Ergo AN OPEN ACCESS JOURNAL OF PHILOSOPHY

Causation, Cluelessness, and the Long Term

SIMON FRIEDERICH
University of Groningen

Agents are said to be "clueless" if they are unable to predict some ethically important consequences of their actions. Some philosophers have argued that such "cluelessness'' is widespread and creates problems for certain approaches to ethics. According to Hilary Greaves, a particularly problematic type of cluelessness, namely, "complex" cluelessness, affects attempts to do good as effectively as possible, as suggested by proponents of "Effective Altruism," because we are typically clueless about the long-term consequences of such interventions. As a reaction, she suggests focusing on interventions that are long-term oriented from the start.

This paper argues for three claims: first, that David Lewis' distinction between *sensitive* and *insensitive* causation can help us better understand the differences between genuinely "complex" and more harmless "simple" cluelessness; second, that Greaves' worry about complex cluelessness can be mitigated for attempts to do near-term good; and, third, that Greaves' recommendation to focus on long term-oriented interventions in response to complex cluelessness is not promising as a strategy specifically for avoiding complex cluelessness. There are systematic reasons why the actual effects of serious attempts to beneficially shape the long-term future are inherently difficult to predict and why, hence, such attempts are prone to backfiring.

Keywords: causation; cluelessness; longtermism; effective altruism; existential risks

1. Introduction

Sometimes it is impossible to confidently predict the consequences of our actions. This can make it difficult to decide which actions are the ones that we ought to perform from an ethical point of view. Indeed, *cluelessness* about our actions' consequences has been identified as a problem for consequentialism (Lenman 2000), the view that an action's ethical status is entirely determined by its consequences. But also non-consequentialist ethical frameworks are poten-

Contact: Simon Friederich <s.m.friederich@rug.nl>

https://doi.org/10.3998/ergo.7428

tially threatened by such cluelessness. According to Mogensen and MacAskill (2021) non-consequentialist frameworks, to the extent that they include some imperative to not cause significant harm, face a "paralysis argument." If we are clueless about the more distant consequences of our actions, including harmful ones, we must avoid *any* actions and are, in that sense, ethically paralysed.

Worries about cluelessness as ethically problematic arise from the perception that cluelessness is widespread. The idea is that almost all our actions plausibly have dramatic long-term causal consequences—e.g. they affect who will be born in the future—and that we are generally clueless about these. Greaves illustrates the idea:

> Suppose … that I pause on my way home from work, in order to help an old lady across the road. As a result, both she and I are in any given place—any given position on the pavement for the remainder of our respective journeys home, for instance—at different times …. As a result, we advance or delay the journeys of countless others, if only by a few seconds … At least some of these others were destined to conceive a child on the day in question, and if so, even our trivial influences on their day will affect, if not whether they conceive, then at least which particular child they conceive …. But once my trivial decision has affected that, it equally counts as causally responsible for everything the child in question does during his/her life …—and of all the causal consequences of all those things, stretching down as they do through the millenia. (2016: 314–315)

Greaves herself believes that cluelessness, though widespread, is not always problematic. Often it is "simple," and then it can be set aside by appeal to an unproblematic principle of indifference. However, she also identifies situations where cluelessness is "complex" and indifference does not apply. This is the case, as she sees it, when proponents of effective altruism try to identify which good causes, aimed at doing near-term good, are most cost-effective. The most attractive response according to her, motivated by "strong longtermism" (Greaves & MacAskill 2021), is to choose actions predominantly based on their expected long-term consequences (Greaves 2020).

Cluelessness is about unforeseeable consequences of our actions. In this paper, I suggest that whether cluelessness is "simple" or "complex" often (though perhaps not always) has to do with *how* our actions might have unforeseeable ethically significant consequences, notably, whether we would cause those consequences *sensitively* or *insensitively* (Lewis 1986; Woodward 2006; 2021) with respect to background conditions. Overall, I argue for the following three claims:

- The distinction between sensitive and insensitive causation can help us better understand characteristic features of situations of simple and complex cluelessness, respectively.
- Greaves' worry that complex cluelessness affects attempts to effectively do near-term good can be mitigated by differentially addressing it when spelled out as a worry about sensitively or insensitively caused unforeseeable consequences.
- Contrary to what Greaves suggests, "going long-term" and selecting actions primarily by focusing on their long-term expected consequences does not necessarily help with the problem of complex cluelessness and can sometimes make it worse.

The structure of the remaining sections is as follows: In §2, I review how Greaves motivates and characterizes the distinction between simple and complex cluelessness. In §3, I outline how the distinction between sensitive and insensitive causation can shed light on simple and complex cluelessness. In §4, I argue that complex cluelessness poses no principled problems for attempts to effectively do near-term good. In §5, I argue that, to the extent that we can perform actions with beneficial long-term consequences, we tend to have complex (near-) cluelessness, or at least "complex uncertainty." Section 6 summarizes the paper's main points.

2. Cluelessness, Simple and Complex

At the heart of the cluelessness worry is the thought that we are usually aware of only a tiny subclass of the causal consequences of our actions, typically those that concern the immediate future. Greaves outlines this idea as follows:

> The argument for [the cluelessness worry] stems from the observation that the relevant consequences include all consequences of the actions in question, throughout all time. In attempting actually to take consequences into account in practice, we usually focus on those effects—let us call them 'foreseeable' effects—that we take ourselves to be able to foresee with a reasonable degree of confidence. […] And while we are arguably correct in thinking that we are justified in being reasonably confident in our predictions of these effects, any choice of one act A_1 over another A_2 inevitably has countless additional consequences that our calculation takes no account of. A butterfly flapping its wings in Texas may cause a hurricane in Bangladesh; so too may my telling a white lie, refraining from telling that lie, moving or not moving my hand; a hurricane will

> certainly affect which other butterflies flap their wings or which other agents move their hands in which ways; and so the effects will ripple down the millennia. (2016: 312f.)

Greaves considers and dismisses two objections against the idea that cluelessness is widespread. The first objection, traceable to Moore (1903) and Smart (1973), is that an action's causal consequences usually peter out towards the future, similar to "ripples on a pond." Greaves disputes this objection by responding as quoted in §1, pointing out how our everyday actions, and indeed the details of how we perform them, influence the very identities of future people. This, according to her, makes it highly implausible that the causal consequences of our actions tend to become ever more insignificant towards the future and thus undermines the "ripples on a pond" objection.

The second objection (Cowen 2006) hypothesizes that the consequences of different candidate actions balance out in the long run in that the bad and good unforeseeable consequences of one candidate action will plausibly be matched by equally bad and good unforeseeable consequences of other candidate actions (Dorsey 2012: 6–7). Greaves argues that this objection is implausible as well. The ethically significant differences between different futures—e.g. which individuals will exist—can be large, and there is no reason to assume that they will balance out.

Greaves suggests another way of alleviating the cluelessness worry, which is more promising. It involves regarding an action's (rationally) *expected* consequences as relevant to its ethical status, not its *actual* consequences. If an agent performs an action that she can reasonably expect to have a morally good outcome, performing the action may be what she ought to do even if, in the end, that outcome, by coincidence, did not materialize and/or a bad outcome materialized instead. If I have a choice between two actions A_1 and A_2, their long-term actual consequences may well be radically different, but their *expected* consequences may well differ only in the near future and be exactly the same in the longer term. Greaves calls this type of cluelessness *simple cluelessness* and characterizes it as follows:

> [C]onsider any possible but unforeseeable future effect $E_1 \mapsto E_2$ [i.e. E_1 occurring instead of E_2] that might … result from my decision to perform act A_1 rather than A_2. For sure, it is possible that: if I did A_1 then E_1 would result and if I did A_2 then E_2 will result (in symbols: $A_1 \mathbin{\Box\!\!\rightarrow} E_1$ & $A_2 \mathbin{\Box\!\!\rightarrow} E_2$). Still, there is no particular reason to think that the correlations between my possible actions and these unforeseeable effects will be that way round, rather than the opposite ($A_1 \mathbin{\Box\!\!\rightarrow} E_2$ & $A_2 \mathbin{\Box\!\!\rightarrow} E_1$). It seems plausible, in that case, that given any credence function that it is rationally permis-

> sible for me to have at the time of decision, my credence in the second correlation hypothesis is exactly equal to my credence in the first correlation hypothesis. But if this is true for all unforeseeable possible effects $E_1 \mapsto E_2$, then the contribution of those unforeseeable effects to the difference in the expected values of A_1 and A_2 is precisely zero. (2016: 317)

Situations of simple cluelessness, according to this characterization, would be those where, if acts A_1 and A_2 can be performed with candidate unforeseeable consequences E_1, E_2, it is just as likely that A_1 will end up causing E_1 and A_2 will end up causing E_2 as it is likely that, vice versa, A_1 will end up causing E_2 and A_2 will end up causing E_1. Indeed, it seems plausible that, in such situations the unforeseeable consequences can be discarded for the purposes of deliberation and one can act purely based on the foreseeable ones.

However, this characterization of simple cluelessness is a very demanding condition and overly restrictive in the light of Greaves' own response to the "ripples on a pond" objection to the cluelessness worry. To see this, consider again the example mentioned in §1, where one considers helping an old lady across the street, supposedly an example of simple cluelessness. For the probabilities assigned to many unforeseeable consequences it may indeed not make any difference whether one helps the lady or not, e.g. to whether one will still be alive in thirty years, will be married, happy etc. But, for instance, for the probabilities assigned to whether, five years later, one will have befriended the lady, met her children, or have fallen in love with her son it will make a difference. This observation suggests that what intuitively makes Greaves' example of helping an old lady across the street into a case of simple cluelessness is that we have no reason to think the unforeseeable consequences of A_1 will be substantially *better* than those of A_2 (and vice versa), not that the probabilities of all those unforeseeable consequences are exactly the same.

We can adapt Greaves' criterion of "complex" cluelessness (quoted below) and incorporate the above suggestion, to get the following attractive criterion for *simple cluelessness*:

> For some pair of actions of interest A_1 and A_2,
>
> (SC1) We have *no* reason to think that the unforeseeable consequences of A_1 would systematically tend to be substantially better than those of A_2;
>
> (SC2) We have *no* reason to think that the unforeseeable consequences of A_2 would systematically tend to be substantially better than those of A_1.

This criterion, like Greaves' criterion of complex cluelessness to be quoted below, refers to how good the consequences of actions A_1 and A_2 are. To the extent that

this makes the criterion a natural fit only with consequentialism and limits the scope of the considerations that follow to consequentialism, one may read them as presupposing consequentialism.

It is not difficult to find examples of cluelessness where SC1 and SC2 do not apply. For example, when a couple deliberates whether to have children, they may have good reasons to expect that the unforeseeable consequences of having children will be mostly good, and they may have reasons to expect that the unforeseeable consequences of having children will be mostly bad, but find it very difficult to weigh those reasons. Similarly, when a politician deliberates whether to go to war with a neighboring country, she may have good reasons to expect that the unforeseeable consequences of going to war will be mostly good and good reasons to expect that the unforeseeable consequences of going to war will be mostly bad, but weighing the pros and cons of military aggression in the light of the overall unforeseeable consequences may be extremely difficult for her. Greaves characterizes situations like these as cases of *complex* cluelessness. According to her, plausibly, as it seems to me, in such situations the following three conditions hold (Greaves 2016: 323):

> For some pair of actions of interest A_1, A_2:
>
> (CC1) We have some reasons to think that the unforeseeable consequences of A_1 would systematically tend to be substantially better than those of A_2;
>
> (CC2) We have some reasons to think that the unforeseeable consequences of A_2 would systematically tend to be substantially better than those of A_1;
>
> (CC3) It is unclear how to weigh up these reasons against one another.

Simple cluelessness as characterized in terms of SC1 and SC2 is meant to be the natural complement to complex cluelessness so characterized.

Complex cluelessness, according to Greaves, raises problems in particular for adherents of Effective Altruism, who focus on maximizing expected good with their limited resources. Effective Altruism has been influential in championing an approach to selecting charities to which one may donate based on expected number of lives saved (or expected quality-adjusted life-years saved). Cost-effectiveness is empirically measured using randomized control trials. According to Greaves, such trials measure only a small part of the causal consequences of the donations. The donations likely have further, unforeseeable, consequences, and these plausibly dominate the overall value created or destroyed by the donations. Worryingly, those consequences could not only affect how well the donations fare

compared with each other. They could even make the net effect of donations—or, more generally, attempts to do good—negative, i.e. overall harmful.

Greaves illustrates her considerations with the case of a child saved by donations to the *Against Malaria Foundation*, one of the charities that score highest in terms of cost-effectiveness according to independent charity evaluator GiveWell (2021). As she argues[1]:

> Averting the death of a child … has knock-on effects that have not been included in this calculation. … [T]he intervention in question also has systematic effects on others, which latter (1) have not been counted, (2) in aggregate may well be far larger than the effect on the child himself of prolonging the child's life, and (3) are of unknown net valence. The most obvious such effects proceed via considerations of population size. … Assuming for the sake of argument that the net effect of averting child deaths is to increase population size, the arguments concerning whether this is a positive, neutral or negative thing are complex. But, callous as it may sound, the hypothesis that (overpopulation is a sufficiently real and serious problem that) the knock-on effects of averting child deaths are negative and larger in magnitude than the direct (positive) effects cannot be entirely discounted. Nor (on the other hand) can we be confident that this hypothesis is true. And, in contrast to the 'simple problem of cluelessness', this is not for the bare reason that it is possible both that the hypothesis in question is true and that it is false; rather, it is because there are complex and reasonable arguments on both sides, and it is radically unclear how these arguments should in the end be weighed against one another. (Greaves 2016: 324–325.)

In §4 I argue that Greaves' worries about complex cluelessness as potentially undermining the attempts of effective altruists and others to do near-term good can be mitigated. To prepare the ground, §3 takes a fresh look at the distinction between simple and complex cluelessness and suggests that the distinction between sensitive and insensitive causation can help us better understand typical features of situations of simple and complex cluelessness, respectively.

3. The Role of Sensitive and Insensitive Causation in Simple and Complex Cluelessness

In this section I argue that simple and complex cluelessness can often be related to whether the agents at issue have reasons to believe that their actions may

1. See (Mogensen 2021a) for similar considerations.

insensitively (as opposed to *sensitively*) cause ethically significant unforeseeable consequences. If they have such reasons, the cluelessness is typically (but, as we will see, apparently not always) complex. Otherwise it is typically simple.

The distinction between sensitive and insensitive causation refers to the *background conditions* with respect to which causal relations obtain. A causal relation is sensitive if even rather small changes in background conditions disrupt it. Otherwise it is insensitive. The distinction is investigated in-depth by Woodward (2006; 2021: ch. 6), who credits Lewis (1986) and introduces it as follows:

> The counterfactual dependence of effects on their causes is such an obvious feature of many examples of causation that it is easy to miss the fact that there is another feature having to do with counterfactual structure that plays an important role in such examples. This feature has to do with the sensitivity of the causal relationship (and, more specifically, the sensitivity of certain of the counterfactuals associated with it) to changes in various other factors. Broadly speaking, a causal claim is sensitive if it holds in the actual circumstances but would not continue to hold in circumstances that depart in various ways from the actual circumstances. A causal claim is insensitive to the extent to which it would continue to hold under various sorts of changes in the actual circumstances. The sensitivity of counterfactuals is understood similarly. (2006: 2)

When we think of causation, we usually have in mind insensitive causal relations. An example of insensitive causation is causing a lamp to light up by hitting the light switch. This causal link holds in a large variety of conditions, independently of the weather outside, of the moods of people in the room, or the color of the light bulb. It does, however, depend on electricity provision and the light bulb being intact, and in that sense it is not completely insensitive. Another example of (somewhat) insensitive causation is someone's smoking causing them to develop lung cancer. Smoking causes cancer in humans with a variety of genetic and environmental predispositions, not just humans with, say, highly specific personal histories.

Examples of sensitive causation, on the other hand, can be found in the scenarios put forward to illustrate simple cluelessness. In fact, Lenman motivates the cluelessness worry by pointing out that "some causal systems are known to be *extremely sensitive* to very small and localized variations or changes in their initial conditions" (2000: 347, emphasis mine). To return to Greaves' example that involves her helping an old lady across the street: Suppose that, as a consequence of how this affects traffic, two people meet on the same day who end up having a child, Max, who is alive thirty years from now. The counterfactual conditional "Had Hilary not helped

the old lady across the street, Max would not be alive thirty years from now" is plausibly true. However, to the extent that there really is a causal link between these two events, it holds only because an enormous variety of other facts occur: the ones that, however indirectly, had to be in place, besides Hilary's helping the old lady across the street, to result in Max's conception, birth, and staying alive.

One can construct situations of simple cluelessness that do not fit this pattern, i.e. situations of simple cluelessness where an agent has reasons to expect that their actions will cause ethically significant or otherwise important unforeseeable consequences *insensitively* with respect to background conditions. A situation of this type might be the following: Alice tries to catch a boat for an important overseas trip and comes to a fork in the road. She does not know which of the two options leads her to the harbour, there is no signpost, and the entire situation seems symmetric between the two roads, so she has no reason whatsoever to prefer either road. Accordingly (to the extent that it ethically matters whether she catches the boat), she experiences simple cluelessness in that the situation fulfills SC1 and SC2. Still, whether she will catch the boat depends insensitively on which option she chooses, and she is aware of this, though not of which option will make her catch the boat and which will make her miss it. It is interesting to note that, for this scenario to exemplify simple cluelessness, it has to be constrained—perhaps somewhat unrealistically—in that Alice must not have any reason whatsoever to believe that either road is the correct one, notably, no competing reasons drawing in opposite directions.

To the extent that cases like these are somewhat special and artificial, and that it is otherwise characteristic for simple cluelessness to arise from the fact that many phenomena are, as Lenman puts it, "extremely sensitive to very small and localized variations or changes in their initial conditions," viewing it through the lens of sensitive causation helps us further appreciate why it has no great ethical significance. For if an action A causes some outcome E in a way that is extremely sensitive with respect to background conditions, then many further, otherwise highly contingent details of the background conditions are just as crucial for E to occur as A itself is. Focusing on A as "the" cause of E is, in that sense, rather arbitrary.[2] Consider again the case of Max, whose parents

2. An exception to this statement might be situations where a systematic and successful effort is undertaken by competent agents to establish and stabilize highly specific background conditions. Friederich and Mukherjee (2021) suggest that high energy physics accelerator experiments can be seen along these lines as stabilizers of very specific background conditions that allow researchers to probe highly sensitive causal relations and, thereby, identify elementary particles with short lifetimes. Absent such highly special circumstances, one possible reaction to (highly) sensitive causal relations, considered by Woodward (2021: 274) and in line with some people's intuitions regarding the nature of causation, is to regard them as not genuinely causal at all. On such a view, the alleged unforeseeable candidate consequences in situations of simple cluelessness tend to actually not be causal consequences, so there is actually no cluelessness at all in these situations.

met due to how Hilary's helping an old lady across the street influenced traffic on that day. Clearly, it is entirely arbitrary to consider Hilary's helping the old lady as "the" cause of Max's being conceived (and being alive thirty years from now). Myriad other factors that also influenced Max's parents and the traffic on that day, including countless actions by third actors such as traffic participants, could just as well be regarded as "the" cause of Max's existence, most of them equally arbitrarily. Holding agents ethically responsible for sensitively caused consequences of their actions would lead to a hopeless proliferation of ethical responsibilities and, thereby, ultimately undermine the very notion of ethically assessing actions in terms of their consequences.[3]

Things tend to be very different in situations of complex cluelessness, where, by CC1, CC2, and CC3, an agent has conflicting reasons that are difficult to weigh as to which of their candidate actions A_1 and A_2 will have better unforeseeable consequences. We can expect that in many such situations the agent has at least a rough idea of *how and why* A_1 and A_2, respectively, might end up having better or worse unforeseeable consequences. And for this to happen, the ways in which, as the agent has reasons to assume, their actions might end up causing ethically significant unforeseeable consequences will have to be at least somewhat predictable, which means that, absent highly specific "stabilizers" of background conditions (see fn. 2), the causation involved will have to be at least somewhat *insensitive* with respect to background conditions.

To illustrate this with an example, consider again the couple that deliberates whether to have any children and suppose that they contemplate this question predominantly in terms of how having children will influence their own future wellbeing. (Analogous considerations can be applied to other aspects of the question that they may also contemplate.) Rationally, they assume that the ways in which having children will make them happy or unhappy will be similar to the ways in which having children makes other couples happy or unhappy, so they are aware of ways in which having children may have good or bad unforeseeable consequences for their future wellbeing. For instance, having children might make the couple experience more of a sense of purpose in life. Or having children might create financial worries for them. These ways in which having children might influence their future wellbeing are, to some extent, typical, and they correspond to *insensitive* causes of future wellbeing. Nevertheless, it may be hard or even impossible for the couple to comparatively weigh the reasons for believing that having children will be good for their future wellbeing and those for believing that it will be bad for their future wellbeing. As a result, they may

3. This observation might also go some way towards deflecting the paralysis argument by Mogensen and MacAskill (2019) mentioned in §1. If agents in general cannot be held responsible for sensitively caused consequences of their actions, the force of this argument diminishes. Investigating this suggestion in detail, however, is beyond the scope of this paper.

end up in a state of *complex cluelessness* about the unforeseeable consequences of having children, satisfying all the criteria CC1, CC2, and CC3.

Two remarks before closing this section:

First, the distinction between sensitive and insensitive causation is inherently vague and gradual, for there is simply no general objective answer to what counts as "specific" background conditions nor to what counts as a "wide range" of background conditions. But this does not make the distinction between sensitive and insensitive causation unimportant—Woodward (2006; 2021) gives many reasons why it is important—nor does it create any problems for connecting it with the one between simple and complex cluelessness, which is also somewhat vague and gradual.

Second, for the purposes of the following section, it is important to note that the way in which complex cluelessness makes the choice between different actions challenging is due to condition CC3: that it is unclear how to comparatively weigh the reasons for believing that A1 will have better consequences and the reasons for believing that A2 will have better consequences. This can be true even if, in expectation, the unforeseeable consequences A_1 and A_2 are not exactly equally good. Suppose, for example, that one is aware of a complex web of causal paths emanating from A_1 and A_2, all rather insensitive with respect to background conditions, via which A_1 and A_2 can have unforeseeable consequences. Suppose further that one tentatively and defeasibly expects A_1 to have overall better consequences than A_2. It doesn't matter much whether one regards the considerations favouring A_1 and A_2 as not exactly in balance (but still hard to weigh as cases of genuine complex cluelessness), or whether one prefers a different label for them (say, "complex near-cluelessness" or "complex uncertainty"). What makes complex cluelessness ethically problematic is not that considerations favoring A_1 and A_2 are *exactly* in balance—but that it is unclear how to weigh them.

4. Doing Good and Complex Cluelessness

Having explored the role that sensitive and insensitive causation characteristically play in situations of simple and complex cluelessness we are in a good position to address Greaves' worry that complex cluelessness is problematic for attempts to effectively do good. The core of that worry, to recall, is that, for actions that do a large amount of short-term good we have reasons to expect that their overall net-effect is dominated by their long-term consequences and that it is very difficult to anticipate whether those consequences will, in the aggregate, be mostly good or bad. Specifically, Greaves worries that donating to charities

such as the Against Malaria Foundation, which improve health outcomes and save lives in developing countries, might have mixed overall consequences via its impact on population size and conceivably result in an overall bad outcome.

Armed with the distinction between sensitively and insensitively caused consequences, we can address this worry by making it more precise and concrete in three different ways and addressing it in each. As developed in the first way, the worry concerns candidate causal effects that are caused *highly insensitively* with respect to background conditions; in the second, the worry concerns candidate causal effects that are caused *somewhat insensitively* with respect to background conditions; and in the third, the worry concerns candidate causal effects that are caused *sensitively* with respect to background conditions. I will argue that, as fleshed out in the first two directions, the worry must be taken seriously but can be met, whereas, as fleshed out in the third direction, it concerns simple cluelessness and can be neglected.

The first direction in which Greaves' worry can be fleshed out is as a general concern about a large class of attempts to do good, namely attempts to alleviate and eliminate poverty in developing countries and improve health outcomes there. The concern is that such attempts will generally backfire because, if successful, they will ultimately increase ecological pressure, typically by increasing population numbers, and in the end contribute to triggering ecological collapse. "Neo-Malthusian" thinkers such as William Vogt (1948) and Paul Ehrlich (1968) advocated dark views along these lines. (Whether they would have endorsed the position as just sketched is a separate question)

While limited defenses of such dark views may still be viable, a strong response to them is available, appealing to the recent track record of key development indicators (Rosling 2018). Life expectancy, share of people in extreme poverty, child mortality and many other indicators have significantly improved since Neo-Malthusians made their pessimistic predictions, and catastrophes on the scale predicted have not occurred. Technological progress, notably in agriculture, seems to have played a key role in averting ecological collapse. The expert consensus nowadays seems to be that either continued technological progress, a more equitable distribution of resources between developed and developing countries, or a combination of these can realistically stave off Malthusian collapse while development indicators continue to rise. In any case, arguments for and against the Malthusian worry, in the light of evidence that has been accumulated in decades, can be studied, and it is perfectly possible to donate based on one's considered view of the merits of this worry.

The second direction in which Greaves' worry can be spelled out is as a concern that the work of the selected charity might have adverse unintended effects in the respective social, cultural, and economic conditions where the charity operates.

To the extent that these adverse effects supposedly occur systematically in those conditions, the candidate causal links from donations to these effects are insensitive with respect to background conditions. And to the extent that these effects are difficult to predict, they could conceivably give rise to complex cluelessness about the intervention. At the same time, because the candidate adverse effects supposedly occur systematically in the social, cultural, and economic conditions at issue, one can study whether they indeed occur and, if so, to what extent.

Incidentally, such studies have been performed at significant scale for the *Against Malaria Foundation*. The meta-charity GiveWell analyzed a variety of possible paths along which its work might lead to adverse effects[4]; first, by increasing insecticide resistance in malaria-transmitting mosquitoes; second, by distorting local markets for anti-insect nets; third, by leading to problems of unequal treatment in targeted communities; and fourth, by diverting skilled workers from activities where they would be needed more urgently. The analysis concludes that none of these candidate adverse effects are likely to outweigh the beneficial direct effects of the donated nets.

The third direction in which Greaves' worry about complex cluelessness and attempts to efficiently do good can be spelled out is as the idea that this specific charity, or perhaps just *this* specific donation, in *these* specific background conditions, might end up having overall adverse effects because of how it might interact with other events. For example, a bednet donated by me might save the life of someone who, years later, as a result of many coincidences, becomes a brutal military leader who causes untold suffering. Worrying about potential consequences of one's actions along such lines means worrying about effects that are caused in a *highly sensitive* manner. In the light of the considerations in the previous section, such worries can be set aside as ethically irrelevant instances of simple cluelessness.

5. Going Longterm?

Greaves' preferred reaction to the problem of complex cluelessness is to focus on actions that aim "to beneficially influence the course of the very far future of humanity and more generally of the planets in the universe" (2020: §5). This response is motivated by "strong longtermism" (Greaves & MacAskill 2021)—the view that the ethical status of an action is determined mostly by its expected *long-term* consequences. Strong longtermism is plausible, according to Greaves and MacAskill, because, first, the future is potentially very "big" (they see 10^{24}

4. GiveWell (2021: § "Are there any negative or offsetting impacts?"). No such section is included in the, at the time of writing, most up-to-date report on AMF.

humans as a conservative estimate of future potential) and, second, we have reliable and effective means to influence it.

In (Greaves & MacAskill 2021) complex cluelessness is considered as a potential *worry* for longtermism. But in a lecture at the 2020 Effective Altruism student summit Greaves suggests choosing actions based on expected long-term consequences explicitly as a *response* to complex cluelessness. Her reasoning seems to be that, by strong longtermism, expected long-term consequences typically determine an action's overall ethical status anyway, so focusing on those consequences from the very start (and not just as an afterthought) will in general allow one to identify the ethically best actions. The specific actions that Greaves recommends are attempts to reduce risks of human extinction and risks of locked-in dystopian states of affairs.

In the rest of this section I argue that even if (as can reasonably be doubted) longtermism is defensible in our specific historical circumstances, while there can be good reasons to perform certain long term-oriented actions, "going long-term" is not promising as a general overarching strategy specifically for avoiding complex cluelessness.[5]

How can we hope to beneficially influence the long-term future? Clearly we cannot *directly* affect the world a hundred years from now. What we can potentially do is to identify reliable causal chains and, by exploiting them, influence the long-term future *indirectly*. Inspired by the language of causal modelling (Pearl 2009) we can put this by saying that we can directly intervene on variables that describe the world today and hope that, via intermediate variables, the effects of such interventions will propagate to variables that describe the long-term future roughly as intended.

With the challenge of influencing the long-term future laid out in this way, it seems unclear how a randomly picked human from arbitrary historical or geographical circumstances could realistically have hoped to ever so slightly increase the long-term prospects for humanity in any systematic way. It seems difficult to specify any "variables" that, say, a Pleistocene hunter-gatherer or a medieval peasant could have intervened on in order to systematically, however slightly, improve the expected wellbeing of humans several hundreds or thousands of years in her future.

However, our historical circumstances may be special. For us, unlike for the vast majority of agents in the past and perhaps the future, actions might be available to beneficially influence the long-term future in ways insensitive to background conditions. Two candidate features of our specific point in history might make this the case:

5. See (Tarsney 2023) for further considerations on the "epistemic challenge to longtermism," which arises from the difficulty to predict the long-term future.

First, risks to the very survival of humanity may currently be much higher than they have been at any previous point in human history. Ord (2020) argues for this *precipice hypothesis*, highlighting risks from nuclear war, climate change, environmental damage, artificial pandemics, and unaligned artificial intelligence. To the extent that actions are available to humans today that can be expected to significantly and enduringly reduce these risks,[6] we can thereby beneficially affect long-term expectations in ways unavailable to previous generations.

Second, and more speculatively, there is the possibility that progress in transformative artificial intelligence (or some other technology) will in a couple of decades or centuries lead to the "lock-in" of values governing our civilization for millions or even billions of years (MacAskill 2022: ch. 4). According to this idea, the way in which ever more powerful artificial intelligence is shaped by humans today will permanently fix key contingent features of human civilization as it coexists with artificial general intelligence similar to how Pleistocene evolution shaped key psychological features of humans. If this *lock-in hypothesis* holds and humans today can identify actions that influence— in ways insensitive to background conditions—how this lock-in will play out, they can thereby beneficially affect long-term expectations in ways available neither to previous nor future generations. If both the precipice hypothesis and the lock-in hypothesis are true, a particularly valuable *existential safety lock-in* might be achievable.

Assessing whether the precipice hypothesis and/or the lock-in hypothesis hold is beyond the scope of this paper. But if at least one of them is true, or even both of them, then what is potentially at stake in our actions with regard to the long-term future is so enormous that to focus on actions with the best expected long-term consequences seems at least defensible. How such long-term oriented actions will compare with each other and with near-term oriented actions, e.g. ones aimed at improving health in developing countries, may still depend on one's specifically preferred framework of normative ethics—e.g. whether one prefers total utilitarianism, average utilitarianism or some deontological framework[7]—but at least some frameworks of normative ethics will plausibly see the focus on long-term consequences as ethically justified.

However, importantly, the move of "going long-term" as suggested by Greaves does not allow one to bypass the problem of complex cluelessness. To see this, suppose that the precipice hypothesis is true and that there are indeed

6. See (Thorstad 2023) for persuasive considerations about why the precipice hypothesis justifies prioritizing existential risk mitigation over alternatives strategies for doing good only if one is confident that historically highly unusual circumstances obtain, namely, ones in which one can rationally hope to significantly and lastingly reduce existential risks ("time of perils hypothesis").

7. Mogensen (2021b) argues that long-term oriented actions will generically be superior by the standards of arbitrary versions of utilitarianism.

ways in which we can reasonably hope to reduce existential risks, via causal paths that are somewhat insensitive with respect to background conditions. Then the interplay of these paths and relevant others is plausibly complex enough to bring complex cluelessness in.

Take the example of research in solar geoengineering with the goal of potentially using this technology to mitigate the risks from climate change by deflecting some solar radiation that earth is receiving (Reynolds 2019). The causal path through which this research might help mitigate climate change can be sketched as the following sequence; (1) improving our understanding of the feasibility of solar geoengineering; (2) informing policy makers of it as an option; (3) creating concrete plans to implement it; (4) actually implementing it by aerosol injection in the stratosphere; (5) by these aerosols deflecting some solar radiation; finally resulting in (6) atmospheric temperatures being lower than they would otherwise be.

Any of the "variables" in this sketched causal path is plausibly connected with other potentially relevant "variables," and some of these will influence the amount of warming that ultimately occurs. Notably, when policy makers start to regard solar geoengineering as a live option, this may (negatively) influence ambition to reduce emissions, thereby leading to overall higher emissions. As a net-result, research into solar engineering could end up increasing rather than decreasing overall warming. Another causal factor relevant to human welfare is that aerosols will not only reduce temperatures but also affect precipitation. These additional relevant causal paths may well be active for a large variety of background conditions and hence be insensitive. Yet estimating their combined effect, and thereby the overall outcome of solar engineering deployment with respect to human welfare and ecosystems, is extremely difficult. While, in the case of the Against Malaria Foundation, one can investigate candidate adverse side-effects by studying the charity's past record, performing analogous investigations for research into solar geoengineering is difficult. Limited-scale trial runs could already interfere with global morale to phase out emissions. It will therefore plausibly remain difficult to weigh the reasons for and against taking steps towards actually performing solar geoengineering.

Arguably, the complex interplay of distinct and individually insensitive causal paths with a hard-to-predict overall outcome is the rule, not the exception, in existential risk mitigation. There might be exceptions—"no regrets" actions to reduce existential risks that do not give rise to complex uncertainty—but it is surprisingly hard to come up with promising candidates. The very act of drawing attention to some specific underappreciated existential risk is risky: It can alert malevolent actors of options to effectively do large-scale harm that they might otherwise not have become aware of.

It is not difficult to find more examples of how realistic initiatives to reduce existential risks or beneficially shape value lock-in can backfire: For example, initiatives to reduce nuclear weapons stockpiles can end up increasing the probability of nuclear war by decreasing retaliation capacity and thereby making first strikes attractive; research on function-enhanced pathogens to guard against natural or artificial pandemics can cause laboratory outbreaks or inform bioweapons development; and attempts to beneficially shape the development of advanced artificial intelligence by being at the forefront of its development and outcompeting less safety-concerned competitors could increase the risks from artificial general intelligence by causing it to be developed before a robust and applicable consensus about how to avoid its *catastrophic misalignment* (Russell 2019, Christian 2020).

Even innocent-looking suggestions such as "replace fossil fuels for electricity with solar and wind power to mitigate climate change" are not as innocent as they look. To make concrete progress towards implementing this suggestion, a particular mechanism will have to be chosen—say, technology-specific mandates or subsidies, or technology-neutral mandates or subsidies—and an electricity market design or some alternative to an electricity market. There is no obvious no-regrets option here. For instance, technology-specific mandates or subsidies, which exclude more controversial technologies such as nuclear energy or fossil fuels with carbon capture and storage, may be most popular and offer large short-term emission reductions. But strategies that privilege certain technologies may actually make it very costly to eliminate emissions entirely (Jenkins et al. 2018) and, thereby, ultimately create novel roadblocks for decarbonization. The heat with which proponents of different ways of mitigating climate change attack each other in the public sphere illustrates how unobvious it is which initiatives are effective and which counterproductive. There is no reason to suppose that questions about how to best mitigate other existential risks are easier to answer.

6. Conclusion

Cluelessness about the consequences of our actions can complicate their ethical assessment. Some of this cluelessness—"simple cluelessness"—is superficial and poses no great ethical problems. I have argued here that, characteristically (though not universally), in situations of simple cluelessness, the causal paths from our actions to their unforeseeable consequences are highly sensitive with respect to background conditions and hence neither predictable nor controllable for practical purposes, and, due to this—not ethically relevant.

But some of the cluelessness we are facing is "complex" and genuinely complicates the ethical assessment of our actions regarding their consequences. In a typical situation of complex cluelessness, the causal paths from our actions to unforeseeable consequences are rather insensitive with respect to background conditions, but the interplay of these paths is complex and difficult to predict. In such a situation, getting more evidence and reflecting more may lead to a very different assessment, potentially even a complete reversal of one's judgment as to which action one ought to take. This contributes to making complex cluelessness problematic in ways in which simple cluelessness is not problematic.

Equipped with this analysis, I considered Greaves' claim that complex cluelessness is a serious problem for attempts to do as much good as possible with limited resources. According to Greaves, the long-term consequences of the activities of candidate charities are unclear, there are reasons to believe that they are very good and reasons to believe that they are very bad, and it is very difficult to weigh those reasons. I argued that, by reasonable standards, this worry can be met: If it is phrased as a general neo-Malthusian worry about attempts to improve health and wellbeing in developing countries, then a standard battery of compelling responses to neo-Malthusianism apply. If it is phrased as a worry about adverse side-effects or downstream consequences caused *insensitively* with respect to background conditions, candidate adverse side-effects and downstream consequences can be subjected to empirical scrutiny one-by-one. And if it is phrased as a worry about adverse side-effects or downstream consequences caused *sensitively* with respect to background conditions, we are back to unproblematic simple cluelessness.

Finally, I considered the suggestion to focus on actions with the best candidate long-term consequences as a response to complex cluelessness, inspired by strong longtermism. Such a long-term orientation is defensible, I argued, to the extent that the precipice hypothesis and the lock-in hypothesis are defensible, but, contrary to what Greaves suggests, it is not promising specifically as a response to complex cluelessness. Attempts to systematically reduce existential risks will inevitably proceed via somewhat indirect causal paths. How the variables along these paths will interact with each other and with further relevant variables is inevitably difficult to predict. Attempts to seriously tinker with existential risks inherently come with their own risks. If the precipice hypothesis and/or the lock-in hypothesis are true, there are plausibly excellent options for doing good that involve attempts to mitigate existential risks and/or shaping value lock-in. However, to the extent that it is the specter of complex cluelessness that makes one doubt the wisdom of near term-oriented interventions (such as donating to AMF), turning to long term-oriented ones instead is not a promising recipe for avoiding complex cluelessness.

Acknowledgments

I would like to thank Emilie Aebischer, Maarten Boudry, Andreas Schmidt, Max Theissen, David Thorstad, and members of the Global Priorities Institute seminar for discussions and helpful feedback on (presentations of) earlier versions of this paper.

References

Christian, Brian (2020). *The Alignment Problem: Machine Learning and Human Values*. W. W. Norton & Company.

Cowen, Tyler (2006). The Epistemic Problem Does Not Refute Consequentialism. *Utilitas*, *18*(4), 383–399.

Dorsey, Dale (2012), Consequentialism, Metaphysical Realism and the Argument from Cluelessness. *The Philosophical Quarterly*, *62*(246), 48–70.

Ehrlich, Paul (1968). *The Population Bomb*. Ballantine Books.

Friederich, Simon and Sach Mukherjee (2021). Causation as a High-Level Affair. In Jan Voosholz and Markus Gabriel (Eds.), *Top-Down Causation and Emergence* (297–304). Springer.

GiveWell (2021). Against Malaria Foundation – November 2021 version. Retrieved from https://www.givewell.org/charities/amf/november-2021-version

Greaves, Hilary (2016). Cluelessness. *Proceedings of the Aristotelian Society*, *116*(3), 311–339.

Greaves, Hilary (2020). Evidence, Cluelessness and the Long Term. Talk presented at the *Effective Altruism Student Summit* in October 2020. Transcript retrieved from https://forum.effectivealtruism.org/posts/LdZcit8zX89rofZf3/evidence-cluelessness-and-the-long-term-hilary-greaves

Greaves, Hilary and William MacAskill (2021). The Case for Strong Longtermism. Global Priorities Institute working paper No. 5–2021. https://globalprioritiesinstitute.org/hilary-greaves-william-macaskill-the-case-for-strong-longtermism-2/

Jenkins, Jesse D., Max Luke, and Samuel Thernstrom (2018). Getting to Zero Carbon Emissions in the Electric Power Sector. *Joule*, *2*(12), 2498–2510.

Lenman, James (2000). Consequentialism and Cluelessness. *Philosophy and Public Affairs*, *29*(4): 342–370.

Lewis, David (1986). Postscripts to "Causation". In David Lewis (Ed.), *Philosophical Papers* (Vol 2, 184–188). Oxford University Press.

Mogensen, Andreas L. (2021a). Maximal Cluelessness. *Philosophical Quarterly*, *71*(1), 141–162.

Mogensen, Andreas L. (2021b). Moral Demands and the Far Future. *Philosophy and Phenomenological Research*, *103*(3), 567–585.

Mogensen, Andreas L. and William MacAskill (2021). The Paralysis Argument. *Philosophers' Imprint*, *21*(15), 1–17.

Moore, George E. (1903). *Principia Ethica*. Cambridge University Press.

Pearl, Judea (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press.

Reynolds, Jesse L. (2019). Solar Geoengineering to Reduce Climate Change: A Review of Governance Proposals. *Proceedings of the Royal Society A*, *475*(2229), 20190255.

Rosling, Hans (2018). *Factfulness: Ten Reasons We're Wrong About the World—and Why Things Are Better Than You Think*. Flatiron.

Russell, Stuart (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.

Smart, John J. C. (1973). An Outline of a System of Utilitarian Ethics. In John C. Smart and Bernard Williams (Eds.), *Utilitarianism: For and Against* (3–75). Cambridge University Press.

Tarsney, Christian (2023). The Epistemic Challenge to Longtermism. *Synthese*, *201*(195). https://doi.org/10.1007/s11229-023-04153-y

Thorstad, David (2023). High Risk, Low Reward: A Challenge to the Astronomical Value of Existential Risk Mitigation. *Philosophy and Public Affairs*, *51*(4), 373–412.

Vogt, William (1948). *Road to Survival*. William Sloan Associates.

Woodward, James (2006). Sensitive and Insensitive Causation. *The Philosophical Review*, *115*(1), 1–50.

Woodward, James (2021). *Causation with a Human Face*. Oxford University Press.