Simplifying Cluelessness

Philip Trammell

June 5, 2019

Abstract

Given the radical uncertainty associated with the long-run consequences of our actions, consequentialists are sometimes "clueless". Informally, this is the position of having *no idea whatsoever* what to do. In particular, it is *not* the position of facing actions that merely take on wide distributions of possible value. Existing efforts to formalize cluelessness generally frame the phenomenon as a consequence of having imprecise credences. Even if some such framing is ultimately correct, however, it appears, at the moment, not to be particularly effective at communicating the seriousness of the problem clueless agents face. It is not obvious what it means to "have some (or no) idea what to do" when one's credences are imprecise, as the variety of theories of rational choice under imprecise credences testifies. Furthermore, anecdotally, it appears that some people find it difficult to grasp the motivation behind existing theories of imprecise credence, or are not satisfied that imprecise credences could give rise to importantly different decision-theoretic situations from those a rational Bayesian consequentialist faces when his actions merely take on wide distributions of possible value. In this article, therefore, I present a brief sketch of a formal treatment of cluelessness that does not depend on a theory of imprecise credences. In doing so, I do not hope to provide an accurate account of the phenomenon in full detail, but only to convince the reader that there is a real and important fact of consequentialist life which the tools of orthodox epistemology and decision theory cannot handle.

1 Introduction

Most people are roughly indifferent about what happens in a long time. A good consequentialist is not. If we are consequentialists—or even non-consequentialists who care to some extent about consequences—we believe that our choices should be governed in part on the basis of all our actions' consequences, including those billions of years out. Unfortunately, we have no idea what these long-run consequences are. How then do we decide what to do?

Smart (1973) asserts without argument that we can decide what to do on the basis of potential actions' short-run consequences, because the most significant consequences of our actions accrue in the short run. Impacts, he says, wash out over time "like ripples on a pond".

Lenman (2000) points out that this is false. In particular, he points out that every decision that affects the timing of some human conception results in the creation of a human being

with a totally different genome—and that this can ultimately change the numbers, personality-distributions, and actions of all future generations. We are thus always radically uncertain, or "clueless", about our actions' long-run consequences.

Cowen (2006) and Dorsey (2012) argue that, even if our actions have substantial long-run ramifications (e.g. changing the "identities" of all future people past some date), the positive and negative impacts on *value* associated with an action will tend to cancel out. Short-term consequences do not serve well as proxies for total consequences, in other words, but short-term value serves well as a proxy for total value.

Burch-Brown (2014) agrees that we can choose on the basis of relatively direct consequences, but only because uncertainty about long-run/indirect consequences is, *ex ante*, symmetric across acts. When we are deciding whether to conceive a child on a Tuesday or a Wednesday, perhaps, or when we are deciding whether to help an old lady across a street, any chance that one act might have some long-run positive or negative consequence will be counterbalanced by an equal chance that the other will have that consequence. So we can't take comfort in the image of ripples on a pond, but we can still use short-run, or direct, impact as a proxy for *expected* value.

Greaves (2016) argues that this does not fully capture the phenomenon of cluelessness. Our uncertainty in long-run consequences is *sometimes* symmetric across acts; perhaps it is in the cases described above. In these circumstances, which Greaves terms cases of "simple cluelessness", we can indeed use short-run impact as a proxy for expected value. In the more interesting (and perhaps more common) cases, though, there is no such symmetry. When we're deciding whether to give enough to Malaria Consortium as to save roughly one person from dying of malaria—say, $2,000 (GiveWell, 2019)—we must weigh a huge basket of *semi-foreseeable* positive and negative consequences on the present human population, the future human population, farm animals, wild animals, economic growth, global warming, and very much more. In these cases of "complex cluelessness", people are sometimes intuitively inclined to pick out one foreseeable positive consequence (such as the saved child's life) and proceed as if the expected value of the cloud of other consequences were zero. But this is not a legitimate move; if it were, we could just as easily pick out one foreseeable negative consequence (the exacerbation of global warming) and proceed as if the expected value of the cloud of other consequences were zero as well. We would then have to call the expected value of giving to Malaria Consortium positive and negative at the same time. What to do in cases of complex cluelessness is, she says, an open question.

I agree. I think complex cluelessness is a substantive and uncrossed roadblock on the way to a general theory of consequentialist decision-making, and I think it warrants more careful attention than it gets. I also think it's the only kind of cluelessness worth the name, so I will just call it *cluelessness* for the rest of this article, ignoring the simple variety.[1]

[1]Besides Greaves, many in the Effective Altruism community have now begun at least some attempt in the direction of formalizing and studying the phenomenon of cluelessness (or decision-making under cluelessness), including Amanda Askell, Jesse Clifton, Aidan Goth, Milan Griffes, Andreas Mogensen, Christian Tarsney, Brian Tomasik, and Tatjana Visak. There are also of course many professional philosophers and other scholars who have

2 "Just be rational"

The most obvious way in which cluelessness differs from the kind of uncertainty that allows for expected values is phenomenological: the two things feel different. My uncertainty about the number of coin flips out of seven that will come up heads can be characterized by a precise probability distribution whose mean is 3.5. My uncertainty about the value of a die roll can be characterized by a different, but also precise, probability distribution whose mean is 3.5. If saving a life causes Nature to flip a coin seven times and save as many additional lives as came up heads, and failing to save the life causes Nature to do the same, my uncertainty about the acts' indirect effects is "symmetric", and I should save the life. If instead saving the life causes Nature to roll a die and save as many additional lives as the number that lands on top, my uncertainty about the acts' indirect effects is no longer symmetric, but saving the life still has a higher expected value than not doing so. But if saving the life causes Nature to roll a die 40 times and save (or, if the number is negative, end) as many lives as the product of the rolls minus the number of grains of sand on earth, I just don't have an expected value—at least, not immediately—for one of the acts available to me. I don't have any feeling of expectation. With respect to giving $2,000 to Malaria Consortium, this is the predicament my brain is currently in.

Objection #1 – Granted, the Malaria Consortium case feels different somehow from a coin-flip / die-roll case. But we shouldn't leap from there to saying that your brain actually has *no* expectation of the value of giving to Malaria Consortium. Regardless of how uncertain we are, shouldn't uncertainty always take the form of a probability distribution over states of the world? And pathological cases aside, mustn't that distribution have a mean?

Response – Perhaps there is some sense in which my credences *should* be sharp (see e.g. Elga (2010)), but the inescapable fact is that they are not. There are obviously some objects that do not have expected values for the act of giving to Malaria Consortium. The mug on my desk right now is one of them. Upon immediately encountering the above problem, my brain is like the mug: just another object that does not have an expected value for the act of giving to Malaria Consortium. Nor is there any reason to think that an expected value must "really be there", deep down, lurking in my subconscious. Lots of theorists, going back at least to Knight's (1921) famous distinction between "risk" and "uncertainty", have recognized this.

worked on, or are working on, closely related issues in decision theory and epistemology. If I were to write this article properly, without running the risk of saying something obvious (or obviously false), I would of course first have to read and more thoroughly incorporate all the related literature. For a long time, that prevented me from writing this. But it may be a long time before I finish all the relevant reading, and I keep having conversations with people who doubt that there could possibly be anything interesting at play here beyond the familiar fact of uncertainty; so I've decided to go ahead and write my thoughts down here as they are today, so I at least have something to send those who are interested. I'll edit this in the future when I read material that seems relevant enough to incorporate.

Objection #2 – Okay, maybe we sometimes don't experience expected values we can report right away. But surely, in every case, we can report expected values after thinking through the problem in question long enough, right?

Response – I'm not sure. But even if that's true (and I think it probably is, and I'm happy to assume it is), this deliberation could take an arbitrarily long time. So we still face the important question of what to do when we have to make a decision among acts now, or soon, and we're still at the stage where we don't have expected values for some or all of them.

Objection #3 – Okay, maybe we sometimes have to make decisions in the absence of expected values. Can't we still make instantaneous, better-than-chance best guesses?

Response – Not necessarily. More below.

3 Bounded rationality: a tiny overview

The most general possible way to model decision-making is to describe agents as acting according to arbitrary "probabilistic choice functions". Given a set of acts A and possible "contexts" C, agent i's probabilistic choice function $p_i(a, c)$ takes an act $a \in A$ and a context $c \in C$—where the context-description is rich enough to include everything that could possibly influence the agent's decision, including the other acts available—and returns the probability that i chooses a. The literature on probabilistic choice modeling is extensive; see McFadden (1973) for one classic early treatement.

On roughly the other end of the spectrum, a highly restrictive way to model decision-making is to assume that agents are fully "rational", in the sense that they "maximize their expected utility". In the language of Savage's (1954) model and its subsequent extensions to account for information acquisition, this framework posits a Borel space of possible states (S, Σ) (with measurable state-sets termed "events"), a set of possible outcomes X, and a set of possible acts $A \subset X^{\Sigma}$ mapping events to outcomes. It then assumes that for any agent i, there is a probability function over states $\mu_i : \Sigma \to \mathbb{R}$ satisfying the Kolmogorov axioms and a utility function over outcomes $u_i : X \to \mathbb{R}$ unique up to positive affine transformation such that, given any information set $\sigma_i \in \Sigma$ and feasible act-set $A_i \subset A$, we have $p_i(a, (\sigma_i, A_i)) > 0 \to \mathbb{E}[u_i(a(s))|\sigma_i] \geq \mathbb{E}[u_i(a'(s))|\sigma_i]\ \forall a, a' \in A_i$. In other words, an expected utility maximizer acts as if she has a precise probability function and a precise utility function, and she performs an act with positive probability only if the act is one of the expected-utility-maximizing acts available.

There is a wide continuum of possible formal models of behavior more general than the assumption of expected utility maximization but more restrictive than the assumption of arbitrary probabilistic choice functions. Among the earliest and best-known, for example, is Simon's (1957) exploration of "bounded rationality". A more recent, and more predictively successful, is

the "prospect theory" developed by Kahneman and Tversky (1979, 1992, etc.) and subsequent behavioral economists. A final model, sometimes used to capture cluelessness (see e.g. Mogensen, 2019), is a theory of imprecise credences, coupled with a decision rule under imprecise credences such as the "Sen-Walley Maximality Rule" (Walley, 1991). On this account, we act as if we have not a single probability function μ but a set of probability functions M, termed a "representor", such that we choose an act with positive probability only if there is no other act with higher expected utility according to all $\mu \in$ M.

All the models listed above—like essentially all formal models of decision-making—can take on a *representational* or an *action-guiding* interpetation. A model can always propose that agents' behavior be representable a certain way, but a model cannot be action-guiding for agents who lack immediate internal access to the quantities by which they recommend that behavior be representable. For instance, because we rarely have precise probability distributions over the states of the world relevant to the acts before us, or even precise utility functions over outcomes, the model of behavior as expected utility maximization is rarely action-guiding. (As Gilboa (2010, ch. 3) humorously demonstrates, the question "Whom should I date?", even in the absence of uncertainty about the state of the world, is rarely usefully answered with "Whoever maximizes utility".) Friedman (1953) famously points out that we might still expect to find people's behavior roughly representable as expected utility maximization, like we can represent trees as positioning their leaves to maximize expected sun exposure. But that is no help to us in the darkness as we try to make our choice; it is just a prediction, correct or incorrect, that we will not make too big a mistake. Likewise, unless we have precise utility functions, probability functions, *and* risk functions, we cannot be guided by prospect theory, or by any other rank-dependent utility theory, even if we would like to be, and even if our behavior might still be representable in accordance with said theory.

Likewise—though this appears to be less widely appreciated—unless we have immediate access to the precise set of probability functions in our representor, we cannot be guided by any theory of action under imprecise credences. Perhaps there are stylized, Ellsberg (1961)-esque contexts in which we do have this sort of access. We might know that a ball was drawn either from an urn with 70 white balls and 30 black or from an urn with 30 white and 70 black, say, while "not knowing anything about" the probabilities across the urns. We might then be guided by, say, the Sen-Walley maximality rule. But even if so, most contexts are obviously nothing like this. With respect to the potential consequences of giving to Malaria Consortium, I do not have access to a probability distribution, and I do not have access to a representor. I am clueless.

In sum: there is a diverse array of possible contexts we can find ourselves in. There are contexts in which we experience absolutely precise expected utilities for all the acts available to us, contexts in which we find ourselves frazzled, or befuddled, or drifting off to sleep, and contexts everywhere in between. When we experience expected utilities, we have a clear guide to rational action: "maximize expected utility". Theorists also sometimes suggest weaker normative restrictions on behavior, some of which are action-guiding in some somewhat more general

contexts. Ideally, we would like an action-guiding decision theory in every context we could possibly find ourselves in—or at least, in every context in which we are alert enough to think about and apply a decision theory. In particular, if we are consequentialists, we would very much like a compelling decision theory which guides us in contexts of cluelessness. So far, to my knowledge, no one has found one.

4 Coarse partitioning

I have not found one either. But here, I think, is something close enough to convey some idea of what one might look like.

As outlined above, the textbook model of probabilistic beliefs and expected-utility-maximizing behavior requires that your brain assign every available act an expected value out to infinitely many decimal places, and then find the act appealing—or "subjectively choiceworthy"—in proportion to this expected value. This is probably literally impossible. There are finitely many atoms in your brain; given some fact about the discreteness of space which I'm told most physicists believe these days, this implies that there finitely many ways your brain can be arranged in your head. Assuming that you cannot exhibit more mental states than brain states, there are therefore only finitely many degrees of subjective choiceworthiness you can exhibit with respect to the acts available to you. In other words, the extent to which we find acts appealing cannot be perfectly fine-grained.

Similar sorts of filtering happen everywhere. When we store data in a database, numbers are stored out only so many decimal places. Longer decimals are rounded off, and whatever precision we had before the rounding we then lose. If two stored decimals are equal after rounding, and the database is asked which was larger before the rounding, it cannot answer. Likewise, digital pictures of objects of slightly different sizes will "round to the nearest pixel". We, looking at the pictures on a screen, can't tell which object is bigger. Likewise, if we're looking through a sheet with five holes at an object behind it, we don't know how big the object is; we just know how many of the holes it covers. And so on.

The question, then, is not really whether our brains partition the space of possible acts into a finite number of buckets ("extremely subjectively choiceworthy", "very subjectively choiceworthy", etc.), thus rendering some acts equally subjectively choiceworthy even when their expected values would come to differ somewhat over the course of an indefinite period of reflection. The question is just how fine this partition is.

There is a literature on what are called "choice procedures": attempts to decompose the acts of reasoning and choosing into their empirical psychological components. See Apesteguia and Ballester (2012) for an example. I will now propose a basic sketch of a possible choice procedure.

Imagine that what happens when we evaluate acts is something like the following. We are exposed to a wide array of options and a confusing mass of reasons for and against choosing each of them. One option that is always available to us, as long as we are able to perceive

multiple options, is the option of pondering the options. Without yet having weighed any of the reasons for choosing one option over another, we dump all the non-pondering acts—I'll call these the "real acts"—into a single bucket, and the act of pondering stands out alone as the most appealing. After some pondering, some of the real acts stand out as better than the others. We now have three buckets. There's no sense pondering among the items in the bad bucket, but we feel we can do better than picking at random among the items in the good bucket, so pondering the good-bucket options still sits alone in the highest bucket, and we keep pondering. This eventually splits the "good bucket" into a better bucket and a worse bucket as well. And so on. Eventually the (intuited) costs associated with pondering outweigh the (intuited) benefits of further refining the top real bucket, and we pick an act at random from the top real bucket. The partition we wind up with as we carry out a real act, then, is usually fine across the most subjectively choiceworthy real acts, but coarse among the rest.

For illustration, consider what happens when you go to the store to buy jam. The set of acts available to you is something like "buy any basket of goods within your budget". On a moment's reflection, you rank acts of the form "buy one jar of jam" above all the rest. You then walk over to the jam shelf, and you start refining the "buy one jar of jam" partition: *Those are too expensive*, and *I like strawberry more than blueberry*, and so on. When you've narrowed it down to one, or to just a few you think it would be a waste of time to ponder further, you pick one of these at random. If, staring at the jam shelf, you were suddenly forced to choose instead between buying pasta and buying soap, you would feel a bit thrown. This, I propose, is cluelessness with respect to an act-set: the feeling of having to choose among acts which have all been relegated to the same partition-element, and the understanding that additional thought would refine this partition-element considerably.

Note the necessity of both ingredients. All the acts must belong to a single partition-element (i.e. it must not be immediately obvious to you whether to buy soap or pasta), and the option to ponder a moment must be much more subjectively choiceworthy than the option to choose one of the options immediately at random (i.e. it must *be* immediately obvious to you that you might have some use for soap but not pasta, or vice-versa). By contrast, suddenly being forced to choose between buying one lottery ticket and buying another would be mere "simple cluelessness". Here, as in the case of "soap or pasta", both acts belong to the same partition-element, though the acts' outcomes could vary substantially. Nevertheless, here you have an immediate sense that further pondering would not refine the partition-element, and that you should therefore choose immeidately at random.

In the face of cluelessness, if this account is correct, our best bet is to proceed as we proceed when buying jam. That is, we should ponder until it no longer feels right to ponder, and then to choose one of the the acts it feels most right to choose. Lest that advice seem as vacuous as "date the person who maximizes utility", here is a more concrete implicaton. If pondering comes at a cost, we should ponder only if it seems that we will be able to separate better options from worse options quickly enough to warrant the pondering—and this may take some time. Otherwise, we should choose immedately. When we do, we will be choosing literally at random;

but if we choose after a period of pondering that has not yet clearly separated better from worse, we will also be choosing literally at random.[2]

The standard Bayesian model suggests that if we at least take a second to write down immediate, baseless "expected utility" numbers for soap and pasta, these will pick the better option at least slightly more often than random. The cluelessness model sketched above predicts (a falsifiable prediction!) that there is some period—sometimes thousandths of a second, but perhaps sometimes thousands of years—during which these guesses will perform no better than random. Likewise, Bayesian intuitions tell us that remembering a one-off fact that compares soap to pasta along a single dimension, such as taste, will increase the odds that we choose correctly. My intuition is that it will do nothing. An incomplete comparison of this sort is useful if it contributes to a more complete comparison later on, when we have compared the options along enough dimensions that one of them has actually grown more subjectively choiceworthy. But on its own, if it leaves both options in the same bucket, it is worthless. It doesn't give us a clue.

5 Conclusion

If something like the above is correct, we are all always clueless with respect to almost all act-pairs. We are clueless, presumably, among most of the commodity-bundles we could buy every time we enter the grocery store, let alone between arbitrary possible act-pairs, like causing three inches of extra rain on the North Pole and moving Andromeda three inches closer to the Earth. Why, then, do we typically not notice our pervasive cluelessness? Why is it so rare to see concern for the phenomenon of cluelessness, or calls for an action-guiding decision theory in contexts of cluelessness, outside conversations about consequentialist ethics? Why is it so common in the Effective Altruism movement in particular? And as clueless consequentialists in 2019, how long shall we ponder?

Many of the answers, I think, fall directly out of the above formalization. In most decision contexts we find ourselves in, we have already processed the information relevant to the actions available to us until the maximal-subjective-choiceworthiness bucket contains just a few similar options. We only notice that we are in contexts of cluelessness, and we only feel the need for normative guidance in navigating contexts of cluelessness, when we find ourselves actively pondering a large and diverse maximal-subjective-choiceworthiness bucket for a long time. Consequentialism offers atypically tangled webs of reasons for and against choosing each available act, so in the context of consequentialist decision-making, the brain takes atypically long to sort through the options—to compute subjective choiceworthiness out n decimal places, as it were. The top bucket stays big, and the problem stays worth pondering, for longer.

[2]To be precise, note that I am attributing the coarse partitioning of acts to limitations of the brain, rather than to the information set or to any other objective feature of the context. Cluelessness is therefore subjective; one agent might be clueless with respect to a pair of acts while another agent facing the same evidence might not.

As for how long we'll ponder: who knows? A pessimistic possibility is that the evidence bearing on our actions' long-term consequences is so complex, and our reasoning tools are so limited, that we'll have to ponder on a cosmic timescale. But a more optimistic possibility, to which I am more sympathetic than I once was, is that the long experience of cluelessness a modern consequentialist faces is due primarily not to the hopeless complexity of his decision problem, but just to the fact that he recently found that he had to decide among options he had long relegated to a large, low-subjective-choiceworthiness bucket. The observation that consequentialists must optimize for long-term impact comes sudden and jarring, like an announcement that we have to walk home from the grocery store with something very different from what we went in to buy. But eventually, this story goes, we can change focus and narrow down the vast space of available acts in a different way. It will still take a while, since there are quite a lot of options and the problem is quite difficult, but this period of "cluelessness" (rather than the mere wide uncertainty) will not last as long as it first threatens to. Perhaps its end is already near. The top buckets, where our options are finely partitioned, will soon come to consist of (say) individual research projects to consider funding, rather than GiveWell top charities. We will remain clueless about the GiveWell charities, as we have always been about almost everything, but this will no longer be unsettling. In some minimal sense, at least, we will know what to do.

Either way, though, I don't think we are best served in the meantime by the unfollowable order to "just maximize expected value". If we want to recover as quickly as possible, we should at least admit we have a problem.

References

[1] Apesteguia and Ballester (2012) "Choice by Sequential Procedures". Barcelona GSE Working Paper No. 615.

[2] Burch-Brown, J. M. (2014). "Clues for Consequentialists", *Utilitas* 26(1): 1–15.

[3] Cowen, T. (2006). "The Epistemic Problem Does Not Refute Consequentialism", *Utilitas* 18(4): 383–399.

[4] Dorsey, D. (2012). "Consequentialism, Metaphysical Realism and the Argument from Cluelessness", *Philosophical Quarterly* 62(246): 48–70.

[5] Elga, A. (2010). "Subjective Probabilities Should Be Sharp", *Philosophers' Imprint* 10(5): 1–11.

[6] Ellsberg, D. (1961). "Risk, Ambiguity, and the Savage Axioms", *The Quarterly Journal of Economics* 75(4): 643–669.

[7] Friedman, M. (1953). *Essays in Positive Economics* (Chicago: University of Chicago Press).

[8] Gilboa, I. (2010). *Rational Choice* (Cambridge, MA: The MIT Press).

[9] GiveWell (2019). "Your Dollar Goes Further Overseas".
URL: `https://www.givewell.org/giving101/Your-dollar-goes-further-overseas`.
Accessed 10 May 2019.

[10] Greaves, H. (2016). "Cluelessness", *Proceedings of the Aristotelian Society* 116(3): 311–339.

[11] Kahneman, D., and A. Tversky (1979). "Prospect Theory: An Analysis of Decision under Risk", *Econometrica* 47(2): 263–291.

[12] Lenman, J. (2000). "Consequentialism and Cluelessness", *Philosophy and Public Affairs* 29(4): 342–370.

[13] McFadden, D. (1973). "Conditional logit analysis of qualitative choice behavior", in P. Zambreka (ed.), *Frontiers in Econometrics* (New York: Academic Press), 105–142.

[14] Mogensen, A. (2019). "Deep uncertainty about doing the utmost good", manuscript in preparation.

[15] Savage, L. J. (1954). *The Foundations of Statistics* (New York: Wiley).

[16] Simon, H. (1957). "A Behavioral Model of Rational Choice", in *Models of Man, Social and Rational: Mathematical Essays on Rational Human Behavior in a Social Setting* (New York: Wiley).

[17] Smart, J. J. C., and B. Williams (1973). *Utilitarianism: For and Against* (Cambridge, UK: Cambridge University Press).

[18] Tversky, A., and D. Kahneman (1992). "Advances in Prospect Theory: Cumulative Representation of Uncertainty", *Journal of Risk and Uncertainty* 5(4): 297–323.

[19] Walley, P. (1990). *Statistical Reasoning with Imprecise Probabilities* (London: Chapman and Hall).