Consequentialism, Cluelessness, Clumsiness, and Counterfactuals

Alan Hájek (Australian National University)

Global Priorities Institute | March 2024

GPI Working Paper No . 4-2024

Please cite this working paper as: Hájek, A. Consequentialism, Cluelessness, Clumsiness, and Counterfactuals. *Global Priorities Institute Working Paper Series, No. 4-2024.*
Available at: https://globalprioritiesinstitute.org/consequentialism-cluelessness-clumsiness-and-counterfactuals-hajek

GLOBAL PRIORITIES INSTITUTE

UNIVERSITY OF OXFORD

Consequentialism, Cluelessness, Clumsiness, and Counterfactuals

Alan Hájek[1]

Introduction

According to a standard statement of objective consequentialism[2], a morally right action is one that *has the best consequences*. More generally, given a choice between two actions, one is morally better than the other just in case *the consequences of the former action are better than those of the latter*. (These are not just the immediate consequences of the actions, but the *long-term* consequences, perhaps until the end of history.) This account glides easily off the tongue—so easily that one may not notice that on one understanding it makes no sense, and on another understanding, it has a startling metaphysical presupposition concerning counterfactuals. I will bring this presupposition into relief. Objective consequentialism has faced various objections, including the problem of "cluelessness": we have no idea what most of the consequences of our actions will be. I think that objective consequentialism has a far worse problem: its very foundations are highly dubious. Even granting those foundations, a worse problem than cluelessness remains, which I call "clumsiness". Moreover, I think that these problems quickly generalise to a number of other moral theories. But the points are most easily made for objective consequentialism, so I will focus largely on it.

I will consider three ways that it might be rescued:

1) Appeal instead to the *not-too-specific*, *short-term* consequences of actions;
2) Understand consequences with *objective probabilities*;
3) Understand consequences with *subjective/evidential probabilities*.

We will see how central to moral philosophy are foundational issues in probability and counterfactuals.

[1] For very helpful discussion and comments, I am grateful especially to Selim Berker, David Builes, Chris Bottomley, John Cusbert, Justin D'Ambrosio, Nick DiBella, Nicky Drake, Adam Elga, Max Fedoseev, Dmitri Gallow, Brian Hedden, Mikayla Kelley, Daniel Kilov, Alexander Kocurek, Boris Kment, Harvey Lederman, Andrew Lee, Leon Leontyev, Daniel Munoz, Makan Nojoumian, Daniel Nolan, Doug Portmore, Theron Pummer, Pamela Robinson, Geoff Sayre-McCord, Nick Schuster, Wolfgang Schwarz, Nic Southwood, Katie Steele, Daniel Stoljar, Jeremy Strasser, Kramer Thompson, Joshua Thong, Peter Vranas, Isaac Wilhelm, Hayden Wilkinson, Patrick Williamson, Timothy L. Williamson, and audiences at Australian Catholic University, Australian National University, Effective Altruism ANU, Princeton University, Reed College, Stanford University, University of Sydney, University of Washington, the ANU/University of Hawaii workshop, King's College London, London School of Economics, the National University Singapore Presidential Conference on Formal Epistemology, the 2023 Global Priorities Institute Workshop, Oxford, and the Australasian Association of Philosophy 2023 conference.

[2] More specifically: objective *act* consequentialism (as opposed to *rule* consequentialism).

Actual or counterfactual consequences?

Julian Savulescu and Dominic Wilkinson (2019) write: "According to consequentialism, the right act is that act which **has the best consequences**."[3] Similarly, Julia Driver (2012) writes: "consequentialism is the view that the moral quality of an action – for example, the rightness of the action – is completely determined by the action's consequences, relative to **the consequences of alternative actions** open to the agent." We nod along with this statement, even if we think there are problems with the view. Hilary Greaves (2016) characterises (objective) consequentialism thus: "the moral status of an action is determined entirely by how it compares to alternative actions in terms of the goodness of its consequences." And she says more generally: "A1 is objectively c-better than A2 iff **the consequences of A1 are better than those of A2**" (312). ("c-better" is her term of art for the comparison that will matter to consequentialists.) Note that this implies that both A1 and A2 actually *have* consequences. Such statements of consequentialism are entirely familiar. But do they make sense?

Consider a moral choice that you face: you could help an old lady across the street, or you could go to the pub. Let's suppose that in fact you help the old lady. Did you do the right thing? I'm happy enough to allow that there's a fact of the matter of the total value of the consequences of your action.[4] But what about the thing that you did *not* do? You did not go to the pub; how good are its consequences? Taken literally, this does not make sense: there *are no* consequences of an action that is not performed. Immediately we encounter a problem with the very statement of objective consequentialism, and its generalisation (comparing one act with another): they do not make sense as stated so far. At a given choice, you actually perform only

[3] All **boldings** are mine, here and elsewhere.

[4] Maybe I shouldn't be. What are the *consequences* of your action, as opposed to things that merely happen later? For example, one might think that the consequences of your action (in the morally relevant sense) are the things that *would not* have happened *if* you had acted differently—a counterfactual notion. But then the kinds of problems that I am about to raise for counterfactuals will kick in.

Also, there is a concern that the one action that you actually perform is judged by different standards—its *actual* consequences—from all your other possible actions, which are judged *counterfactually*. (The statement of consequentialism that I am about to consider treats all actions uniformly—counterfactually—with no special treatment of *actual* actions.) Now, it might seem that for the action *A* that you actually performed, with its actual consequences *C*, we get the corresponding counterfactual for free: if you had performed *A*, its consequences would have been *C*. After all, you did indeed perform *A*, and its consequences were indeed *C*. But I question the inference from this conjunction to the counterfactual. More generally, I question the validity of the and-to-if schema:

$\underline{p \,\&\, q}$

$\therefore\ p \mathrel{\square\!\!\rightarrow} q$

To be sure, it is endorsed by Stalnaker (1968) and Lewis (1973), among others. It corresponds to 'strong centering' of the similarity relation. But it is also opposed by McDermott (xx), Vessel (2003)—see footnote 12—and others.

one action, so only it *has* consequences.[5] Moreover, being the *only* action that has consequences, it trivially has the *best* consequences—and the *worst*!

So it seems that the statement should instead involve *counterfactual* actions, and the consequences that they *would* have *if* they were performed—I think that is the most charitable understanding. Or we might stipulate a technical usage of "consequences" which builds in their counterfactuality by fiat. (Decision theory sometimes employs such a technical usage—see e.g. Savage 1954.) In any case, let's say it more carefully: a morally right action is one that *would* have the best consequences *if* it were performed. For example, here is Tännsjö's (2013) statement of utilitarianism:

> an action is right if and only if in the situation there was no alternative to it which **would have resulted** in a greater sum total of welfare in the world... This means that if there was something the agent **could have done** instead of the action he or she actually performed which **would have resulted** in a greater sum total welfare in the world, then he or she acted wrongly. (18)

And given a choice between two actions, one is morally better than the other if and only if the consequences of the former action, if it were performed, *would be* better than those of the latter, if *it* were performed. Again, this glides off the tongue—all too easily. Again, I think there is an underappreciated but fatal problem. I will develop the problem in two ways, depending on whether the world is indeterministic or deterministic after the time at which the choice of action takes place.

Indeterminism

First, let's assume that the world is indeterministic thereafter. Indeed, that is rather plausible—it seems that the world has myriad chancy processes, such as coin tosses, lotteries, and various natural processes, not to mention quantum mechanical events. Focus on an action that you did *not* perform—in this case, going to the pub. It is merely hypothetical. *If* you had gone to the pub, how *would* the first non-actual chance process thereafter have turned out? To fix our ideas, let it be the first non-actual coin toss thereafter.[6] 'If you had gone to the pub, the coin *would* have landed *heads*, not tails!' That was a joke—I want you to be struck by the implausibility of this claim, and perhaps even to laugh at it. 'No; if you had gone to the pub,

[5] Marcus Singer (1977) makes this point, although he partly obscures it by saying that "actual consequence utilitarianism requires you to **know** the actual consequences of acts never performed, to compare with the actual consequences of the one performed" (72). Let me emphasize that the point has nothing to do with knowledge—it is about *the consequences themselves* not existing.

[6] I like the simplicity of this example, but if you think there are less controversial examples of chancy processes, feel free to replace it with one of those.

the coin *would* have landed *tails*, not heads!' That was another joke—equally implausible, equally laughable, or so I say. After all, if the coin had been tossed, it *might not* have landed heads, and it *might not* have landed tails—that's what chance is all about. (I don't actually need the appeal to the 'might not' counterfactuals to make the point—chanciness undermines the counterfactuals, especially when the chance is nowhere near 1. But the 'might nots' let me put the point quickly.)

While I disagree with the usual Lewis-style (1973) semantics for counterfactuals, it's a useful heuristic that helps to convey my point. On this approach, the counterfactual 'if it were that *p*, it would be that *q*' is true just in case *all* the most similar *p*-worlds are *q*. Among the most similar worlds in which you go to the pub, some are 'heads' worlds and some are 'tails' worlds—they are not unanimous either way. So it's false that the coin *would* have landed heads, and false that it *would* have landed tails.

Or consider the first (non-actual) lottery to be played after your (non-actual) trip to the pub. Suppose that it would have had some large number of tickets, numbered 1, 2, 3, … 'If you had gone to the pub, ticket #17 *would* have won!' That was yet another joke. And it remains a joke whichever number is claimed to be the hypothetical winner, rather than some other. For any ticket, I say that it is false that *that* ticket *would* have won, if you had gone to the pub. It might not have. In similarity-of-worlds speak: the most similar worlds in which you go to the pub do not all agree on which ticket wins.

I say that the counterfactuals are *false*. But an important alternative view says that they are *indeterminate*. Stalnaker (1981) thinks that there is always a unique most similar ("closest") antecedent world, but in cases like these it is indeterminate what it is. He supervaluates over all of the arbitrary selections of the unique world. For each ticket, a world where *that* ticket wins gets selected by *some* admissible valuation. But since the valuations disagree, there is no fact of the matter of what the winning ticket would be.

I don't want to insist on my view that the relevant counterfactuals are *false*; it suffices that Stalnaker and I agree that the counterfactuals are *not true*. Indeed, Stalnaker can laugh with me at them. For him, it's rather like laughing at "Sherlock Holmes had an odd number of hairs (not even!) at Reichenbach Falls", which is indeterminate on a popular view about truth in fiction. Indeterminacy is problematic enough for objective consequentialists to the extent that they are committed to there being facts of the matter for such counterfactuals. Well, perhaps they could think that there is rampant indeterminacy about whether a given action is morally better than another in almost all cases. But that's a huge bullet to bite: it is often platitudinous what the

right verdict is. For example, it's *true* that donating to Oxfam is morally better than going on a serial killing rampage. Any view that says otherwise is absurd, and perhaps even pernicious.

But I have barely begun. There has recently been a cottage industry of observing just how significant an effect on subsequent history our actions have—even mundane or trivial actions. They have ripple effects. And these are not merely like the ripples on a pond caused by a stone throw, which dampen down and quickly disappear. On the contrary, the ripple effects of our actions are ongoing and amplify. Our actions affect the identities of future people for the rest of history. These identities depend on the fine details of which sperm happens to fertilise which egg on particular occasions; changing the details changes the nature of the conceptions, and hence the identities of the children that are subsequently born. For example, a tiny change in your action may just slightly *delay* the moment of a conception, and that's enough to change the identity of the future child. (See e.g. Parfit 1984, Ch. 16; Greaves 2016.) It is plausible that such conceptions are yet more chance events, more lotteries in a broad sense.

So if you had gone to the pub, consider the first (non-actual) child to be conceived thereafter. I say that it is false of any particular sperm that *it* would have won its race to an egg, its lottery—false that *this* child would have been conceived rather than some other. But I have barely begun. Now consider that hypothetical child's children, and grandchildren, and great grandchildren, and so on, for the rest of history. Consider all the (non-actual) people's interactions with these hypothetical descendants, and all *their* hypothetical descendants' interactions. I find it highly implausible that there is a truth of how all of these chance events *would* have turned out, if you had gone to the pub. But I've still barely begun. It is even less plausible that there is a truth of what the magnitudes and durations *would have been* of all the joys and sorrows of all these non-existent people, or of whatever else count as 'consequences'—after all, these depend on the resolutions of yet more chance processes.

We can all agree that for any action you might perform, there is a strongest true proposition concerning what would have happened. According to one extreme view, this proposition is as strong as (the singleton set of) a single world—soon we will critically discuss this view (under the heading 'counterfactism'). The objective consequentialism that I have been considering, which I take to be quite standard, need not be that extreme, but it is still implausibly close to that extreme, committed to implausibly strong claims about what would have happened. (Again, it had better not settle instead for rampant indeterminacy.) According to another extreme view at the other end of the spectrum, this strongest true proposition is no stronger than a necessary proposition. I think the true view is a moderate one that lies somewhere between these extremes. In particular, while this strongest proposition may resolve various

matters, it does not resolve the outcome of a chancy process (at least where the chances are not near 0 or 1).

For example, suppose a psychopath contemplates hooking up a doomsday device to be activated if a particular fair coin landed heads, but not if the coin landed tails.[7] What *would* happen if he were to perform this action? It's not the case that billions of people *would die*—they might, they might not, each scenario happening with chance 1/2. Objective consequentialism formulated counterfactually fails to capture the wrongness of the action.[8] In fixating on what *would* happen, this consequentialism neglects what *might* happen, and more importantly, the associated chances.

This poses a dilemma for objective consequentialism formulated counterfactually. If it admits that the relevant counterfactuals are not true, then it will go silent in many chancy cases where it should speak, and even shout from the rooftops: its ethics is compromised. If it insists that the relevant counterfactuals are true, as I believe it standardly does, then its metaphysics is compromised.

Let's pursue the latter horn further.

Determinism

So far I've been assuming indeterminism after your pub jaunt; now let's assume determinism, which you might think is the best hope for objective consequentialism. Now, given a precise specification of the initial conditions and the laws of nature, we get an entire history determined thereafter. From a snapshot of the world at the time of your choice of going to the pub, the rest of what happens in the world lawfully follows. Now it might seem more plausible that there's a truth of what would have happened for the rest of history if you had gone to the pub. But now a different problem kicks in: the unspecificity of the antecedent. "If you had gone to the pub …"—somehow or other. Well, how exactly? Now I find it implausible

[7] I am grateful to Wolfgang Schwarz for the argument here and the dilemma in the next paragraph.

[8] Consequentialists can point out that billions of people *would be exposed to a substantial risk of death*. But not hooking up the doomsday device would not expose anyone to such a risk. And so the platitudinous verdict can be upheld. But here is a response. Suppose the psychopath did the deed. People would with certainty be exposed to a risk of death (and that's bad), and moreover, there would be a 1/2 probability that they'd actually die (and that's worse). The concern is that, if we're just pointing to the badness of the risk-exposure, then we'll only say that setting up the doomsday device only has the smaller badness: that of the risk. So only attending to the badness of the thing that would be true, were you to flip undervalues the badness of the act. It ignores the things that merely might be true, were the device set up. Consequentialism formulated counterfactually can only account for the badness of the risk and not the badness of the possible outcome and so must undervalue the risky action's badness. (Thanks to Mikayla Kelley for suggesting how to uphold the verdict, and to Dmitri Gallow for the response.)

that there is a particular way that this loosely specified hypothetical scenario *would* be realised.[9] 'If you had gone to the pub, you would have entered it at 6:03 pm, 24 seconds, and *17* milliseconds.' That was a joke! 'No; you would have entered it at 6:03, 24 seconds, and *18* milliseconds.' Another joke! Again, I am drawing attention to the implausibility of these counterfactuals: they are *implausibly specific*, hitching together an inexact antecedent with an all-too-exact consequent. If you had gone to the pub, you *might not* have entered it at 6:03 pm, 24 seconds, and 17 milliseconds; so it's false that you *would* have done so. And so it goes for any putative exact entry time. Said in Lewisian terms: the most similar 'pub' worlds do not all agree on your exact arrival time.

But yet again, I have barely begun. There is no truth of who the first non-actual child to have been conceived would have been, no truth of which sperm would have fertilised which egg, if you had gone to the pub (somehow or other). And so on, for the subsequent non-actual lineage of this non-actual child, and all the non-actual people with whom they would interact, and all the non-actual people with whom *they* would interact, and so on.

Now, an objective consequentialist might reply that they need not commit to a specific counterfactual history if you had pubbed—a range of counterfactual histories is fine, as long as the total value of their consequences is sufficiently delimited. For example, suppose all of them would have consequences whose total value falls below the (actual) total value of your old-lady-helping's consequences. Then you did the right thing. However the chips might have fallen, helping the old lady was better.

But this is a forlorn hope. Our everyday actions may have radically divergent effects on the total values of subsequent histories—this is bread and butter for the cottage industry. In one hypothetical history after your hypothetical pubbing, the children that happen to be conceived are a series of latter-day Buddhas and Einsteins, and a wonderful world ensues. But with just a tiny tweak to the details of your pubbing, we get another hypothetical scenario in which the children that happen to be conceived are latter-day Hitlers and Stalins, and a nightmarish world ensues. And with other tiny tweaks, we get a broad spectrum of total values in between. Some of these scenarios are comparable in total value to that actually realised by your helping the old lady, some are much worse, and some are much better. So there are not even moderate bounds

9 Cf. Hare's (2011) discussion of a Wheel of Fortune with alternating red and black sectors—the result of a spin is determined by the precise force imparted to it initially. Consider the question "What would have happened if you had spun the wheel?" Hare replies: "because the condition 'if you had spun the wheel' is underspecified, there is no fact of the matter about whether you would have gotten a red, if you had spun the wheel, and no fact of the matter about whether you would have gotten a black, if you had spun the wheel".

on the total value that would have resulted if you had gone to the pub, and there is no truth of whether helping the old lady was better or worse.[10]

Yet as before, the exact details make all the difference. As it might be: if you had entered the pub at the 17-millisecond time, a series of moral saints and geniuses would have been created; but if you had entered the pub at the 18-millisecond time, a series of moral monsters and charlatans would have been created. (Take these claims with the generous lumps of salt that they deserve, but they should convey my point for now. Soon I will throw more salt of my own at them.) There is acute sensitivity to the exact initial conditions of the value of what would follow thereafter; indeed, it suffices that there *might* be such acute sensitivity. So even under determinism with an unspecific starting point, there is no fact of that value, nor even moderate bounds on it. Instead, there is a vast portfolio of live possibilities regarding what it might be. The exact details make a world of difference—we get huge world-differences.

As I have been understanding objective consequentialism, it is committed to the truth of a staggering set of counterfactuals concerning the consequences of non-actual actions. I say that such counterfactuals are false; followers of Stalnaker say that they are indeterminate; either way, they are *not true*. Such counterfactuals had better not be the foundation of morality.

Cluelessness and clumsiness

At this point, a consequentialist might reply that there *is* a truth of how exactly you would have entered the pub—the exact initial conditions associated with your pubbing, from which the rest of (counterfactual) history deterministically follows. Let's suppose this, for the sake of the argument now (I won't be so concessive soon). There is still a serious problem for objective consequentialism, one worse than I think has previously been recognised.

The recent literature on objective consequentialism makes much of the so-called problem of *cluelessness*: we can never have the faintest idea which action would have the best consequences. (Lenman 2000, Cowen 2006, Burch-Brown 2014, Greaves 2016, Mogensen 2020.) Let's suppose that your 17-millisecond arrival time would initiate a history rich with latter-day Buddhas and Einsteins, while your 18-millisecond arrival time would initiate one darkened by latter-day Hitlers and Stalins. The problem is supposed to be that we could never *know* these facts, or have justified belief in them. This is an *epistemic* problem.

[10] In this case, we are comparing an actual action (helping the old lady) with a non-actual action (going to the pub). Still less is there a truth of which of *two non-actual* actions would be better. Then we have *two* large spans of possible histories thereafter, and *two* broad spectra of their total values. Still less do we have one action determinately better than the other (one spectrum sitting entirely above the other).

But this understates how bad the situation is for objective consequentialism. For suppose that we could somehow solve the problem of cluelessness—say, God tells you these alleged facts. Then what? The trouble is that it is simply *not under your control* to realise these conditions in one precise way rather than another. Much as you may want to arrive at the 17-millisecond time (say), you cannot so finely tune your actions so as to do so, rather than arriving at the 18-millisecond time. You are *clumsy*. When it comes to these extremely fine-grained actions, you are a klutz. By the standards of acute sensitivity to the exact initial conditions of subsequent history, you are ham-fisted, unable to steer things exactly *this* way rather than a closely-neighbouring *that* way. These exact arrival times are not genuine *options* for you: you cannot decide to realise one rather than another. And the "actions" that consequentialism evaluates should not be mere behaviours; they should be options that you can decide among. What God tells you is not *action*-guiding in a sense that we should care about.

As usual, I have barely begun. What would happen immediately after your arrival? Suppose that if you were to move your hands in exactly *such-and-such* a way for the next second, a Pan-glossian world would follow; but if you were to move them in an adjacent *so-and-so* way, a Pain-lossian world would follow. But you do not have such fine motor control over your hands as to direct them the first way rather than the second. And so it goes for the rest of your bodily movements, for the rest of your time in the pub—and thereafter.

We can put this as another dilemma, depending on how fine-grained the objects of moral evaluation are. First horn of the dilemma: suppose that they are somewhat coarse-grained things, like 'you go to the pub'. Then it is plausible that that these really are options of yours. But it is *not* plausible that there is a single complete counterfactual history thereafter, given the unspecificity of such options. Second horn of the dilemma: suppose the objects of moral evaluation are precisely specified 'options', like

> 'you go to the pub at 6:03, 24 seconds, and 17 milliseconds (not 18!) and move your hands in *such-and-such* a way for the next second (not *so-and-so*!), and …'.

Then it is perhaps more plausible that there is a truth of the entire counterfactual history thereafter. (I am about to question even this.) But it is *not* plausible that these 'options' are things that you can decide among; they are not genuine *options* that you can deliberately realise or not by an act of your volition. (Hence my scare quotes around 'options'.) Your clumsiness

renders you unable to realise one of these 'options' instead of its near-neighbours by an act of your will. Either way, objective consequentialism founders.

There is no 'sweet spot' between the dilemma's horns at which objective consequentialism might find refuge. Indeed, the horns overlap.[11] It is plausible that the objects of moral evaluation are rather fine-grained options. For example, you can choose to enter the pub not merely somehow or other, but in rather specific ways—say, within intervals of a second. As such, these really are options for you. But any finer graining is beyond your control—for example, entering within intervals of a tenth of a second. At this level of resolution you are clumsy, and these are no longer genuine options for you. But still these 'options' are too coarse-grained to yield unique counterfactual histories, even under determinism. There is still enough wiggle-room even at this level of resolution to allow radically divergent counterfactual histories, with a vast range of possible consequences. Thus, far from finding a 'sweet spot' here, this level of resolution lands consequentialism on both horns.

Here is another way to put the dilemma. A counterfactual that mismatches an unspecific antecedent with a specific consequent is false, I say. We can alleviate this mismatch by either making the antecedent more specific (strengthening it), or by making the consequent less specific (weakening it). The stronger we make the antecedent, the less wiggle-room there is in what happens subsequently. (At an extreme, we specify the initial conditions of your pubbing *to infinite precision*, maximally constraining what would happen thereafter under determinism.) But then the problem of clumsiness becomes progressively worse. (At this extreme, it is maximally implausible that you can fine-tune your pubbing with such precision.) The weaker we make the antecedent, the more possible future histories are left open. (At the other extreme, we leave the initial conditions of your pubbing maximally unspecific: *you do so somehow or other*.) But then it is less plausible that there is a truth about the right thing to do. (At that extreme, the span of possible future histories is greatest, including wonderful ones that are superior to that of helping the old lady, and horrific ones that are inferior, and consequentialism gives no verdict on what you should do.) Summarising this dilemma: The more we strengthen the antecedent, the more we generate the problem of clumsiness; the more

[11] Thanks here to Wolfgang Schwarz.

we weaken the consequent, the less plausible it becomes that there is a truth of consequentialism's verdicts.[12]

It's not just about you

But yet again, I have still barely begun. I have engaged in an act of pretence in order to convey a point, but the reality is worse. So far I have pretended that under determinism, an exact specification of your pubbing would determine a unique counterfactual history thereafter. Hilary Greaves (2016) explicitly endorses a generalisation of this thought:

> "Assume determinism. Then, for any given (sufficiently precisely described) act A, there is a fact of the matter about which possible world would be realised – what the future course of history would be – if I performed A."

But this is far from clear to me. For you are just a tiny part of the world; in the sweep of world history, you are just a speck. (I'm sorry if this comes as news to you!) Even describing precisely the details of your going to the pub—the millisecond of your arrival, your exact hand movements for the next second, etc.—falls stupendously short of determining the *entire world's* initial conditions at that time. And under determinism, it's the initial conditions of *the entire world* and the laws that entail the rest of history. The initial conditions of *a miniscule*

[12] Vessel (2003) discusses an example in which determinism is assumed, and a powerful demon approaches a utilitarian called "Sam" and invites him to flip a fair coin, with radically different consequences depending on whether Sam accepts or not, and if he does, how the coin lands. Whether he acted in accordance with the utilitarianism that he endorses turns on the truth values of:

"If Sam were to flip the coin, then it would come up heads."

"If Sam were to flip the coin, then it would come up tails." (105)

Vessel points out that "Lewis's account entails that neither of the counterfactuals is true" (110). Firstly, the common antecedent is "extremely underspecified": there are many different ways of fully specifying the initial conditions of Sam's hypothetical flip, some leading to heads and some leading to tails at the most similar antecedent-worlds. Secondly, "given Sam's inability to ensure that his coin tosses produce the results of his choice, there don't appear to be any factors that would influence the similarity relation to grant any special priority (or 'closeness') to heads-worlds over tails-worlds" (109). More generally, Vessel observes: "Humans are, by and large, clumsy animals. We simply don't have the ability to ensure the outcomes of our choices in many cases." (110-111).

I agree with all these points—indeed, I generalize and strengthen them in several ways. I don't assume Lewis's account (on the contrary!), or indeed any particular account of counterfactuals. The counterfactuals are false (not merely not true, I say) because the antecedent is too weak to rule out live alternatives to the comparatively strong consequents: if Sam were to flip the coin, it *might not* come up heads, and it *might not* come up tails. The falsehood of act-consequence counterfactuals generalizes beyond far-fetched cases to *all* cases in the long term, including mundane choices such as helping an old lady vs going to the pub, both under determinism and indeterminism. And clumsiness figures in my central dilemmas for objective consequentialism. But I part company with Vessel on an important point. I have emphasized how these observations spell the downfall for objective consequentialism, as formulated with such counterfactuals. Vessel, by contrast, is sympathetic to objective consequentialism, and his focus is entirely different: indeed, his main concern is to argue *from* objective consequentialist reasoning to the failure of implication from p & q to $p \mathrel{\square\!\!\rightarrow} q$, and for a Lewis-style semantics for counterfactuals with weak rather than strong centering. That is, he takes objective consequentialism as a *premise* and argues for a conclusion concerning a property of the similarity relation and the associated counterfactual logic—a project completely different from mine, and indeed at odds with mine.

part of the world, such as the arrival of one person in a pub, does not even come close to being a sufficient input for such an entailment. After all, what's going on elsewhere would also play a huge role in determining the identities of the hypothetical people that we are supposed to imagine. And these hypothetical people would interact in all sorts of ways, creating more hypothetical people who would interact in yet other ways, and so on. And never mind the people—don't forget about the hypothetical dogs and frogs, bees and trees, photons and protons, …, not to mention hypothetical pandemics and natural disasters. This is an even more dramatic version of the problem of *unspecificity* of the counterfactual's antecedent. To get an entire history to follow from the specification of initial conditions under determinism, you would need to specify an entire time-slice of history—not merely some tiny part of a portion of a fragment of a time-slice of history: the details of your pub-going. Even the exact specification of your pub-going is just a miniscule sliver of the slice that's required.[13]

[13] But Greaves has a seemingly strong reply. When tweaking things so that you hypothetically go to the pub rather than help the old lady, we should hold fixed most of what actually happens. We imagine a local change that secures the truth of the counterfactual's antecedent, but everything else that is not impacted by that change is kept the same. This is like the reasoning that underpins counterfactuals of the kind made famous by Morgenbesser. A fair coin is about to be tossed. You are offered a bet on heads, but you decline; the coin is then tossed, and it comes up heads. We say:

"If you had bet on heads, the coin would still have landed heads; so you would have won."

Similarly, if you had gone to the pub in an exactly specified way, almost everything else would still have happened as it actually did. Or so the reply goes.

Morgenbesser-style counterfactuals are too big a topic for a full discussion here, but some quick rebuttals are pertinent. Firstly, defenders of Morgenbesser-style intuitions typically assume that the things to be held fixed are 'causally independent' of the antecedent. But an action generates fluctuations in the surrounding electromagnetic field that are propagated at the speed of light, impacting everything in the action's light-cone. Never mind that the causal influence may be small (however we might characterize *that*); its mere existence defeats causal independence. 'Causal independence' is an *absolute term*, in Unger's (1975) sense (like 'flat' or 'certainty'). Much more is causally dependent on your actions than you might think. As I have said, the cottage industry recognizes this with its emphasis on how consequential tiny tweaks to one's actions may be.

Secondly, everybody faces the hard problem of how to think about counterfactuals under determinism. Take your pick: if you had gone to the pub, either a law of nature would have been broken, or the initial conditions of the world and all subsequent history would have been different. Each option seems crazy: it's highly unintuitive that such a minor change in what actually happens would have such huge repercussions! Lewis (1979) famously opts for the former solution. He posits a 'small miracle' shortly before the time of your choice. That may seem reasonable on his deflationary view of laws of nature: they are merely regularities in the best systematisation of the world. Then it may not seem like a big deal to break a law: it just tweaks what these regularities are. However, on a beefier conception of laws as *governing* what happens, it is a big deal. (See Chen & Goldstein 2022.) How could your going to the pub result in a difference in one of *those*? Dorr (2016) opts for the latter solution. The Big Bang and everything thereafter would have been different, but minimally so among all ways that lead to your going to the pub in the exactly specified way. I find this more plausible: changing historical facts is not as radical as changing the laws. On this solution, we do *not* hold fixed most of what actually happens; on the contrary, we imagine changing all of it.

Either way, it is not clear that there is a fact of the matter of which possible world would be realized if you had pubbed. Either the laws (Lewis) or the initial conditions (Dorr) must be changed, but I question that there is a fact of the matter of exactly *how* they must be changed. There may be many ways of demarcating the exact extent of the 'small miracle', or of minimally changing the Big Bang, so as to get you to the pub in a particular way. (Compare: many roads lead to Rome. Seeing you at a particular spot in the Colosseum does not determine exactly how you got there.) And the different ways will have different ramifications for what happens

Previously I found it highly implausible that there is a fact of the matter of exactly *how* you would have entered the pub: 'If you had gone to the pub, you would have entered it at 6:03 pm, 24 seconds, and 17 milliseconds', and other jokes. (I didn't say they were *good* jokes!) Objective consequentialism à la Greaves is committed to something I find even more implausible: 'If you had gone to the pub, the entire world history would have been __' (with exactly one way of filling in the blank). Your action would make a tiny contribution to world history given everything *else* that would be going on. All the more I find it highly implausible that merely fixing your action nails down the rest of history. Again, this is not a problem of cluelessness, a problem that it is hard to know this history; it is not merely an epistemic problem. Rather, it is a metaphysical problem: there is no particular history that *would* ensue (although of course, this implies that there is nothing to know). The world itself is clueless.[14]

Greaves had us assume determinism, arguably the best case for there always being a unique world that would be realized if one performed some non-actual action; I have argued that this best case is not good enough for this metaphysical claim. (Of course, any *actual* action is realised in a unique world, the actual world; we need not assume determinism for that.) A number of other authors go even further, making or presupposing the claim on behalf of consequentialism even without assuming determinism.[15] All the more we should question this.

So I submit that that the metaphysical foundations of objective consequentialism are highly dubious. The statement of it may glide easily off the tongue, but when we scrutinise it, I think

---

subsequently. Then there is no fact of the matter about what the future course of history would have been if you had gone to the pub, *pace* Greaves.

This interacts with the problem of the unspecificity of the antecedent and the problem of clumsiness. There are many ways for you to go to the pub somehow or other. Different realisations of this antecedent, each of which *might* have occurred, correspond to different 'small miracles' or different changes to the Big Bang. (Rome is spread out. Some roads there lead to the Colosseum; others lead to St Peter's Basilica.) But the more specific the antecedent, the more the problem of clumsiness bites.

14 Lenman suggests this point: "Perhaps such talk of massively complex historical counterfactuals is metaphysical nonsense on stilts and there is nothing here for even God to know" (352). But he does not develop it beyond this sentence, his focus being on cluelessness instead. Much of this paper is a detailed defence of this point.

[15] Here is a small sample (not all of the authors are consequentialists, but they are stating what they take consequentialism to presuppose):

"It is plain that when we assert that a certain action is our absolute duty, we are asserting that the performance of that action at that time is **unique** in respect of value...It can, therefore, be unique only in the sense that the whole world will be better, if it be performed, than if any possible alternative were taken." Moore (1903, p. 147).

"The morally relevant outcome of an action is **the possible world** that would be actual if the action were performed." Carlson (1993, p. 10).

"The outcome of an act is **the possible world** that would be actual if the act were performed." Gustafsson (2019, p. 195).

"An act's outcome is **the possible world** that would be actual if it were performed." Portmore (2011 p. 34).

"It is *right* for S to do A (S *ought* to do A or S *should* do A) iff no **total state of affairs** that would be a consequence of S's doing any alternative to A would be better than **the total state of affairs** that would be a consequence of S's doing A." Sosa (1993, p. 101).

it falls apart. But suppose that my critique is mistaken, and that somehow or other the foundations are in good order despite everything that I have said. Then I still submit that objective consequentialism has jaw-dropping metaphysical commitments that have not previously been appreciated. We should have done a double, triple, and quadruple take at the very statement of it, rather than nodding along with its easy, breezy wording. And only then should we have begun to discuss familiar objections.

Counterfactism

Leaving moral philosophy for a moment, a number of authors in the counterfactuals literature go further still, assuming that a unique world would be realised *for any counterfactual antecedent*—not just antecedents concerning agents' actions. At numerous points I have been sharing with you my incredulous stares at the counterfactuals to which objective consequentialism is committed. But perhaps I have dismissed them too quickly. After all, a view about counterfactuals that needs to be taken seriously regards such counterfactuals to be in good order. According to the view,

> *for any antecedent A, there is an entire world w such that 'if A were the case, w would be the case' is true.*

This generalises Greaves' claim above in two ways: it does not assume determinism, and it quantifies over *all* antecedents *A*, not just those associated with an agent's actions. I call this (in my 2020) *counterfactual plenitude.*

Notice that counterfactual plenitude is incompatible with Lewis's (1973) rejection of the *limit assumption*: that for any *A*, there is at least one closest *A*-world. (Consider his putative counterexample: 'if I were taller than 7 ft, then I would be …'—there is an infinite sequence of ever-closer worlds where I am taller than 7 ft, but none closest, so there is no candidate for '*w*'.) But counterfactual plenitude has been defended by a most impressive line-up of philosophers. They go back to the medieval Molinists (who were primarily concerned with agents' actions—see Molina 1953, Suarez 1856-1878). In modern times they include the likes of Hawthorne (2005), Moss (2013), both of whom find inspiration in Stalnaker (1968); and also Schulz (2017), Stefánsson (2018), and to some extent Bradley (2012, 2017). Moreover, most of them (all but Stalnaker) are committed to what I call *Primitive Counterfacts Realism*:

> *There exist primitive modal facts that serve as truth-makers for all counterfactual claims.*

Call the conjunction of these italicised theses *counterfactism*, and proponents of it *counterfactists*.[16] I have argued (2020) that counterfactual plenitude entails primitive counterfacts realism, so really my target is counterfactism. And as I say, I take it seriously.

But what of my 'jokes': 'If you had gone to the pub, the coin *would* have landed *heads* (not tails!)', 'If you had gone to the pub, you would have entered it at 6:03 pm, 24 seconds, and 17 milliseconds (not 18 milliseconds!), and so on? I invited you to laugh with me at these counterfactuals. Counterfactists can laugh too, but not for my reasons. For them, the force of the jokes, to the extent that they have any force, is *cluelessness*. It's laughable to make such claims when one so obviously is not in any position to know them. The laughter should be directed at a pragmatic defect (unassertability) rather than a semantic one as I claim (falsehood). Indeed, the counterfactuals themselves may well be true, and in any case counterfactuals just like them *are* true: it's just a matter of getting the details in the consequent *right*. That is, one just has to state the counterfact that obtains. Of course, doing so is beyond our ken—we're clueless. But that's an epistemic problem, not a metaphysical problem. Or so say counterfactists. It's rather like epistemicism about vagueness, according to which there is a fact of the matter of how many grains of sand mark the sharp boundary between 'non-heap' and 'heap', although we can never know what this number is (and it would be laughable to assert what it is).

I devote an entire paper (2020) to presenting arguments for counterfactism, and then arguing against it. I cannot reprise all of that material here. But one important argument for it is based on a suggestive analogy between future contingents and counterfactuals. Most of us think that there are various true statements about the future: 'the sun will rise tomorrow', 'the world's population will grow', and what have you. Moreover, many of us think that such statements can be true even if the propositions expressed are chancy. For example, there is *some* chance that the sun will explode in the next few hours and *not* rise tomorrow (and the world's population will not grow soon thereafter!). Nevertheless, assuming that it in fact does, as is extremely likely, the prediction about its doing so is true now. Even more obviously chancy claims, like 'this coin toss will land heads' may be true now, provided the chanciness is resolved as claimed. We might think of there being various possible futures, which we might model as a tree, and a 'Thin Red Line' that traces the true future branches. (See Belnap and Green 1994, although they reject this view.) Counterfacts are the analogues of the Thin Red Line for alternative possibilities. Let me add on behalf of counterfactism that 'would' is the

[16] Stefánsson coined the word "counterfacts".

past tense of 'will'. So it is natural to think of counterfactuals, which we express with 'woulds', as typically making predictions relative to non-actual starting points.

Having granted that the analogy between counterfactuals and predictions is suggestive, I now want to undercut it. For starters, most predictions have their moment of reckoning.[17] As I am about to toss a coin, I predict "It will land heads". Then, we simply wait and see whether the prediction comes out true. And there is no mystery about its truth-maker: it is a perfectly mundane, Humean supervenient fact (in the sense of Lewis 1986). But a counterfactual about a coin toss that never takes place has no such moment of reckoning. The truth-maker of 'if the coin *had been* tossed, it *would have* landed heads' is altogether more mysterious. According to counterfactism, it is a primitive modal fact—not mundane at all, not Humean supervenient, not determined by the non-modal facts. Indeed, it contradicts the dictum that *truth supervenes on being*[18]—truth supervenes on what objects exist and what properties they have. I don't know how even to begin to describe what such a putative counterfact would look like. All the more, I question consequentialism's putative naturalistic credentials.

Moreover, counterfactism is committed to a spectacular proliferation of such primitive modal facts. For each antecedent *A*, there is a corresponding 'thin red line': an entire future history that would have been realised if *A* had been the case. That's a *lot* of thin red lines—infinitely many! Moreover, presumably the counterfacts could have been otherwise. For each true counterfactual about how they could have been, there is a further counterfact. And presumably each such further counterfact could have been otherwise, and so on—an infinite regress of primitive modal facts. The ontological commitment of counterfactism goes way beyond that of future contingents having truth values. In any case, counterfactism has the problem (as I see it) of implausibly specific counterfactuals in spades. Even a counterfactual with the maximal possible mismatch in specificity between antecedent and consequent is true by counterfactist lights: 'if something had been different from how it actually is, the world would have been *w*', for a unique *w*. I think that's false; but even someone who thinks that it is indeterminate, or that it is a truth gap, agrees with me that it is *not true*.

[17] Most, but not all. Consider Dummett's example: "A city will never be built here". Still, there is nothing mysterious about its truth-maker—it is a perfectly mundane, Humean supervenient fact, in keeping with what I say in the rest of this paragraph about predictions that do have their moment of reckoning.
[18] Thanks to David Builes for this way of putting the point.

Context-dependence and objective consequentialism

But it arguably gets worse. Almost all philosophers think that counterfactuals are *context-dependent*: the proposition expressed by a counterfactual can change, depending on the context in which it is uttered—what is salient, what serves one's conversational purposes, the operative standards of precision, the stakes, whether back-tracking resolutions of the similarity relation are appropriate, and so on. (See e.g. Karen Lewis 2014, 2016). (This is another striking disanalogy between future contingents and counterfactuals: nobody thinks that the truth of 'wills' depends on what is salient, and so on.) If counterfactuals are context-dependent, counterfactism is even more profligate. Now we have a thin red line for each antecedent *and* for each context in which a counterfactual could be uttered—each with its own primitive truth-maker!

The problem of context-dependence for counterfactism quickly turns into a problem for any version of objective consequentialism formulated counterfactually. This is the case whether or not it is committed to maximally-specific counterfactuals that counterfactism embraces. Suppose that counterfactuals about what the consequences would be if an action were performed are context-dependent. Then it seems that objective consequentialism formulated counterfactually is committed to the context-dependence of corresponding moral evaluations: whether an action is right, or whether it is better than another. Moreover, it is committed to their context-dependence *in the same ways that counterfactuals are context-dependent*: depending on what is salient, what serves one's conversational purposes, the operative standards of precision, the stakes, whether back-tracking resolutions of the similarity relation are appropriate, and so on. I question whether moral evaluations are context-dependent at all, but in any case I seriously doubt that they are context-dependent *in the same ways*.

Generalising the problem

The heart of the problem is objective consequentialism's appeal to facts regarding the long-term consequences of actions that are not performed. According to this theory, they are all that matters: the moral status of an action is entirely determined by such consequences. But a version of the problem will arise for any theory for which such consequences count for *something*. As Greaves writes: "any plausible moral theory will agree that considerations of consequence-goodness are at least morally relevant—that they should be taken serious account of, both in moral decision-making and in moral evaluation, as at least one important factor" (312).

Even most deontological theories give *some* moral weight to the consequences of actions. For example, a deontologist may acknowledge that while one has a pro tanto duty to keep a promise, this may be overridden if the consequences of doing so are sufficiently dire. Rawls (1971) writes:

> deontological theories are defined as non-teleological ones, not as views that characterise the rightness of institutions and acts independently from their consequences. All ethical doctrines worth our attention take consequences into account in judging rightness. One which did not would simply be irrational, crazy. (30)

And Hursthouse (1999, 33) writes: "Though it is sometimes said that deontologists 'take no account of consequences', this is manifestly false, for many actions we deliberate about only fall under rules or principles when we bring in their predicted consequences." Now, perhaps a rabid deontologist would agree with Kant that one's duties can never be overridden—for example, that one must tell the truth about one's friend's location even when a murderer asks for it. But surely a more sensible deontologist will allow that at least *some* consideration must be given to consequences. And to the extent that these are long-term consequences, the problems that I have raised for objective consequentialism will kick in. It makes no sense to speak of the consequences of an action that is never performed. And it is a jaw-dropping metaphysical commitment to appeal to counterfactuals about what the long-term consequences *would be* if the action were performed. To the extent that deontological theories make such an appeal, their foundations are also suspect.

Likewise, virtue ethics must traffic in the consequences of actions to some extent. Hursthouse continues: "A surgeon who subscribes to virtue ethics has the same problem: she may not doubt that charity, which is concerned with others' good, is a virtue; her doubt is over whether the consequences of the operation will be that her patient is benefited or harmed." Similarly, Hursthouse and Pettigrove (2018) write: "It should go without saying that the virtuous are mindful of the consequences of possible actions. How could they fail to be reckless, thoughtless and short-sighted if they were not?" But again, an action that is *merely* possible does not have consequences at all. And long-term (for those who are not "short-sighted") counterfactual consequences are as suspect here as they were before.

The problems of cluelessness and clumsiness will also recur for any moral theory to the extent that it acknowledges the significance of long-term consequences and attempts to characterise it with counterfactuals. Again, there is acute sensitivity of the consequences to the exact way an action is realised. To the extent that the consequences matter, this acute sensitivity will matter. Yet you are clueless about them. And you are clumsy: you have limited control

over your actions, and you cannot as an act of the will realise them one particular way as opposed to some closely neighbouring way. As before, a dilemma arises for these alternative theories. If the objects of moral evaluation are rather coarse-grained options, then it is implausible that there are facts of the matter about how they would counterfactually be realised; if they are very fine-grained 'options', then it is implausible that you can decide to realise one rather another, and as such it is implausible that they really are *options* for you at all. Moreover, as before, there is no 'sweet spot' between this dilemma's horns.

And consider again the best case for there being a fact of the matter of what would happen were you to act in some (non-actual) way: determinism. Again, we may specify as precisely as we like the details of your hypothetical action, but what would then happen depends on much more than you. For the rest of history to follow, a snapshot of the entire world at that time is needed, but all we have is a selfie.

I don't want to overstate this. To be sure, deontology and virtue ethics are not as beholden as objective consequentialism is to exactly what the consequences would be. But given how perilously the consequences may vary depending on the fine details of what you do and what the rest of the world does, even various versions of these ethical theories risk foundering.

Where do we go from here?

So far, my discussion has been almost entirely critical. How did we get into this predicament? Let's return to objective consequentialism, as I have been understanding it. Three moving parts led to the problems that I have raised for it:

1. It appeals to the *highly specific, long-term* consequences of actions;
2. It understands consequences with *counterfactuals*;
3. It is *objective*.

Accordingly, I want to briefly consider three ways we might rescue it:

1) Appeal instead to *not-too-specific, short-term* consequences of actions;
2) Understand consequences with *objective probabilities*;
3) Understand consequences with *subjective/evidential probabilities*.

I think the first way still fails, but that the next two are more promising.

<u>1) Not-too-specific short-term consequences of actions</u>

I found it highly implausible that there is a fact of the matter of highly-specific counterfactual consequences, until the end of time, that *would* have transpired if you had performed some action. But it is much more plausible that there is a fact of the matter of the *not-too-specific short-term consequences* of your action.

I am deliberately staying vague about both parts of this proposal, but each is meant to earn its keep—let me illustrate how. If you were to help the old lady across the street, she would be significantly happier for the next few minutes. If you were to go the pub, you would have fun for the next hour or two. And so on. Such claims are straightforward common sense.[19] But they cease to be if we either increase the specificity too much or extend the time horizon too much. If you were to go to the pub, exactly *this* sequence of neuron-firings would occur in your brain for the next hour or two?—Too specific. If you were to go to the pub, you would have fun for the next hour or two, and you would eventually go on to have two children, and then thirty or so years later have a total of five grandchildren, and then thirty or so years later have a total of eleven great-grandchildren, and then … (and so on, until the end of time)?—Too extended a time horizon. The trick is to stay within common-sensical limits on how specific the counterfactual consequences are, and for how long.

The serious metaphysical problem that I raised for long-term objective consequentialism seems to be alleviated. There is no longer a striking mismatch between the strength of the antecedent and that of the consequent: we have now greatly weakened the consequent. While we're at it, apparently the problems of cluelessness and clumsiness are alleviated too. Common sense has it that we often know such simple counterfactuals. Moreover, it seems that the fine details of how you act don't matter much for the not-too-specific, short-term consequences. You could help the old lady across the street in a myriad of ways, and still bring about her being significantly happier for the next few minutes. You could arrive at the pub at a wide range of times, and still have fun for the next hour or two. Clumsy you might be, but it shouldn't matter—the unspecific, proximate consequences should turn out much the same. As long as the counterfactuals are suitably circumscribed—which is what the not-too-specific short-termist wants—all is well. Or so it seems.

For what it's worth, I think that even such not-too-specific short-term counterfactuals are false, but that's because of my particular (some might say peculiar) views about counterfactuals, which I have not presupposed in the bulk of this talk. I think that *most*

19 Lenman (2000) advocates short-termism: "insofar as the agent's concern is with consequences at all, it is with visible consequences that he or she should be, even indirectly, concerned." (364)

counterfactuals are false, because of either chanciness of their consequents or unspecificity of their antecedents. If you were to help the old lady, she *might not* be significantly happier for the next few minutes: there would be some chance of your offending her, or of her falling badly, or of her having a heart attack, or what have you. If you were to go the pub (somehow or other), you *might not* have fun for the next hour or two: you might get into a fight with someone, or a friend might break some bad news to you, or what have you. So I still think that the requisite counterfactuals for objective consequentialism are false even on its commonsensical not-too-specific, short-termist formulation. However, this time I acknowledge that *I* am the one with the surprising metaphysical commitments—my view flies in the face of common sense—whereas previously it was clearer that the long-term objective consequentialist had them.

Be that as it may, the respite for objective consequentialism is short-lived. Recall the psychopath contemplating hooking up a doomsday device to be activated if a coin landed heads, but not if the coin landed tails. The potential consequences are short-term, and may be rather unspecific: the world would be destroyed (somehow or other) immediately if the coin landed heads; business as usual (more or less) if the coin landed tails. So, what *would* happen if the psychopath did his deed? It's not the case that billions of people *would die*—they might, they might not, each scenario happening with chance 1/2. Even short-term, not-too-specific consequentialism fails to capture the wrongness of the deed. This version of consequentialism might fare better than a long-term-specific version in more mundane cases, such as helping an old lady or going to the pub, but it is still untenable.

2) Understand consequences with *objective probabilities*

I think that *probabilities* are in better order than counterfactuals. Consequentialism should be formulated in probabilistic terms, following Frank Jackson (1991) and Greaves (2016). In particular, the notion of *expected value* (*expectation*) is probabilistic: it is a weighted average of possible values, the weights being conditional probabilities. The value of a given possibility is multiplied by its conditional probability of being realised, given an action. The expected value is the sum of such products. In your choice between helping the old lady and going to the pub, the right action is the one that has higher expected value. We get rid of the problematic counterfactuals altogether.

Of course, this shifts the problem of understanding consequences counterfactually to the problem of understanding probabilities. But we have a vast literature on the interpretation of probability to draw on. In particular, we may distinguish *objective probabilities*, also known

as *chances*; *subjective probabilities*, also known as *credences*; and *evidential probabilities*, also known as *degrees of confirmation*. (See Hájek 2023.) I suggest, then, that objective consequentialism should be formulated in terms of *objective probabilities*. A morally right action is one that maximises expected value, where the weights are conditional chances (at the time of choice). More generally, given a choice between two actions, one is morally better than the other just in case the former's expected value is greater than the latter's.

Will the problems that I have raised for objective consequentialism formulated with counterfactuals reappear when it is stated in terms of objective chances/expectations? No, or at least not to the same extent. There is so much to say about this, so little time! But briefly:

What about non-actual actions? I observed that there are no consequences of an action that is not performed. But there are *chances* of consequences conditional on such an action—*actual* chances, features of the actual world. For example, there are chances at the time of your choice of various consequences given that you go to the pub, even if you do not in fact go to the pub. Indeed, most conditional chances have conditions that are not realised.

What about the metaphysics? I argued that counterfactual-based consequentialism has jaw-dropping metaphysical commitments, on pain of rampant indeterminacy. Chance-based consequentialism, I submit, is in better order. We have a vast literature on the interpretation of *objective* probability to draw on: frequency interpretations (actual and hypothetical), propensity interpretations (frequency-based and non-frequency-based), best systems interpretations (Lewisian and Mentaculus), symmetry-based accounts (the method of arbitrary functions), and more. I won't choose among them here—I have expressed some of my (dis)preferences elsewhere.[20] (For example, I am no fan of frequentism!) Here I merely point out that there is much to say philosophically in favour of a commitment to chances. We are plausibly committed scientifically to chances anyway. Statistical mechanics and quantum mechanics are explicitly probabilistic theories, and it is natural to understand their probabilities objectively. Far from being suspicious of chances, I think that they are indispensable. In any case, I need only a commitment to *comparative* chances—one chance being greater than another—and that's less of a commitment than to numerical chances.

What about rampant indeterminacy? I argued that on the view that counterfactuals with chancy consequents or unspecific antecedents are indeterminate, all too many consequentialist counterfactuals will be indeterminate—there will be no verdicts on which action is better than

[20] References to relevant work by Venn, von Mises, Popper, Giere, Lewis, Loewer, Strevens, and my anti-frequentism papers can all be found in Hájek 2023.

another, even when the answer is obvious (e.g. helping the old lady versus serial-killing). Likewise on the view that these counterfactuals have no truth values. One might level a similar charge against my chance-based account: that for many of these cases there are no *chance* values, and hence no verdicts.[21] I assume that chances are abundant—they attach to all propositions of interest to us here. And I am in good company. I take Lewis (1980, 1986) to have a similarly expansive view of chances. Loewer (2020) is even more explicit about this in his Mentaculus account of chances—indeed, he regards it as a selling point of his account that it applies so broadly (and so do I). Soon we will see the influential idea that rational credences are expectations of corresponding chances. But if a given chance is undefined, so too is the corresponding expectation: a weighted average of values is undefined if one or more of the values themselves are undefined. Yet credences are a dime a dozen—they can attach to pretty much any proposition that one can comprehend. So it seems that this idea is committed to chances being a dime a dozen too, or at least to rational credences being committed to their being a dime a dozen too.

Moreover, I think that conditional chances are even more abundant than unconditional chances, since the former can be well defined even when the latter are not. For example, we might question whether free actions such as your going to the pub have chances. But various chances *conditional* on your pubbing are well defined. Here are two simple examples: the chance of your going to the pub *given* that you go to the pub (= 1), and the chance of a coin toss landing heads *given* that you go to the pub (= ½). (See Hájek (2003) for more discussion.) So all the more the requisite conditional chances are out there, wherever the objective consequentialist needs them to be. Still more the requisite *comparative* conditional chances are out there—they may be defined while numerical conditional chances are not.

What about clumsiness? While *outcomes* are often acutely sensitive to initial conditions, the *chances of outcomes* are more robust. The outcome of a given coin toss may depend sensitively on exactly how it is tossed, while the chances of the outcomes are resiliently ½. The chance of heads does not depend on who is tossing it, or how high they toss it, or when they toss it, and so on (within reasonable limits). Similarly, the outcome of a given spin of a wheel of fortune landing in a red rather than a black sector may depend sensitively on the initial force that is imparted to it, but the chance of red is comparatively invariant. (See Strevens 2003, 2013 for detailed discussions of both of these cases.) Indeed, often chances are so invariant that they get enshrined in the laws of nature while the outcomes that they attach to are seemingly random—

[21] Mikayla Kelley and Daniel Nolan have raised this concern to me.

think of the probabilistic decay laws of radioactive particles, Born's rule in quantum mechanics, and Mendel's law of independent assortment. Chances are such powerful theoretical tools largely because of their stability when compared with the outcomes themselves. Conditional chances are more robust. Inequalities between conditional chances, and correspondingly between expectations, are far more robust still. After all, a chance inequality or an expectation inequality can persist through huge changes in the values themselves, and even through huge changes in the value-differences

I take these points to apply to the cases we have considered (while admitting that it is a little harder to substantiate these points). Small variations in your exact arrival time at the pub will make no difference, or almost no difference, to the conditional chances of latter-day Buddhas or Hitlers being eventually conceived; still less to the comparative conditional chances, the inequalities between them, and thus the comparative expectations. So your clumsiness about your exact arrival time will not matter to these chances, and hence to the corresponding expectations. It's as if God would tell you: "Don't worry, you can relax when it comes to the comparative expectations—it all comes out in the wash!". And so it goes for almost all your choices.

What about cluelessness? Thanks to the stability of conditional chances, the relevant facts about them are more easily known than the corresponding outcomes. To be sure, we will typically have no idea what the *values* of the requisite conditional chances are. (Coin-tossing, wheel-of-fortune-spinning, and probabilistic laws are special cases, though they display our good epistemic standing to at least some chance values.) But we need not know the values themselves in order to know whether one action is better than another—that's a matter merely of whether the expectation of the former is *greater than* that of the latter, whether this *inequality* holds between them. And we will often have clues about that.

Expectation is a measure of the centre of location of a distribution. We can know that the centre is *shifted* one way or another by an action without knowing by how much, let alone where the centre was or is. We know that delaying your departure time for a city drive until after rush hour will decrease the expected duration of your trip. We know that increasing your salt intake increases the expectation of your blood pressure.

Or consider a *quincunx* (the curiously-named gift to Scrabble players), or *Galton board*: a vertical board with interleaved rows of pegs. Many balls are dropped from above, and they bounce their way to the bottom, where they are collected in small bins. The trajectory of any particular ball is utterly unpredictable—it gets buffeted left or right seemingly at random on its way down as it strikes the pegs on successive rows. At the level of what an individual ball does,

we are clueless! Yet collectively the balls almost invariably form a bell-shaped pattern across the bins, corresponding to the binomial chances across these resting points. (YouTube has many videos of this delightful phenomenon.) Suppose you release the balls over one point of the board; the bell shape of balls that eventuates will be centred directly under this initial release point. Now move the release point to the left; the bell will then be shifted to the left. Far from being clueless about this, one can be almost certain of it.

I think that something similar is often true of the chance distributions over the consequences of one's actions. Consider the chance distribution over consequences of, say, donating to Oxfam, and compare that to the chance distribution over consequences of serial-killing. One can be almost certain that the latter distribution is 'shifted to the left' of the former—the distribution of consequences is centred at a lower value for the latter than the former—despite one's being clueless about exactly what the future trajectory of one's actual action will be. And this means that one is not clueless after all about which action would be better—on the contrary, one can be almost certain that the former is better. (Cf. Shiller 2021.)

What about clumsiness? While *outcomes* are often acutely sensitive to initial conditions, the *chances of outcomes* are more robust, and the *conditional chances* are even more robust. Much more robust still are *comparative* conditional chances, and correspondingly *comparative* conditional expectations. Small variations in your exact arrival time at the pub will make no difference, or almost no difference, to the conditional chances of latter-day Buddhas or Hitlers being eventually conceived. So your clumsiness about your exact arrival time will not matter to these chances, and hence to the corresponding expectations. It's as if God would tell you: "Don't worry, you can relax when it comes to the comparative conditional chances—it all comes out in the wash!". And so it goes for almost all your choices.

The solutions to the problems of clumsiness and cluelessness are closely related. Both turn on the comparative *insensitivity* of chances to initial conditions, which we are typically unable to precisely fine-tune and of which we are ignorant, and the much greater insensitivity of comparative conditional chances.

In any case, I think that the prospects for solving the problem of cluelessness may be better still for subjective consequentialism—to which I now turn.

3) Understand consequences with *subjective/evidential probabilities*

Now the obvious move is to replace the objective probabilities with *subjective* probabilities—those of the agent in question. This brings us to Jackson's *decision-theoretic consequentialism*. The morally right action for this agent is the one that maximises subjective

expected value, where the weights are her conditional credences (at the time of the choice). And more generally, in Greaves' words (2016, 316): "Act A1 is subjectively c-better than A2 iff the expected value of the consequences of A1 is higher than the expected value of the consequences of A2 (where both expectation values are taken with respect to the agent's credences at the time of decision)".

Again, the problems that I have raised for objective consequentialism formulated counterfactually are mitigated, or even entirely solved. Again, non-actual actions are not a problem. Credences of consequences conditional on such an action can easily be well-defined. Indeed, most conditional credences have conditions that are not realised.

The metaphysics of credences is arguably in even better order than that of chances. While some authors are skeptical about the existence of degrees of belief (Harman, Byrne, Holton, Horgan, …), they are in the minority. Credences are such a staple nowadays of formal epistemology and decision theory, not to mention economics, psychology, and computer science, that I feel I can appeal to them with impunity; and if I can't, I'm in excellent company. (See Eriksson and Hájek 2007 for more defence of credences.)

It is even clearer than it was for objective expectations that clumsiness is not a problem for subjective expectations. For almost any rational agent, small variations (and even large variations) in their exact arrival time at the pub will make no difference to their conditional credences of most subsequent events—for example, latter-day Buddhas or Hitlers being eventually conceived. It would take extraordinary evidence for a rational agent's credences in these possibilities, and all the more for inequalities among expectations, to be so acutely sensitive to the initial conditions. (Perhaps God could provide such evidence, but that certainly would be extraordinary!) And so it goes for almost all your choices.

I have left the problems of rampant indeterminacy and cluelessness for last. After the quote just given, Greaves writes: "We can never have even the faintest idea, for any given pair of acts (A1, A2), whether or not A1 is subjectively c-better than A2." This means that you, for example, can never have the faintest idea whether *the expected value* of the consequences of A1 is greater than that of A2. How could that be? Expected value is a very simple formula: a sum of products of credences and values. How could you not have the faintest idea whether one such sum of products is greater than another? Either you don't know what some of your credences are, or you don't know what some of the values are. (I set aside the possibility that you don't know how to do the elementary arithmetic if the sum is finite, which it plausibly is for agents like us.). The problem of cluelessness was originally supposed to be a problem about *unpredictability of events in the world*. But now it seems to have morphed into something

else—it's no longer cluelessness about what happens in the world, but rather what is happening in your head! Greaves' 'cluelessness' problem is orthogonal to Lenman's.

Indeed, 'cluelessness' does not seem like the right word for Greaves' problem, while 'indeterminacy' seems more apposite. Assuming that we have introspective access to our own credences (and values)[22], we are not clueless about the requisite expectations when they exist—we can just calculate them by plugging in the appropriate credences (and values). The problem is more that our credences *may not exist*, or they may exist but be so *imprecise* as not to yield a determinate ordering among our options. For example, our credences for an additional 22nd century dictator conditional on helping the old lady or going to the pub might be imprecise over wide intervals, yielding correspondingly wide expectation-intervals for each of these options, which may overlap. Then we might say that it is indeterminate what is the right thing to do—neither is determinately better than the other, nor are they determinately equally good.[23] Moreover, we may well know that this the case. Far from being clueless, we may be all too aware of our predicament! But it may not be such a predicament after all. Often it *is* indeterminate what the right thing to do is, by one's own lights.[24] Indeed, too much determinacy might be problematic in its own way.

'Cluelessness' suggests that there is a fact of the matter regarding what the right thing to do is, but you don't know what it is. But if it is indeterminate what it is, there is no fact of the matter. There is nothing to know. Not even God could tell you what is the right thing to do. Compare: God can't tell you that Sherlock Holmes had an even number of hairs when he met Watson, nor that he had an odd number of hairs, and it would be odd to say that you are 'clueless' about which it is.

So far we have been following Greaves and Jackson in assuming that the credences in question are those of the agent at the time of decision. It is natural to understand this literally as the *human* agent, warts and all, who may be less than ideally rational in various ways—forgetful, subject to biases, computationally limited, poor at assessing their evidence, and so on. Indeed, plausibly their credences do not even obey probability theory. Perhaps instead we

---

[22] If we lack introspective access to our own credences, a different problem of 'cluelessness' is: what is the right thing to do *by our own lights* when we clueless about what those lights are?! But even if we lack *perfect* introspective access, we surely have *some* access. It seems to be an overstatement to say that we are *clueless* in this sense.

Ignorance of the objective *value* of various consequences is another problem of 'cluelessness' for objective consequentialism, one that has been discussed rather less than the problem that Lenman raised.

[23] There are various approaches to decision theory with imprecise credences. This is determinately not the place to get into the weeds, but I hope I have said enough to indicate a possible concern.

[24] Not of course when it comes to easy cases, such as whether to donate to Oxfam or to go on a serial-killing rampage.

should be thinking of an *ideally rational* agent. In that case, their credences do obey probability theory (I will assume). However, according to radical subjectivism there are no further constraints—anything goes, as long as they obey probability theory. This is far too unconstrained to serve as a basis for a *moral* theory worthy of the name. After all, not anything goes when it comes to morality.

It is more plausible, then, that there are further constraints on rational credences. Greaves appeals to the Principle of Indifference, which we may regard as providing *evidential probabilities* or degrees of confirmation. I am skeptical of her appeal, even in the "simple" cases of cluelessness to which she argues that it applies (see Hájek MS). But I take seriously the idea that ideally rational credences align with evidential probabilities. We might aspire to be like such agents, but we fall short. Now we could say that another 'cluelessness' problem arises: we don't know what the ideally rational credences are, given our evidence. But again, this is not the original cluelessness problem, a problem about *unpredictability of events in the world*. It is an important problem nonetheless, both for epistemology and ethics. After all, various epistemic virtues have been thought to be truth-conducive—e.g., simplicity, explanatory power, and fertility—and to the extent that they are, they should be reflected in evidential probabilities, although these are difficult to quantify, and thus difficult to turn into probabilistic constraints.

The Principal Principle

Given the centrality of objective chances in consequentialism as I have formulated it, it is natural to seek a constraint that is based on chances. Here it is—the so-called *Principal Principle* (Lewis 1980, here simplified):

$C(A \mid ch(A) = x) = x$, for all A and for all x where this is defined.

Here, 'C' is the credence function of a rational agent, and '*ch*' is the objective chance function. The idea is that a rational agent strives to track the chances with her credences. When she knows what they are, her credences align exactly with them. When she is unsure what they are, she uses her credences over various hypotheses of what they are, and takes an expectation of them.

Now that we have a bridge between chances and rational credences, I can recapitulate some points that I made earlier about objective expectations, and parlay them into corresponding points about subjective expectations. I argued that chances of outcomes are less sensitive to initial conditions than the outcomes themselves are. Knowing this fact about chances, rational credences, and thus rational subjective expectations, will correspondingly be less sensitive to

initial conditions than the outcomes themselves. Even less sensitive will be conditional credences, and much less sensitive will be comparative conditional credences, and thus inequalities between conditional-credence-weighted expectations. We thus bolster our solution to the problem of clumsiness. I also argued that we can often be almost certain about inequalities between objective expectations—for example, almost certain that the expectation is greater for donating to Oxfam than it is for setting up a probabilistic doomsday device. Rationality then requires us to reflect this inequality in our subjective expectations. Subjective consequentialism then bids us to donate. Moreover, far from being clueless, we can know this.

That said, I admit that lots of cases are harder. But we already knew that: sometimes our credences, and plausibly even ideally rational credences, do not yield a determinate verdict about which of our options is best. Subjective consequentialism should recognise that.

To be sure, the Principal Principle is just one evidential probability constraint. The program of evidential probability has roots in the work of Pascal and Laplace, with notable contributions by Keynes, Carnap, and Williamson. But it is still at a rudimentary stage of development in epistemology—and so it is in moral philosophy, to the extent that the latter needs to be informed by the former.

Conclusion: consequentialism reformulated

It is time to put all this together. Objective consequentialism should not speak of "the consequences" of actions that are not performed, let alone comparisons involving them, since this is nonsense. Nor should it traffic in counterfactuals about what the consequences of such actions "would be", since this betrays a commitment to dubious metaphysics.

Rather, objective consequentialism should be formulated probabilistically. The objective moral value of an action is the conditional-chance-weighted average of the value of its possible consequences. An objectively right action is one that maximises this quantity. And one action is better than another just in case the former's objective moral value is greater than the latter's. Replace the chances by credences (either actual or idealised) and we get subjective consequentialism. The distinction between the two kinds of consequentialism neatly follows from the corresponding distinction between two kinds of probabilities.

This brings me back full circle to where I began. I said that foundational issues in probability and counterfactuals bear crucially on moral philosophy. We have seen that the very formulation of consequentialism, and of various other moral theories, bring in their train a host of such issues.

But I have barely begun …

REFERENCES

Belnap, Nuel and Mitchell Green (1994). Indeterminism and the thin red line. *Philosophical Perspectives*, Vol. 8, Logic and Language. Ridgeview Publishing Company, 365-388.

Bradley, Richard (2012). Multidimensional possible-world semantics for conditionals. *Philosophical Review*, 121(4), 539-571.

Bradley, Richard (2017). *Decision theory with a human face*. Cambridge: Cambridge University Press.

Burch-Brown, Joanna M. (2014): "Clues for Consequentialists", *Utilitas* 26 (1), 105-119.

Carlson, E. 1995. *Consequentialism Reconsidered*. Dordrecht: Kluwer.

Chen, Eddy Keming and Sheldon Goldstein (2022): "Governing Without A Fundamental Direction of Time: Minimal Primitivism about Laws of Nature", in Ben-Menahem, Y. (ed.) 2022, *Rethinking the Concept of Law of Nature*, Jerusalem Studies in Philosophy and History of Science. Springer, Cham, 21-64.

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

Dorr, Cian (2016): "Against Counterfactual Miracles", *Philosophical Review* 125, 241–86.

Driver, Julia (2012): *Consequentialism*, Routledge.

Eriksson, Lina and Alan Hájek (2007): "What Are Degrees of Belief?", *Studia Logica* 86, July, (Formal Epistemology I), 185-215, ed. Branden Fitelson.

Greaves, Hilary (2016): "Cluelessness", *Proceedings of the Aristotelian Society* 116, No. 3, 311-339.

Gustafsson, Johan E. (2019): Is Objective Act Consequentialism Satisfiable? *Analysis* 79 (2), 193-202.

Hájek, Alan (2003): "What Conditional Probability Could Not Be", *Synthese*.

Hájek, Alan (2023): "Interpretations of Probability", *The Stanford Encyclopedia of Philosophy* (Winter 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/win2023/entries/probability-interpret/>.

Hájek, Alan (MS): "Cluelessness and Indifference".

Hare, Caspar (2011): “Obligation and Regret When There is No Fact of the Matter About What Would Have Happened if You had not Done What You Did”, *Noûs* 45 (1), 190-206.

Hawthorne, John (2005): “Chance and Counterfactuals”, *Philosophy and Phenomenological Research* 70 (2), 396-405.

Hursthouse, Rosalind (1999): *On Virtue Ethics*, Oxford: Oxford University Press.

Hursthouse, Rosalind and Glen Pettigrove, "Virtue Ethics", *The Stanford Encyclopedia of Philosophy* (Winter 2018 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2018/entries/ethics-virtue/>.

Jackson, Frank (1991): “Decision-theoretic Consequentialism and the Nearest and Dearest Objection”, *Ethics* 101 (3), 461-482.

Lewis, David (1973): *Counterfactuals*. Blackwell and Harvard University Press.

Loewer, Barry (2020): “The Mentaculus Vision”, in V. Allori (ed.) *Statistical Mechanics and Scientific Explanation: Determinism, Indeterminism, and Laws of Nature*, Singapore: World Scientific, 3–29.

Mogensen, Andreas L. (2021): “Maximal Cluelessness”, *The Philosophical Quarterly* 71 (1), 141-162.

Molina, Luis de (1953). *Liberi arbitrii cum gratiae donis, divina praescientia, providentia, praedestinatione et reprobatione Concordia*, John Rabeneck (Ed.), Oña and Madrid.

Moore, G.E. 1903. *Principia Ethica*. Cambridge: Cambridge University Press.

Moss, Sarah (2013). Subjunctive credences and semantic humility. *Philosophy and Phenomenological Research*, 87(2), 251-78.

Portmore, D.W. (2011). *Commonsense Consequentialism: Wherein Morality Meets Rationality*. New York: Oxford University Press.

Railton, Peter (1984): “Alienation, Consequentialism, and the Demands of Morality”, *Philosophy & Public Affairs*, Vol. 13, No. 2, 134-171.

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

Savage, Leonard J. (1954): *The Foundations of Statistics*, New York: John Wiley and Sons.

Savulescu, Julian and Dominic Wilkinson (2019): “Consequentialism and the Law in Medicine”, in *Philosophical Foundations of Medical Law*, T.C. de Campos., J. Herring, and A.M. Phillips (eds.), Oxford: Oxford University Press. Available at:

https://www.ncbi.nlm.nih.gov/books/NBK550266/#:~:text=According%20to%20consequentialism%2C%20the%20right,some%20ethical%20relevance%20of%20consequences).

Miriam Schoenfield (2012): “Chilling out on epistemic rationality: A defense of imprecise credences”, Philosophical Studies 158 (2):197-219.

Schulz, Moritz (2017). *Counterfactuals and probability*. Oxford: Oxford University Press.

Shiller, Derek (2021), “Chance and the Dissipation of our Acts’ Effects”, *Australasian Journal of Philosophy* 99(2), 334-348.

Sidgewick, Henry (1874): *The Methods of Ethics.*

Sosa, David (1993). Consequences of consequentialism. Mind 102: 101–22.

Stalnaker, Robert C. (1968). A theory of conditionals. *Studies in Logical Theory, American Philosophical Quarterly*, Monograph: 2, 98-112.

Stalnaker, Robert C. (1980). A defense of conditional excluded middle. In *Ifs*, Harper W.L., Stalnaker R. and Pearce G. (Eds.), Dordrecht: Springer, 87-104.

Stefánsson, H. Orri (2018). Counterfactual skepticism and multidimensional semantics. *Erkenntnis*, 83, 875-898.

Stevens, Michael (2003): *Bigger than Chaos: Understanding Complexity through Probability*, Harvard University Press, Cambridge, MA.

Stevens, Michael (2013): *Tychomancy: Inferring Probability from Causal Structure*, Harvard University Press, Cambridge, MA.

Suarez, Francisco (1856-1878), De gratia. In *Opera Omnia*. Paris.

Tännsjö, Torbjörn (2013): *Understanding Ethics*, Edinburgh University Press.

Vessel, Jean-Paul (2003): “Counterfactuals for Consequentialists”, *Philosophical Studies* 112 (2), 103-125.