The Problem of Complex Cluelessness: what is it and what can we do about it?

11-2-2018

Siebe Rozendal

Course: Cluelessness and consequentialism (Tutorial)

Supervisor: dr. Andreas Schmidt

University of Groningen

Faculty of Philosophy

5,714 Words

Abstract In this paper, I consider whether we are fundamentally clueless about the value of the total consequences of our actions, and if so, what we can do to improve our epistemic situation. I briefly discuss what Greaves calls 'simple cluelessness'. Afterwards, I discuss 'complex cluelessness', which is the problem that in many decision situations we do not know how to compare the expected values of different acts if all possible consequences are considered. I conclude that 'complex cluelessness' is under-described, and the problem (or set of problems) needs to be disentangled before we can make much progress on it. I suggest some questions that seem important to address and offer different options of how to address them.

I. Introduction

Cluelessness is a problem for ethical theories that hold that whether one ought to do an act is, at least partly, a function of the consequences of that act. If we are clueless about how good the consequences are of two different acts under consideration, we are clueless about what we ought to do. Of course, no one can foresee the future, but many philosophers believe that it is possible to assign precise probabilities to all different possible values of outcomes of an act. One can then calculate the act's *expected value*: the sum of values weighted by their probabilities of occurring. However, if we cannot estimate the values of different outcomes or cannot estimate their probabilities of occurring or both, we are clueless about what we ought to do. Acts have many consequences, some of which very unpredictable, which make it difficult or impossible to estimate values of possible outcomes and their respective probabilities. To some, cluelessness is seen as an argument against consequentialism (Lenman, 2000). To others, it is not a problem (Burch-Brown, 2014; Mason, 2004), or a challenge to be overcome (Greaves, 2016). I belong to the last category.

If cluelessness is real there is much to worry about. Any agent who wants to promote some value, be it utility, equality, perfection, or a plurality of values, would have no clue as to how to promote those values if all consequences of an act are considered. Much cluelessness stems from empirical, rather than moral, uncertainty, but I believe this is not a reason to ignore the problem, because there are such strong implications if the problem is real. There is a distinction between problems of simple and complex cluelessness by Greaves (2016) that I want to follow in this paper. Most of the literature is concerned with what she calls 'simple cluelessness', of which I will give a short description. However, I will give some reasons that indicate that there is a plausible solution to simple cluelessness, and that the problem of 'complex cluelessness' is more pressing. The structure of this essay is different from most philosophical essays. In sections III and IV, I try to clarify the vague problem that is complex cluelessness and identify some possible means of approaching the problem. Nearly everything is therefore very tentative and should be taken as a call for more research rather than as strongly supported advice to deal with the problem of complex cluelessness.

II. The problem of 'simple cluelessness'

The problem of simple cluelessness is described most thoroughly in Lenman (2000), although it has been brought up before. Lenman argues that most actions are, directly or indirectly, identity-affecting. Any identity-affecting action has as part of its consequences all the

consequences that came about by those persons existing, instead of other persons existing.[1] In such a way, the merciful bandit Richard who chooses not to kill a pregnant villager Angie has as a consequence that her descendant 100 generations later is Adolf Hitler, and therefore Richard's act of mercy results in a genocide. Yet before we condemn Richard, we cannot even claim that this genocide was morally worse than if Richard had killed Angie, for that act might have brought about the existence of Malcolm the Truly Appalling, who brings about a much worse state of the world. These consequences are completely unforeseeable given the unpredictable nature of complex systems, and agents have only limited time and computational power to deliberate. For many acts, Lenman concludes, the value of their foreseeable consequences is greatly outweighed by the value of their unforeseeable consequences, and there are many consequences that cannot be observed even after the fact (*ex post*) because causal networks are so complex. Therefore, the argument goes, the foreseeable consequences only give us a very weak reason to maximize goodness of consequences, and we might better pursue strategies that are inherently valuable, such as being happy, being a good friend and a good community member.

A number of reasons have been given for dismissing the problem of simple cluelessness. I shall discuss four. First, J.C. Smart (1961) has suggested that consequences fade out over time like ripples in a pond. This is largely an empirical question, but it seems that for many consequences, especially identity-affecting ones, the amount or size of the consequences increases rather than decreases over time (Greaves, 2016; Lenman, 2000). Second, one might suppose that the sum of values of all unforeseeable consequences tends to be zero. But Greaves (2016) argues that we have no reason to suppose that they do, and that we have some reason to believe that they do not tend to be zero, based on the theory of random walks.[2] The third response grants some strength to the simple cluelessness problem, but limits its scope a little bit. Cowen (2006) argues that some acts have foreseeable consequences that are big enough to outweigh the unforeseeable ones. For example, a nuclear bomb being detonated in Manhattan, the extermination of the USA, or extinction of 99.99% of all sentient life in the universe. He calls the view that we should focus on positively affecting these big events Big Event Consequentialism. I agree with Cowen that some events have big enough foreseeable consequences that we have sufficient reason to believe that they outweigh unforeseeable ones, although it is unclear when an event

[1] Note that even if actions are not identity-affecting, the argument may still work if consequences keep affecting other consequences further in time. However, the case for identity-affecting acts is strongest, so I will focus on those.

[2] For a mathematical explanation, see Greaves (2016, p. 314).

becomes ‘big enough’. Furthermore, even these big events might have good consequences. For example, near extinction events might reduce the risk of future extinction by giving humanity a hard lesson of the risks we face.

The last and most promising line of argument I shall discuss is concerned with the Principle of Indifference (POI). It means that when we have no more evidence for P than for Q, our belief that P will happen should be equal to our belief that Q will happen, and because in the situation of simple cluelessness we have absolutely no evidence for either P following from act A being more likely, or P following from not-A (where P is some merely plausible story of what might happen). Therefore, our belief that P will happen given A should be equal to our belief that P will happen given not-A. If the POI is correct, then the expected value of an act is not affected by unforeseeable consequences.

However, Lenman (2000) claims that we cannot invoke the POI, as it would require an arbitrary partitioning of the set of possibilities, before we assign equal credences to each part. There are many ways to partition a set of possibilities, and there seem to be only arbitrary ways to partition without access to the probabilities of different elements. Many other authors have argued against invoking the POI as well (Greaves, 2016). But Greaves (2016) and Mason (2004) both claim that in this specific case, the POI is applicable, as there is a natural symmetry between something happening and something not happening that makes the partition into P given A and P given not-A nonarbitrary. This natural symmetry consists in that any plausible story that an act A results in large unforeseeable consequences P has a precise counterpart, namely that refraining from the act (i.e. not-A) results in the large unforeseeable consequences P. For example, Richard’s murdering of Angie might lead to a series of consequences that finally result in Hitler’s genocide, but his not-murdering might - via the same causal chain - finally result in Hitler’s genocide. Thus, the equal credences in P following from A, as P following from not-A, result in an unchanged expected value of A.

There is more to say about this and I admit that the debate around simple cluelessness has not been settled. Nevertheless, I find Greaves’ response most plausible and I want to focus on complex cluelessness instead for the rest of this paper. This strikes me as an important problem which is more likely to be taken seriously outside of academia as well, as it strikes people as more than a philosophical oddity. Furthermore, it has received very little attention so far in academic philosophy.

III. The problem of complex cluelessness

Complex cluelessness is different from simple cluelessness. While simple cluelessness is concerned with the consequences of either performing or omitting an act, complex cluelessness is concerned with how consequences of different acts weigh up against each other. Furthermore, whereas there is evidential symmetry in simple cluelessness, the evidence for the different consequences in complex cluelessness is not symmetrical. The reasons that hold that one act will lead to good consequences overall are different from the reasons that another act will lead to good consequences overall. It is unclear how to weigh these reasons against each other. Greaves (2016, p. 9) formulates the following propositions to describe complex cluelessness:

> For some pair of actions of interest A_1, A_2,
>
> (CC_1) 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;
> (CC_2) 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;
> (CC_3) It is unclear how to weigh up these reasons against one another.

As an example, we may suppose that we have $1 million to spend. If we spend it on insecticide-treated bed nets distributed in a neglected area where malaria is problematic, we may reasonably expect that we avert roughly 270 – 310 lives of children under the age of five, or something of that order of magnitude (GiveWell, 2017)[3]. However, these couple of hundred children will grow up and affect among others population growth, economic growth, climate change, and meat consumption. These "indirect" effects are of an uncertain value and uncertain size, and some of the effects themselves are not clearly positive or negative. For example, on a total utilitarian view, population growth may be favored, as long as lives are worth living and it does not result in human extinction. Economic growth may affect the speed of technological development, which might be too fast too control, resulting in global catastrophe or extinction of humans. Alternatively, we may spend our $1 million on bio-technological research on in-vitro meat, which might be able to radically change the meat industry, substituting meat from slaughtered and badly treated animals with meat grown from a couple of stem cells. This will affect animal wellbeing, farmed animal populations, human health, and climate change. Again,

[3] Obviously, even these numbers have a lot of uncertainty, but these numbers have some empirical grounding (Givewell, 2017) and these effects are on an observable time scale.

it is unclear how our donation will affect these factors. Even more difficult is it to compare the goodness of one option's consequences to the goodness of the other option's consequences.

III.2 Disentangling complex cluelessness

So far, Greaves' conception of complex cluelessness is the only one I am aware of (although I do not exclude similar concepts being discussed in different terms elsewhere), and her description is rather short. Furthermore, she writes that she does not know the answer to the question "what is the right theoretical description of complex cluelessness?" (Greaves, 2016, p. 321). Therefore, there is much room to clarify the problem, which I attempt to do now. I take a rather formal and abstract approach to this problem, and this is possibly the wrong choice. It is plausible that cases of complex cluelessness should be examined individually, and one should look for practical solutions sooner rather than later. It is even possible that this latter approach would allow us to, over time, generalize our findings into a framework to deal with cases of complex cluelessness. Nevertheless, a general framework would have much merit and the abstract way, if successful, seems to be the shortest path towards such a framework.

Below, I will first define a number of sets. I will use these sets to distinguish different uncertainties we face as decision makers. I make some simplifications which I discuss in section V.2. In section IV I will discuss four out of the six uncertainties in more detail. My method is to investigate multiple uncertainties shallowly, rather than defining a narrow problem and discussing it in-depth. Although this is uncommon in analytical philosophy, I believe it is the best approach considering the lack of attention that this problem has received.

Assume determinism and act-consequentialism. There is a finite set of possible acts for an agent at a specific point in time. Acts are not mutually exclusive.

$$A\ \{A_1, A_2, \ldots, A_n\}.$$

At any point in time, an agent can select acts from *A* to form a strategy: the set *S* contains all possible strategies, which are mutually exclusive. The size of the members of *S* depends on the properties of the choice-situation: if the agent has a lot of time and resources, strategies include more acts.

$$S\ \{S_1, S_2, \ldots, S_n\}$$

Every strategy in *S* has a corresponding member C_i in the set of possible consequences

$$C\ \{C_1, C_2, \ldots, C_n\},$$

such that C_1 is a set that includes all the actual consequences that arise from S_1, C_2 all actual consequences from S_2, and so forth. C has a subset of foreseeable consequences

$$C_{foreseeable}\ \{C_{1\_foreseeable}, C_{2\_foreseeable}, \ldots, C_{n\_foreseeable}\},$$

of which every $C_{i\_foreseeable}$ is a subset of C_i.

There is another set which contains all foreseen consequences, foreseen by an agent at a specific point in time:

$$C_{foreseen}\ \{C_{1\_foreseen},\ C_{2\_foreseen}, \ldots, C_{n\_foreseen}\}$$

This is not a subset of *C* or $C_{foreseeable}$, because an agent might mistakenly foresee consequences that will not happen. Furthermore, the sets $C_{i\_foreseen}$ contain all consequences about which the agent has *any* credence that they will follow from an act in *A*, and thus these sets are able to contain mutually exclusive consequences.

The set $C_{unforeseen}$ contains all the members of *C* that are not in $C_{foreseen}$.

$$C_{unforeseen} = C \setminus C_{foreseen}$$

However, knowing the relations between different sets of consequences is not enough if we want to bring about good consequences. We also need to know the value of the consequences. If we can aggregate the values of each C_i into a single metric (e.g. goodness, choiceworthiness, or utility), there will be a total order possible of all elements in *C* and $C_{foreseeable}$ according to their values. If we can do the same for each $C_{i\_foreseen}$, and weight the different elements in each set by the agent's credence in them, we have an expected value ordering. As a result, we would have three totally ordered sets: *V (C)*, *V* ($C_{foreseeable}$), and *EV* ($C_{foreseen}$). Ideally, the order of $EV(C_{foreseen})$ completely matches the order of *V(C)*, such that every $C_{i\_foreseen}$ has the same ranking as C_i. This would mean that whenever we believe that S_i has a higher expected value than S_j, S_i *will* have better consequences. But this is impossible if our expected value is realistic. Take the case from Jackson (1991) of three possible drugs to cure Jill, who has a minor rash: drug A will cure her partially, either drug B or drug C will completely cure her and the

other will kill her, but we do not know which one is which. The expected value order would be $A \succ B \sim C$, but A will never bring about the best consequences.

Thus, we need to substitute *V(C)* for another ordered set that we want to approach as much as possible. Let us define $EV(C)$ as the ordered set that would be ranked in terms of expected value[4] by an ideal observer who was completely rational and omniscient about all moral and nonmoral facts. This circumvents cases such as Jackson's. We want our expected value ordering $EV(C_{foreseen})$ to consistently and closely match $EV(C)$ every time we pick a strategy. If it would perfectly match, there would be a correlation of 1.

There are some problems with this approach. First off, I am not sure which relationship between the two ordered sets we want to optimize. Because the strategies are all mutually exclusive, we can only choose one strategy, and we will choose whatever seems best to us. Therefore, what matters is that our top strategy is good, and not so much the correlation, or "how close one ordered set is to the other"[5]. Second, the sets are different size, because we have no expected value for unconsidered strategies. We could treat unconsidered strategies as having an expected value of 0. Alternatively, they could be treated as indeterminate. Third, cardinal information about the expected values is ignored in ordered sets. How important this is depends on what the cardinal information is: if many strategies have net negative consequences and only a few have net positive consequences, that implies a different selection strategy than if there are mostly strategies with positive consequences, and only a few with very bad ones.

Now that I have discussed some elements of the problem of complex cluelessness, let us look which uncertainties result in cluelessness. Greaves writes that complex cluelessness is that, for any pair of acts, we do not know which act of the two produces better consequences all things considered. I think we can distinguish the following uncertainties, that together create this state of cluelessness.

1. We do not know all the strategies that are available to us, and we might miss the strategies with the most valuable consequences.
2. We do not know which consequences follow from which strategies.

[4] This allows the possibility of objective probabilities, such that even an ideal observer would not know which total consequences will obtain for sure. However, this account does not depend on objective probabilities existing.

[5] For an overview of different methods to measure and improve closeness (distance), see (Truchon, 2007).

3. We do not know how the consequences of different strategies compare with one another in terms of value.
4. Because of 2 and 3, we have no idea how to order our considered strategies on the basis of their expected value. Not even if we only consider two very limited strategies.[6]

If it is possible to compare $EV(C_{foreseen})$ with $EV(C)$ in some way, then there are also the following uncertainties:

5. We do not know how close our current expected value orderings of foreseen consequences is to the expected value ordering of an ideally rational and completely informed observer.
6. We do not know how to make our ordering closer to that of an ideally rational and completely informed observer.

As I see it, there are roughly two approaches: either 1, 2, 3, and 4 are resolved, which automatically resolves 5 and 6, or 5 and 6 are resolved directly, and it is not necessary to resolve uncertainties 1 to 4. I do not know whether it is possible to resolve 5 and 6, and I do not know how I would approach it. I am not even sure that it makes sense to talk about comparing $EV(C_{foreseen})$ and $EV(C)$.[7] In the rest of this paper, I will focus exclusively on uncertainties 1 to 4.

IV. Resolving uncertainties 1 to 4.

In this section, I shall explore some ways to reduce the first four uncertainties from the previous section in a practical way. Complementary to these practical approaches are more theoretical ones such as this paper, which are needed to improve our understanding of what the uncertainties are. These are not to be taken as policy recommendations, but rather as a start for discussion and further idea development.

IV.1 Are we missing the most valuable strategies?

The first uncertainty is that we do not know whether we have considered most high-value strategies. Our cluelessness about what will bring about the best consequences may be a direct result from a too narrow consideration of strategies. We might be unable to decide between S_1 and S_2, but if would only consider S_x, we would know that S_x would be best all things

[6] This is the cluelessness that Greaves talks about.
[7] I discuss a further problem in the conclusion.

considered. On the other hand, broad consideration might also introduce cluelessness. We might know which intervention is best at reducing child mortality from malaria, but if we try to compare that to improving farm animal welfare or the reduction of the risk of extinction, we would not know anymore what is best. Nevertheless, we want to do what is best all things considered, and not just what is best at reducing child mortality from malaria.

What could S_x look like? And how could we find it? S_x could be a recombination of existing strategies or something completely new. In the effective altruism community, a community of people who aim to do the most good, a common approach is to first select a "cause area" (e.g. global poverty, animal welfare, long-term future), and then compare strategies only within that area. If their underlying assumption is correct – that there are large differences in value between cause areas – then it pays off to focus on a single cause area. A set of cause areas that jointly exhausts all possible strategies ensure that, at least at the first step which is cause area selection, no strategies are missed out on. The cause areas of effective altruism are not jointly exhaustive: many presently existing humans are not considered, and therefore strategies such as battling smoking or depression may be overlooked. Furthermore, cause areas are large and need to be divided further into smaller sets. If we want to consider all possible strategies, every step of narrowing down should result in a jointly exhaustive set of strategies of the superset. It need not be a partition, as overlap of subsets would not be too harmful. Yet, sets can be carved up into many different ways and I cannot provide an a priori way to do this.

IV.2 Which consequences follow from which strategies?

This question is very broad and there are already many incentives to predict consequences accurately. Therefore, we should focus on consequences with large values which will dominate our expected value ordering.[8] We should discover the ones with the largest value and estimate their probabilities (i.e. estimating the probability of them happening, given a certain strategy is executed, for every strategy) more accurately. What kinds of strategies have large value consequences? An obvious candidate is strategies large in scale. However, if one "big strategy" can be pursued, this means that others could have been pursued also. So, we should look for strategies that have large value consequences, *relative to* the same decision situation (i.e. same agent, same endowments of time, skills, resources, etc.).

[8] This is especially the case if consequences are power-law distributed, where only a few consequences account for most of the expected value of strategies. However, we do not (yet) know if this is the case.

At least three kinds of strategies have potentially large value consequences relative to other strategies. First, some strategies have more long-lasting effects than others. For example, helping humans will probably have more consequences than helping nonhuman animals. However, the value of long-running consequences depends on the average expected value of consequences. If the consequences have an average expected value of zero, such as in the case of identity-affecting actions in 'simple cluelessness', it does not matter that the consequences are large in size or number. For these consequences to matter, they must either have an average negative expected value (so we can avoid them) or an average positive expected value (so we can promote them). This brings us to the second kind of strategy: some strategies are affect more aspects in the moral domain than others. For example, strategies with consequences that touch on rights or well-being of morally relevant individuals. Which strategies fit the bill depends on the scope of the moral domain and on answers to questions such as "which entities require moral consideration?" A third kind of strategy is when some possible futures are excluded. Choices that we make now affect the range of choice we will have in the future. Accidentally limiting this range should be avoided. One option that seems to keep our options open is reducing the risk of extinction. However, a narrower set of choices could be beneficial if we have eliminated negative possible futures. But until we become less clueless, it is too difficult to determine whether we are closing off positive or negative possible futures.

IV.3 How valuable are certain consequences?

Consequences are valuable in two ways. First, they matter by how valuable the states of the world are that are brought about. Thus, if I save a child's life, the consequences of a life saved is valuable intrinsically. Second, consequences matter extrinsically by virtue of other states of the world they lead to. The child whose life is saved will contribute to or decrease value from the world, and this is the extrinsic value of the consequence of a life saved. To determine the expected value of a strategy, we need to know which consequences are brought about, and how valuable these consequences are intrinsically and extrinsically. Determining the intrinsic value of consequences is difficult because we have *axiological uncertainty*: we are not sure what is valuable and how much (or at least there is no consensus about this). Determining the extrinsic value of consequences is complicated further by the fact the we do not know which consequences follow from which, which is discussed in the previous section.

Resolving axiological uncertainty is difficult, as any ethics scholar can attest to. However, a fruitful approach might be *metanormativism*: the view that there are second-order norms that govern action that are relative to a decision-maker's uncertainty about first-order normative claims" (MacAskill, 2014, p. 2). Although the field is still in its early stages, one popular view holds that different ethical (and thus, axiological) theories can be aggregated into a single expected choice-worthiness function which should be maximized. For example, if we have an 80% credence in an axiological view that the intrinsic value of a set of consequences is 10, and 20% in a view that holds its value is minus 100, its expected choiceworthiness is minus 12. For an overview of the issues in dealing with moral uncertainty, see Bykvist (2017).

Another thing to look out for is *crucial considerations*. "A crucial consideration [..] would be a consideration that radically changes the expected value of achieving [a high-level] subgoal" (Bostrom, 2014). What this means is a small amount of considerations can completely overturn our expected value assessments. For example, realizing that insects are sentient in conjunction with believing that sentience is a sufficient condition for moral status would radically alter the valuation of our possible strategies. Therefore, the potential existence of crucial considerations that are relevant for a particular strategy should make us very uncertain about the expected value of that strategy. This would be reflected in the resilience level of the expect value, rather than in the size of the expected value itself.[9] A low resilience level entails that the number (specifically the probability component) is very subject to change when particular new information comes in.

IV.4 Comparing the expected value of strategies

One problem for constructing an expected value calculation was that we are so uncertain that we cannot assign precise credences to values of different possible outcomes. However, contrary to orthodox Bayesianism, different authors have argued for the use of imprecise credences when faced with deep uncertainty (Romeijn & Roy, 2014). Imprecise credences mean that belief states are represented by a set of probability functions, rather than by a single probability function. Greaves (2016) explores what the implications of imprecise credences might be. She concludes that for acts we are clueless about they are most likely indeterminate. That is, they are neither permitted, required, nor forbidden. However, I think that if we grant that the moral

[9] For an explanation of how probabilities reflect evidence, see (Joyce, 2005)

status of an act or strategy is indeterminate that does not solve the problem of complex cluelessness, but it might capture it.[10]

V. Conclusion

I have discussed two forms of cluelessness: simple and complex. It seems that simple cluelessness has a plausible response, namely that evidential symmetry about different stories warrants the Principle of Indifference (POI). However, complex cluelessness is a tougher nut to crack. It seems that in at least some cases, we do not know which strategy is best to do. I have considered some uncertainties that create this form of cluelessness, and I have explored some ways of dealing with those uncertainties. I believe it is highly unlikely that all my points are sound or useful, and I encourage further exploration of this important topic. I will now discuss some practical implications and ways in which my account of complex cluelessness can be improved.

V.1 Implications for complex cluelessness cases

I believe that there are some recommendations that can be made for cases of complex cluelessness, although they are very tentative. First off, try to not bring about consequences that are irreversible. Although this may be hard to do, avoiding human extinction or Orwellian scenarios that are hard to unilaterally escape from are good strategies to strive for. Second, we should develop methods to not overlook valuable strategies. This requires both rigor and creativity, two qualities that maybe have not been simultaneously employed enough, because they are so opposite from each other.

If we are clueless in *every* decision situation, it appears the only thing with any expected value is more research. It would be worthwhile to find out, amongst others, how the expected value of possible total consequences ($EV(C)$) is distributed. Is it a Gaussian or a Pareto distribution, and with which properties (e.g. mean, variance, and for Pareto: which alpha)? Answering this question would affect our decision procedure strongly. Another worthwhile research direction is reducing our moral uncertainty (specifically axiological) and improving our methods to deal with moral uncertainty. However, I cannot rule out the possibility that we are clueless about whether this research brings about good consequences. Maybe we really are clueless even about whether trying to resolve complex cluelessness brings about good consequences.

[10] Greaves does not present this a solution to the problem of complex cluelessness and admits that the implications of imprecise credences for moral theory should be further explored.

The potential existence of omnipresent cluelessness should not keep one from acting morally. Maybe we are not actually clueless and should just follow common sense. But if we are clueless, it is better to attempt to act morally, so that we can evaluate whether our strategies have worked once we have the tools necessary for good evaluation. Furthermore, the strategy that seems most likely to bring about good consequences is promoting moral behavior, altruism, and value as such (Williams, 2013). This seems especially valuable when behavior is promoted that tracks intellectual and moral progress such that new insights will be incorporated.

V.2 How my account of complex cluelessness can be improved

To round off, I want to discuss ways in which my treatment of complex cluelessness falls short, can be improved, or is flat out wrong. First off, I think that it is worthwhile both to think in terms of acts that are available to an agent, as well as thinking in terms of strategies that are available to an agent. Thinking in terms of acts is valuable because one can consider combining acts that complement each other. For example, both investing in technological development of a potentially dangerous technology as well as investing in safeguards for that technology. Thinking in terms of strategies has the benefit that one can assess the consequences from performing all the acts. The consequences of performing only act A_1 and the consequences of performing only act A_2 cannot simply be aggregated. The acts might interact when performed together, producing different consequences.

I have made a number of simplifications. With regards to acts, I have not addressed that it is arbitrary to individuate an act in a certain way: it is always possible to define an act in a more atomic way. With regards to strategies, I have not addressed that strategies should be *ordered* sets, because one does not perform all acts simultaneously and the order matters both theoretically and practically. For example, first researching whether a vaccination works and then setting up a vaccination program has different consequences if performed the other way around. Furthermore, strategies are not executed in a vacuum, but the choice of strategy affects the payoffs, possible strategies, and expected success rates of strategies for different agents, a fact long established by game theorists. Incorporating this fact makes our decision situations more complex. Another aspect of strategies is that "figuring out which strategies are available" and "figuring out how valuable some strategies are" are in themselves acts that belong to certain strategies, which makes things more complex. I have also not touched upon the nature of agents and decision situations and how that matters for cluelessness. Agents with more time, resources,

or groups of agents obviously have more strategies available to them, which might affect their uncertainty either positively or negatively.

The last weakness I want to discuss concerns the expected value ordering of foreseen consequences, $EV(C_{foreseen})$. To consequentialists, it seems strange to have side constraints to expected value. For example, if we think that "strategy S_1 has the highest expected value, considering the consequences that we foresee. However, in the past it was always our second-highest ranked strategy that produced the best outcomes so we should choose that", the extra reason that was offered after constructing an initial expected value ordering should also be included in the expected value! In other words, it would be irrational to not execute the strategy that has the highest expect value. This leaves open the question whether it makes sense to ask how $EV(C_{foreseen})$ relates to $EV(C)$, which is the ordered set that would be ranked in terms of expected value by an ideal observer who was completely rational and omniscient about all moral and nonmoral facts. All I can say is that we are far away from resolving the problem of complex cluelessness.

References

Bostrom, N. (2014). Crucial Considerations and Wise Philanthropy. Retrieved from https://www.effectivealtruism.org/articles/crucial-considerations-and-wise-philanthropy-nick-bostrom/

Burch-Brown, J. M. (2014). Clues for Consequentialists. *Utilitas*, *26*(1), 105–119. https://doi.org/10.1017/S0953820813000289

Bykvist, K. (2017). Moral uncertainty. *Philosophy Compass*, *12*(3), 1–8. https://doi.org/10.1111/phc3.12408

Cowen, T. (2006). The Epistemic Problem Does Not Refute Consequentialism. *Utilitas*, *18*(4), 383–399. https://doi.org/10.1017/S0953820806002172

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

Jackson, F. (1991). Decision-Theoretic Consequentialism and the Nearest and Dearest Objection, *101*(3), 461–482.

Joyce, J. M. (2005). How probabilities reflect evidence. *Philosophical Perspectives*, (19).

Lenman, J. (2000). Consequentialism and Cluelessness. *Philosophy & Public Affairs*, *29*(4),

342–370.

MacAskill, W. (2014). *Normative Uncertainty*. University of Oxford.

Mason, E. (2004). Consequentialism and the Principle of Indifference. *Utilitas*, *16*(3), 316–321. https://doi.org/10.1017/S0953820804001190

Romeijn, J.-W., & Roy, O. (2014). Radical Uncertainty: Beyond Probabilistic Models of Belief. *Erkenntnis*, *79*(6), 1221–1223. https://doi.org/10.1007/s10670-014-9687-9

Smart, J. J. C. (1961). An outline of a system of utilitarian ethics.

Truchon, M. (2007). *Aggregation of Rankings : a Brief Review of Distance-based Rules and Loss Functions for the Expected Loss Approach*.

Williams, E. G. (2013). Promoting Value As Such. *Philosophy and Phenomenological Research*, *87*(2), 392–416. https://doi.org/10.1111/j.1933-1592.2012.00601.x