral philosophy, and naturalism in epistemology.

[21] However, an additional consideration is that being considerate even on occasions where that does not seem obviously warranted helps forming a habit of considerateness. Having acquired this habit may help you overcome the temptation of being inconsiderate in situations where that would be clearly harmful. (This point thanks to a conversation with Shamil Chandaria.)

[22] While the question of how to resolve such trade-offs between different kinds of considerateness is important, it is outside the scope of this article.

Let us look at some specific mechanisms through which considerateness can be costly.[23]

Shaping incentives

Being considerate often makes others happier to interact with you. That is normally good, but in some circumstances may not be desirable. If people find you extremely helpful when they ask you about frivolous matters, they will be incentivized to keep asking you about such matters. If you would prefer them not to, you should not be quite so helpful.

Similarly, a possible reason to be actively inconsiderate is to disincentivize undesirable behaviour (cf. the notion that we shouldn't always tolerate the intolerant). Common sense morality might agree: when someone behaves badly it is often thought that they have waived their right to considerate treatment. We think that this could be a good reason to drop certain types of considerateness—friendliness, respectfulness, cooperativeness—which are most salient to the relationship with the person being interacted with. It doesn't seem like a good reason to drop types of considerateness—intellectual honesty, integrity—which are likely to be interpreted as indicative of your behavior in general, and which have strong effects on social capital. We further suggest that if you are inconsiderate for this reason, it is helpful to make that explicit, in order both to strengthen the incentive effect and to avoid damaging reputation or social capital.

Getting attention

Acting to attract attention through, e.g., sending multiple emails on a topic, is often an imposition on the people whose attention you are grabbing.[24] It may therefore to some extent be inconsiderate. At the same time, it can be valuable to attract attention to important but neglected issues. Hence being considerate through refraining from such behavior carries opportunity costs.[25] Unlike cases where it is just extra work to be considerate, there may be no way to get the good effects without the bad in these cases. Hence there is some irreducible tension.

In some cases, the act of being inconsiderate itself may attract more attention (if it is particularly unusual). This is sometimes taken advantage of by social activists.[26] Our guess is that the indirect costs of inconsiderateness mean that such cases are rarely worth it for effective altruists.

---

[23] Note that several of these mechanisms aren't unique to action within a community. They effectively are general reasons against considerateness.

[24] Though not always. Setting off a fire alarm is a violent way of grabbing attention, but is considerate toward their considered interests if there is actually a fire.

[25] In the context of social movements, see Owen Cotton-Barratt (2015) for discussion of the trade-offs between attracting attention and building a good reputation.

[26] PETA is a prominent example.

Unreasonable preferences

On our definition, to be considerate means to treat others, in personal interactions and communications, as they would like to be treated. This raises the question what to do if people's preferences are unreasonable. For instance, suppose that someone demands that their colleagues in an open office be completely silent. Should such preferences be met?[27]

This question has to be decided on a case-by-case basis, but in general, it seems that if the preferences really are unreasonable—which of course in itself often is contentious—then they should not be met. For instance, doing so could encourage people to develop ever more demanding preferences.

In some circumstances, there could be strong countervailing considerations, however. For instance, in some cases failing to meet such preferences could lead to a significantly worse relationship with the person(s) with the preferences. If that relationship were important enough that might outweigh the costs of complying. But in general, it seems that only satisfying reasonable preferences with respect to considerateness is a good rule of thumb.

Short-term time horizons

Being considerate can be seen as a capacity-building investment. For instance, improving the effective altruism community's reputation, or its social capital, can increase the effectiveness of future work. Just like other kinds of capacity-building, the value of being considerate therefore depends on how you weigh work done in the shorter term with work done in the longer term. If you think that it is urgent to maximize the effective altruism community's direct impact on the world in the next few years, for instance through work on global poverty, animal welfare, or existential risk, being considerate may be relatively less important. If you rather think that most of the value of the effective altruism community lies several decades ahead, the case for considerateness is strengthened.

This point particularly affects benefits of considerateness that might unfold over a long time, like improving societal social capital, whereas it can be less strong for benefits like personal reputation. Short-term time horizons lower the benefits but don't make them negative, so just raise the relative importance of opportunity costs.

Concluding discussion: implications for effective altruism

With respect to prioritization between causes such as global poverty, animal welfare, and existential risk, the effective altruism community has been at pains to take indirect effects

[27] A similar problem is posed by preferences which are viewed as reasonable in some cultures, but as unreasonable in others.

into account. For instance, the meat eater problem (cf. Scott Weathers) and effects of technological progress on existential risk have been discussed in significant detail.[28]

With respect to considerateness, the effective altruism community has less often explicitly incorporated indirect effects into their analyses.[29] While a handful of steps toward increasing considerateness have been made (see, e.g., William MacAskill on the Centre for Effective Altruism's guiding principles), it may be argued that this has led to a relative neglect of these effects in practice as well.

In this article, we have taken an initial stab on these questions. We have argued that the indirect effects of considerateness are amplified when you act as a member of the effective altruism community. This means that though considerateness carries costs, and though there are some circumstances in which you should not be considerate, overall the positive effects of considerateness turns out to be surprisingly strong.

The extent to which this entails high levels of considerateness partly depends on how demanding morality is: how much it requires us to give up for the sake of others. Some ethical theories (e.g., many versions of consequentialism) demand that we take good opportunities to help others if we can do so at comparatively small cost for ourselves. For instance, they demand that we donate significant parts of our income to distant strangers. It seems plausible that such theories similarly demand high levels of considerateness. Indeed, it is plausible that such theories do not converge with common sense morality, but that they demand us to be significantly more considerate than common sense morality has it.[30]

Our analysis is quite cursory, however, and several caveats are warranted. First, future research and experiences are likely to uncover additional considerations of relevance for how considerate effective altruists ought be be. This means that we should emphasize the tentativeness of our conclusions in favor of high levels of considerateness. Second, one should be careful not to draw too quick conclusions regarding particular behaviors from our general argument for considerateness. It could very well be that some considerate behaviors are actually not that important. Others could be underrated.

[28] Although we think there is still room for valuable work to be done investigating these issues further.
[29] As we saw in footnote 11, there have a been some informed comments, though.
[30] There are also outside view-reasons to reject strong convergence with common sense morality on how considerate we ought to be. The premises and styles of reasoning employed by, e.g., consequentialism and common sense morality, are so different that strong convergence by accident seems implausible (see Gregory Lewis and John Halstead). To justify strong convergence, we would need some mechanism which would explain it. Some such mechanisms have been propounded. One of them is that following common sense norms is simple and cheap (see, e.g., Roger Crisp, (1992), p. 159). Another is that our moral intuitions over time are shaped by consequences. For instance, Joshua Greene (2002, pp. 340-342) has argued that common sense intuitions about drunk driving have converged with utilitarian judgements. However, none of these mechanisms seem sufficiently powerful to justify a strong convergence.

With those caveats in mind, let us look at some examples of implications of our current beliefs for how effective altruists ought to behave:

- We suspect that effective altruists should be much more intellectually honest than common sense morality requires you to be. On this view, just abstaining from falsehoods isn't enough. On the contrary, you should go further and, e.g., clearly state important counter-arguments to your views.

- We suspect that effective altruists should be much more collaborative than common sense morality requires you to be. For instance, it may be that you should spend time sharing advice with other community members, even if this brings few rewards to you.

- We suspect that effective altruists should be more modest than common sense morality requires you to be. For instance, it could be that effective altruists should go much further to avoid ostentatious displays of status than common sense morality requires.

If this is right, the effective altruism community should also probably provide incentives to its members to behave considerately: e.g., by rewarding considerateness.

The indirect effects of considerateness and its absence may be so important that they in some cases could swamp the direct effects of work on standard effective altruist causes. It is not outlandish to think that charity evaluators may have a greater impact through raising the level of intellectual honesty and norms of acceptable reasoning, than through helping the causes they are officially focused on.

The community should consider to do more work directly aimed at the positive effects of considerateness. Some such work is already being carried out, notably on community reputation and community social capital (Julia Wise (2016a), (2016b), (2016c), Owen Cotton-Barratt (2014), William MacAskill, Jeff Kaufman). Less work has been done on improving societal social capital through, e.g., norm-setting (though see Julia Galef's update project).[31] However, it could very well be that such norm-setting (e.g., of epistemic norms) is so valuable that it should be its own cause.

Further students of considerateness might help by exploring its facets, focusing on more specific behaviors such as honesty, openness, friendliness, and cooperativeness. The community might benefit from having psychologists, sociologists and other relevant experts contribute to this discussion. To the extent that we should judge the moral value of actions based on outcomes, the question of what norms and other decision rules to live by is, as

[31] The fourth set of effects concerns individuals' reputation within the community. It could be argued that this is best left to individual members.

Toby Ord (2009, p. 12) has pointed out, an empirical question. It deserves significant further study.

Bibliography

Alexander, Scott (2014a). Living by the sword.

Alexander, Scott (2014b). In favor of niceness, community, and civilization.

Arnold, Raymond (2017). Project Hufflepuff: Planting the Flag.

Askell, Amanda (2017). Act utilitarianism: criterion of rightness vs. decision procedure.

Baggini, Julian (2016). If you want people to trust you, stick to commonsense morality.

Blume, Hannah (2010). A Suite of Pragmatic Considerations in Favour of Niceness.

Christiano, Paul (2014). On Progress and Prosperity.

Christiano, Paul (2016). Integrity for consequentialists.

Crisp, Roger (1992). Utilitarianism and the Life of Virtue. *The Philosophical Quarterly*, 42, pp. 139-160.

Cotton-Barratt (2014). Keeping the effective altruism movement welcoming.

Cotton-Barratt (2015). How valuable is movement growth?

Dostoevesky, Fyodor (1866/1956). *Crime and Punishment.* (Random House.)

Everett, Jim, David Pizarro, and Molly Crockett (2016). Inference of trustworthiness from intuitive moral judgments. *The Journal of Experimental Psychology: General*, 145, pp. 772-787.

Galef, Julia. The Update Project.

Halstead, John (2017). Comment on Amanda Askell's "Act utilitarianism: criterion of rightness vs. decision procedure".

Greene, Joshua (2002). *The terrible, horrible, no good, very bad truth about morality and what to do about it.* (Ph.D. dissertation, Princeton University.)

Kaufman, Jeff (2016). Concerns with Intentional Insights.

LaVictoire, Patrick (2011). Consequentialism Need Not Be Nearsighted.

Lewis, Gregory (2016). Beware suspicious and surprising convergence.

MacAskill, William (2017). Introducing CEA's Guiding Principles.

Mill, John Stuart (1838). "Bentham", in John M. Robson (ed.) Collected Works (vol. X), (Toronto: University of Toronto Press), 1985, pp. 75–116.

Ord, Toby (2009). *Beyond Action: Applying consequentialism to decision making and motivation*. (D. Phil. dissertation, University of Oxford.)

Rawls, John (1971). *A Theory of Justice.* (Belknap).

Rowling, J. K. (1997-2007). *The Harry Potter series.* (Bloomsbury Publishing.)

Scheffler, Samuel (2003). *The rejection of consequentialism.* (Oxford University Press, Oxford.)

Schubert, Stefan and Ben Garfinkel (2017). Hard-to-reverse decisions destroy option value.

Sidgwick, Henry. (1907). *The methods of ethics*, 7th edition, (London: Macmillan).

Tomasik, Brian (2013/2016). Why honesty is a good policy.

Yudkowsky, Eliezer (2008a). Ends Don't Justify Means (Among Humans).

Yudkowsky, Eliezer (2008b). Protected From Myself.

Weathers, Scott (2016). The Meat Eater Problem: Developing an EA response.

Wise, Julia (2016a). Against segregating EAs.

Wise, Julia (2016b). Guidelines on depicting poverty.

Wise, Julia (2016c). Making EA groups more welcoming.