ACT CONSEQUENTIALISM WITHOUT FREE RIDES

FORTHCOMING IN *PHILOSOPHICAL PERSPECTIVES*

ABSTRACT. Consequentialist theories determine rightness solely based on real or expected consequences. Although such theories are popular, they often have difficulty with generalizing intuitions, which demand concern for questions like "What if everybody did that?" Rule consequentialism attempts to incorporate these intuitions by shifting the locus of evaluation from the consequences of acts to those of rules. However, detailed rule-consequentialist theories seem ad hoc or arbitrary compared to act consequentialist ones. We claim that generalizing can be better incorporated into consequentialism by keeping the locus of evaluation on acts but adjusting the decision theory behind act selection. Specifically, we should adjust which types of dependencies the theory takes to be decision-relevant. Using this strategy, we formulate a new theory, *generalized act consequentialism*, which we argue is more compelling than rule consequentialism both in modeling the actual reasoning of generalizers and in delivering correct verdicts.

0. INTRODUCTION

One of the great difficulties of modern moral theory is the synthesis of two influential, yet often adversarial, ideas. Consequentialism, on the one hand, requires an exclusive focus on consequences in determining moral wrongness. On the other, generalizability (or, "universalizability," as it is sometimes called), in its pre-theoretic form, requires a concern for the question, "What if everyone did that?" These two ideas, though not incompatible, have nevertheless proved difficult to combine. In contrast, there is no shortage of consequentialist theories that ignore generalizing. For example, a simple and popular form of consequentialism, act consequentialism, in most contexts has no concern for the generalization question whatsoever: according to the standard act consequentialism, generalizing one's act is not relevant to evaluating its expected consequences.

The desire for a consequentialist theory that generalizes explains much of the allure of rule consequentialism.[1] A basic idea of rule consequentialism is that we can incorporate generalizing into consequentialism by developing principles for the evaluation of sets of rules. However, it has proven difficult to spell out rule-consequentialist theory in detail. When this has been attempted, as with modern "acceptance-based" rule consequentialisms like that of Brad Hooker (2000), the result is theories that struggle to avoid arbitrariness in the presentation of essential principles. Nearly all theorists accept that, at least in comparison to act consequentialism, fully specified rule-consequentialist theories are on shakier footing when it comes to worries about arbitrariness.

Our proposed diagnosis is the following: the conceptual apparatus that rule consequentialism employs to generalize consequentialism is not sufficient for the task. Just as advances in decision theory, such as expected-utility theory, were required to formulate consequentialist theories that succeed in dealing with *uncertainty*, advances in decision theory may

[1] A similar motivation is behind generalizing alternatives to motive and virtue consequentialism. Examples include Adams' (1976, Sec. VI) "universalistic motive utilitarianism" (See (1976, Sec. VI)) and Bradley's (2005, Sec. IV) "universal virtue consequentialism." (See footnote 4 for more.)

also be required to formulate a consequentialist theory that succeeds in dealing with *generalization*. This essay explores this possibility.

Traditional expected-utility versions of act consequentialism are a marriage between a decision theory and an account of the good. Different decision theories claim different types of dependencies between act and state are decision-relevant. Evidential decision theory, for example, claims that all evidential dependencies are decision-relevant, while causal decision theory claims that only causal dependencies are decision-relevant. Since both types of dependency fail to account for generalizing intuitions, so does traditional act consequentialism.

Generalizing intuitions, such as that tacitly expressed by the "what if everyone did that" question, concern what we call *algorithmic* dependencies. Algorithmic dependencies are, roughly (for now), dependencies between the outputs of agents' decision procedures. Our theory, *generalized act consequentialism*, combines an account of the good with a decision theory that focuses on algorithmic dependence (instead of evidential or causal). We argue that generalized act consequentialism best elucidates the motivations and aspirations of consequentialist generalizers.

Section 1 develops the tension between standard consequentialist theories and generalizing intuitions in more detail and introduces the free-rider problem as the characteristic manifestion of the conflict. Section 2 discusses rule consequentialism and some of its problems. Section 3 recounts how advances in decision theory solved consequentialism's problem with uncertainty and argues that the same can be true of generalizability. Section 4 presents standard theories of act-state dependence, while section 5 presents our account of algorithmic dependence and uses it as an engine for generalized act consequentialism. Section 6 further contrasts generalized act consequentialism with rule consequentialism, and section 7 wraps up.

1. Consequentialism and Generalizability

A moral theory is a form of consequentialism if it holds that normative properties depend only on consequences. Consequentialist moral theories can evaluate the normative aspects of many things, such as acts, character traits, practices, and institutions. The most prominent consequentialist theory, act consequentialism, evaluates moral wrongness in terms of the goodness of the consequences of acts. Similarly, direct motive and virtue consequentialism evaluate moral wrongness in terms of the consequences of an individual's motives and virtues, respectively.[2]

The impulse to generalize presents a difficulty for most consequentialist theories. The characteristic manifestation of this difficulty is *the free-rider problem*. Here is one example:

> **Voting:** Molly favors Stephen over Leopold for president because she thinks Stephen will better increase aggregate welfare. However, Molly knows that it is very improbable that the election will be decided by exactly one vote. Because it is inconvenient for Molly to get to the polls, she is considering staying home, since voting doesn't seem like an effective allocation of her resources despite her moral preference for Stephen over Leopold.

From a standard act-consequentialist perspective, it might be true that Molly ought to stay home.[3] If her vote does not make a difference, then the consequences of staying home

[2]The indirect versions of motive and virtue consequentialism evaluate *acts* on the basis of the consequences of the motives or virtues that cause them.

[3]Both direct and indirect motive and virtue consequentialism can also face versions of the free-rider problem. To create a free-rider problem for motive consequentialism, we can stipulate that Molly's motive has no relevant

are superior to those of voting: Molly gets to use the time she would have spent voting doing something else, and the result of the election is unaffected. Nevertheless, if all, or even a substantial subset, of her fellow Stephen supporters decide in this way, then that could make a difference to the outcome of the election. Thus, the concern arises: many people find it natural to ask Molly to consider what would happen if others decided as she does.

The term "generalization" came to prominence in the influential work of Marcus Singer (1961), which helped set the agenda for the development of contemporary rule-consequentialist theory. Singer argued that the question "What if everyone did that?" is central to moral thought and discussion. The implications of the question, according to Singer, can be formalized into what he calls "the generalization argument":

> If everyone were to do that, the consequences would be disastrous (or undesirable); therefore, no one ought to do that. (Singer, 1955, 361)

Singer took this argument to be behind common exhortations against free-riding, such as "What would would happen if no one voted?" and "If everyone refused to serve, we would lose the war" (p. 361).

Singer's work generated an unusually large number of reactions and responses throughout the 1960s and 70s. At that point, early rule-consequentialist proposals were often referred to as "general utilitarianism" or "utilitarian generalization" (this was partly due to Singer, but also due to the terminology of early rule-consequentialist theorists like Harrod (1936) and Harrison (1953).)

Around the same time that Singer started writing about generalization, the term "universalizability" was introduced to ethical discussion by R.M. Hare (1955). Hare's interest was not in what Singer called "the generalization argument," but instead in a certain logical thesis about evaluative sentences. Hare's principle of universalizability built on Sidgwick's (1884, 379): "If a kind of conduct that is right (or wrong) for me is not right (or wrong) for someone else, it must be on the ground of some difference between the two cases, other than the fact that I and he are different persons." Thus, Hare's principle of universalizability was similar to what Singer called "the generalization principle" (to distinguish it from his generalization argument): "What is right for one person is right for any similar person in similar circumstances" (Singer, 1985, 47).

---

consequences other than her action of not voting. If the action has good consequences, then the individualistic versions of both direct and indirect motive consequentialism approve of Molly staying home. Similarly, the individualistic version of virtue consequentialism defines virtues by the consequences of the individual agent possessing a particular character trait. If the consequences are sufficiently positive, then the trait is identified as a virtue. Thus, a free-rider problem for this theory can be constructed in a similar way.

In an attempt to avoid the free-rider problem, motive and virtue consequentialists have suggested generalized formulations of their theories. Adams (1976, Section VI) introduces the idea of a "universalistic motive utilitarianism," which focuses on the consequences of everyone, or of a large majority, sharing the agent's motives. Similarly, Bradley (2005, Section IV), building on individualistic formulations of virtue consequentialism by Zagzebski (1996), Thomson (1997), and especially Driver (2001), suggests that a virtue consequentialism can embrace what he calls "universalism," and identify character traits as virtues if they would have positive consequences when everyone, or a large majority, possesses them. These theories, however, have never been worked out in detail. Nevertheless, as Adams (1976, 480) predicts, it is likely that they will suffer the same formulation problems as rule consequentialism. We discuss these problems in section 2.

A battle between Hare and Singer ensued over the proper terminology for the principle. Hare (1963, 39) recognized that 'universality' and 'generality' "are often used interchangeably," but he insisted that they are distinct, because generality admits of degrees while universality does not.[4]

These historical facts have left some confusion about the best way to match terms to concepts. We propose to use the term 'universalizability' to refer to the concept identified by Hare and called 'the generalization principle' by Singer. This appears to be the most common usage today. Meanwhile, 'generalization' should be reserved for talking about the concept associated with what Singer called 'the generalization argument' and its characteristic use in exhortations against free riding. This also seems to correspond to the way this term is commonly used.[5]

Standard act consequentialism appears to respect universalizability. Indeed, universalizability has been seen as a desirable feature of utilitarianism since Hare. In contrast, standard act consequentialism does not respect generalizability. It has no concern for the question, "What if everyone did that", interpreted as a question about the consequences in a possible world in which everyone acts like the agent does. It thus can evaluate free riding as morally permissible.

Singer's generalization argument identified a common source of dissatisfaction with act consequentialism, but it left much unspecified. What are we supposed to imagine, exactly, when we imagine "everyone" "doing" the act in question, and what does it mean for the imagined consequences to be "disastrous (or undesirable)"? Rule consequentialism attempts to answer these questions in more detail. We turn to this strategy in the next section.

2. Rule Consequentialism

The main idea of rule consequentialism is that we can incorporate generalizing intuitions into consequentialist theory by shifting the locus of evaluation from acts to rules. Rule consequentialists argue that an act's moral standing cannot be evaluated by looking just at the act itself — we must instead determine the act's relation to a set of ideal rules, or "moral code." While this idea might appear promising, it has proven difficult to formulate precisely.

To pick out the relevant rules, rule consequentialists need to specify several variables, including, i) the attitude people have toward the rules (e.g., conformity or acceptance), ii) the percent of the population that has this attitude, iii) whether the actual or expected consequences matter, and iv) whether, and in what way, transition costs matter to the calculation. For example, Brad Hooker's (2000, 32) influential rule-consequentialist theory of moral wrongness is:

---

[4]Hare also argued that 'specific' is the opposite of 'general' and 'singular' the opposite of 'universal,' and so since 'general' and 'singular' have different meanings so must 'general' and 'universal.' To this, Singer (1985, 49) retorted, "What is singular is that he thinks this is illuminating."

[5]Potter and Timmons (1985, xii–xiii) propose that we distinguish between "non-substantive" and "substantive" universalizability principles. They call Hare's concept a "non-substantative" universalizability principle because it "does not entail, either alone or together with other non-moral premises, any moral conclusions of the sort that something (some action, person, state of affairs) has a certain moral property." Non-substantive universalizability principles, they claim, are principles of "ethical consistency." By contrast, "substantive" universalizability principles, like Singer's generalization argument, "set forth a standard or test for determining in connection with other non-moral information, the moral acceptability of something." One problem with this classification is that in later work Hare claimed that his principle of universalizability does indeed entail moral conclusions. He wrote, "the requirement to universalize our prescriptions generates utilitarianism" (Hare, 1981, 11).

> An act is wrong if and only if it is forbidden by the code of rules whose internalization by the overwhelming majority of everyone everywhere in each new generation has maximum expected value in terms of well-being (with some priority for the worst off). The calculation of a code's expected value includes all costs of getting the code internalized. If in terms of expected value two or more codes are better than the rest but equal to one another, then one closest to conventional morality determines what acts are wrong.

Like most rule consequentialists, Hooker focuses on the acceptance of a code — or, using the terminology he favors, its "internalization" (which requires the agent to have certain moral reactive attitudes to the code) — rather than compliance with it. He does this because the acceptance or internalization of a code can have consequences separate from merely complying with it, and especially so in situations involving deterrence or coordination. This desire to account for all the consequences of a code's acceptance follows Brandt's (1963, 120–25) influential defense of rule utilitarianism.[6] Also following Brandt (1967, 48), Hooker (2000, 82) defines an "overwhelming majority" as 90% of the total population.

In contrast to historical objections to rule consequentialism, the objections that are most relevant for our purposes do not concern the general motivations of the project.[7] Indeed, we are in favor of the attempt to create a consequentialist-friendly theory of generalizability, and especially so given the prominence of the generalizability argument in pre-theoretic moral reasoning. Our worries concern the formulation of the theory — the way in which rule consequentialists have attempted to introduce generalizing to consequentialism.

The first problem is arbitrariness. Why does Hooker build 90% internalization into the theory instead of 100%? The answer is that imagining 100% internalization results in the selection of a code unfit for deterring would-be offenders. This is because the internalization of rules for punishing and deterring those who have not internalized the code involves some cost. If we imagine that there are no people to deter, then the internalization of such rules is a pure cost. Yet, it is important that the chosen code have something to say about the moral permissibility of punishing and deterring potential offenders.

This feature of rule consequentialism is not easily avoidable. Whether a theory focuses on internalization, acceptance, or compliance, the theorist must specify what percentage is relevant to the evaluation. And since 100% is problematic — because it makes punishment

[6]Indeed, arguments in favor of acceptance-based rule consequentialism have been so influential that compliance-based rule consequentialism has been virtually abandoned in modern moral theory. We believe this is justified. The theory we present in this paper, generalized act consequentialism, while significantly different from both acceptance and compliance-based rule consequentialism, is nevertheless more similar in spirit to the more popular acceptance-based version. (See footnote 22 for more.)

[7]Historical objections to rule consequentialism include the charges that it collapses into act consequentialism, has incoherent motivations, and that it involves rule worship. These objections were once thought fatal to the theory, but recent responses by rule consequentialists have brought this view into question. Against the collapse objection, rule consequentialists point out that the internalization of rules has consequences separate from the agent's actions, and that internalization of the rule "maximize expected consequences" *does not* bring about the best consequences. The charge of incoherence points out that if the ultimate and overriding goal of rule consequentialism is to maximize the good, then it is incoherent for it to recommend actions that are known not to maximize the good. Hooker (2000, 99–102) responds to this by claiming that the ultimate goal of rule consequentialism is not to maximize the good but to *match and tie together our moral convictions* and to specify *what is impartially defensible*. If these are the fundamental goals of rule consequentialism, Hooker argues, then there is nothing incoherent in it sometimes recommending not doing what maximizes the good. Always maximizing the good, after all, is inconsistent with many of our moral convictions. Finally, against the charge of rule worship (Foot 1985, 196; Smart 1956, 24–5), Brandt (1992, 150–151) responds that the ideal rules would include something like a "prevent disaster" rule to override other rules in situations in which following them would have disastrous consequences.

and prevention mechanisms morally impermissible — rule consequentialists must find some non-arbitrary way of specifying a different rate.[8]

Proponents of rule consequentialism take the arbitrariness worry seriously.[9] Michael Ridge (2006, 245) writes that 90% seems to be a number "pulled out of thin air" and that "defenders of rule-utilitarianism seem slightly embarrassed by this feature of their theory." Indeed, the problem is so pronounced that it is one of the two reasons that Hooker refuses to endorse rule consequentialism in the afterword to *Ideal Code, Real World*. He writes, "I am unsure what to think. […] My formulation of rule-consequentialism selects rules by reference to the expected value of their being internalized by the 'overwhelming majority' in each future generation. I am worried by the question of what exactly constitutes the overwhelming majority" (188–9).

Ridge (2006) proposes that rule consequentialists can avoid arbitrariness by adopting his "variable-rate" rule consequentialism. This theory recommends evaluating the expected value of a moral code at *every* level of acceptance (i.e., the expected value at 1%, 2%, 3%, and so on), and selecting the moral code that has the greatest average expected value. By doing so, Ridge claims, the rule consequentialist can avoid the arbitrariness associated with selecting a specific percentage of acceptance.

However, it is not clear that Ridge's proposal does manage to avoid arbitrariness. Holly Smith (2010, 416) writes, "It is true that testing a code by its acceptance-[value] at exactly ninety per cent acceptance seems arbitrary — but why test its acceptance-[value] at every possible level, when in reality it must be accepted at some level or the other?" As an alternative, Smith suggests "optimum-rate" rule consequentialism, which selects the moral code with the greatest maximum expected value at any single percentage. So, on Smith's optimum-rate theory, a moral code that has a maximum expected value of 10 (at, say, 60% acceptance) would be superior to a moral code with a maximum expected value of 9 (at, say, 95% acceptance).[10]

Recently, Dale Miller (2020) has argued that *any* rule-consequentialist theory that focuses on percentage of acceptance or compliance will suffer from arbitrariness or other problems, and that rule consequentialists should therefore stop the search for the "right" acceptance or compliance rate. Instead, he proposes "uniform-moral-education rule consequentialism," which posits that the ideal code is the one that would maximize expected value *to teach*. On Miller's proposal, therefore, a code is "accepted" if it is uniformly taught; there is no need to specify the level of acceptance.

A second problem concerns the general format of the rules themselves. As Hooker (2000, 5) notes, many acts seem morally permissible but would be disastrous if everyone were required to do them. For instance, it is often permissible to sit aboard a particular canoe, but if everybody sat aboard that same canoe at that same time, it would sink. Instead, Hooker

---

[8]Other generalizing consequentialisms, such as Adams' "universalistic motive utilitarianism" or Bradley's "universal virtue consequentialism," (see footnote 3) also face pressure to posit a level of generalization less than 100%. Regarding motive consequentialism, Adams (1976, 480) writes, "If we try to state it as the thesis that motives are better, the greater the utility of *everybody's* having them on *all* occasions, we implausibly ignore the utility of diversity in motives." Similarly, problems with formulating virtue consequentialism in terms of everyone sharing some character trait causes Bradley (2005, 288–9) to suggest that attributions of virtue must always be relative to a person or population.

[9]Our discussion of recent rule-consequentialist theories in this section benefited greatly from Miller (2020).

[10]Part of the motivation for Smith's theory is the idea that reasonable moral codes will tend to increase in expected value as the level of acceptance increases until reaching a turning point at which the costs associated with inducing further acceptance would override the benefits. So, in the example presented here, the first moral code may be a complex difficult-to-teach code that reaches the turning point at 60% acceptance while the second code may be a simple easy-to-teach code that does not reach the turning point until 95% acceptance.

suggests that the rules should specify what people can *feel free* to do. However, this fails to fix a related concern: the consequences of everyone feeling free to do something change with people's desires. Such consequences also change along with environmental factors. For example, as an area becomes more populated, or as the existing population becomes more interested in hunting, the negative consequences of everyone feeling free to hunt become greater. How does rule consequentialism's evaluation of moral codes adjust to factors that change within-generation? If the evaluation focuses on the large-scale inculcation of rules into new generations, this is unclear. At the least, rule consequentialism is faced with a curious tradeoff between nimble but difficult-to-inculcate rules and clumsy but easy-to-inculcate rules. (We return to this problem in Section 6).

The final problem for rule consequentialism is the most fundamental. Despite its initial motivation, detailed formulations of rule consequentialism do not capture the basic intuition behind the generalization argument. When Molly is considering whether to stay home from the polls and she asks herself "what if everyone did that?" she is not inquiring about the consequences of teaching the next generation of people to stay home from the polls. Rather, she is wondering what it would take, right then and there, to "do her part" in helping to elect Stephen. Similarly, when we pose the same question to a litterer, liar, or cheater, the question does not concern the consequences of rule-inculcation. Rather, in each instance, the generalization question seems to concern a hypothetical change to the *current* situation: what if these were the rules that people, including people in current and past generations, generally followed? (Indeed, many people assume, falsely, that modern rule-consequentialist theories are an attempt to answer this hypothetical question). Without a connection to this question — the *real* generalization question, in our opinion — rule consequentialism seems unmotivated, and especially so in comparison to act consequentialism.

It seems that rule-consequentialist theory has become progressively less connected to generalizing intuitions as theorists grapple with the problem of specifying a non-arbitrary level of acceptance. Simply specifying that 90% of people accept the code, as Brandt and Hooker do, may feel arbitrary, but it at least seems to be in the ballpark of what generalizers have in mind when they ask "what if everyone did that?" Variable-rate and optimum-rate rule consequentialism are designed to be less arbitrary, but this is at the cost of moving further away from the spirit of the generalization argument. Perhaps Miller's uniform-moral-education rule consequentialism is the best theory yet at avoiding arbitrariness, partly because it abandons the attempt to specify a level of acceptance entirely — Miller focuses instead on what moral code it would maximize expected value to teach. Miller's theory seems to answer a question like, "what if we tried to teach everyone to do that?" This is not, we submit, what people like Molly have in mind when wondering whether morality requires them to vote.

We suggest that the problem is due to a mismatch between the spirit of the generalization question, "what if everyone did that," and the way that rule consequentialism has been developed. The person asking this question need not be seen as requesting a shift in evaluation from acts to rules. Instead, they are best interpreted as appealing to a noncausal dependency between acts. This realization motivates us to explore an alternative approach to generalization that focuses on theories of act-state dependence rather than procedures for evaluating sets of rules. Our proposed theory, generalized act consequentialism, attacks the problem by adjusting the decision theory of consequentialism. This provides the conceptual machinery to capture the spirit of the real generalization question while avoiding arbitrariness and accounting for rapid changes to sociological and environmental conditions.

3. Decision Theory and Consequentialism

Expected-utility versions of act consequentialism include two components: an expected-utility theory, such as causal or evidential decision theory, and an account of the good. Such a theory, viewed either as a criterion of rightness or a decision procedure, is advocated by, inter alia, Smart (1973); Gruzalski (1981); Zimmerman (2005); Timmons (2002); Mason (2003); Oddie and Menzies (1992), and Jackson (1991). As Jackson (1991, 463–64) puts it, "The proposal is to recover what an agent ought to do at a time according to consequentialism from consequentialism's value function [...] together with the agent's subjective probability function at the time in question in the way familiar in decision theory, with the difference that the agent's preference function that figures in decision theory is replaced by the value function of consequentialism. [...] Whereas decision theory enjoins the maximization of expected utility, consequentialism enjoins the maximization of expected *moral* utility."[11]

The difference between causal and evidential versions of act consequentialism is often noted by ethicists interested in expected-utility act consequentialism but rarely discussed in detail.[12] In contrast, here we explore the possibility that the decision theory of act consequentialism is the key to generalizing it. To create a precise and well-motivated generalized consequentialism we replace the decision theory of standard act consequentialism with one suitable for generalization.

Before we discuss the finer points of different candidate decision theories, however, we pause in this section to reflect on the conceptual machinery decision theory has already contributed to act consequentialism.[13]

Suppose you are an act consequentialist, but you have no understanding of expected-utility theory. You are uncertain what particular consequences might result from various acts and are trying to figure out how to decide what to do. You may start by invoking various candidate methods for determining what to do in different types of situations. For instance, you might say, correctly, that if one is certain act A will have better consequences than any other act, then one should perform A.

But most real-life cases come with no such guarantees, and you must give some advice when the outcome is uncertain. Perhaps, at first, you think that if A will probably have a better outcome than B, then you should choose A. But, of course, it's easy to come up with counterexamples. For instance, consider the following case, adapted from Jackson (1991):

> **Drugs Mark i :** Beverley is treating her patient Wesley for a rare and irritating skin condition. She can either give him drug X or Y. Drug X will relieve Wesley's condition without curing it and carries no other side effects. Drug Y is very likely to completely cure Wesley's condition, but it also has a non-negligible chance of resulting in instant death.

Even without a formal decision theory, the answer here is clear. Beverley should give Wesley drug X (at least on the most reasonable interpretations of the vignette). However, giving drug Y will probably result in a better outcome than X will.

So, perhaps you propose that if A will probably have a better outcome than B but doesn't also carry the potential to be massively worse, then you should choose A. However, once again, it's easy to provide a counterexample:

---

[11] For simplicity, Jackson (and we as well) restrict attention to maximizing versions of consequentialism.

[12] See, e.g., Jackson (1991, 464, fn. 4).

[13] Our hypothetical narrative in this section draws on Feldman (2006)'s historical account of the incorporation of expected-utility theory into utilitarianism.

DRUGS MARK II: Beverley is treating Will for a second type of skin condition that causes itchiness and moderate pain. She can either give him drug *Q* or *R*. Drug *Q* will not cure the condition, but will make the condition only mildly painful without doing anything for the itchiness. Sixty percent of the time, Drug *R* will also make the condition only mildly painful but it will also make it very slightly less itchy than it is currently. Forty percent of the time, however, it will make it significantly itchier than it is currently, slightly increase the pain, and also cause moderate daily headaches.

There's no potential for disaster in this case, but presumably Beverley should nonetheless prescribe Drug *Q*. So, we need to modify our decision principle further.

Eventually, without invoking expected utility, you will end up with semi-adequate principles for dealing with uncertainty. They might include the principle that an agent should do *A* instead of *B* iff *A* has "greater net expectable utility" than *A* (Brandt, 1959, 382)[14] or "*A* will produce more probable benefit than *B*" (Smart, 1973, 30). These principles, at least some of the time, produce the wrong recommendations.

Expected-utility theory tells you, roughly, to weight the goodness of possible consequences by their probability given an action. Using this theory, you can formulate precisely when an unlikely but massively good potential consequence from *B* would outweigh the fact that *A* will usually have better consequences.

Expected-utility theory thereby provides the conceptual apparatus necessary for act consequentialism to deal with uncertainty about outcomes. The decision theory itself does not tell us what "the good" is, nor does it tell us how to aggregate the good across individuals. Indeed, in its pure form, decision theory remains almost entirely neutral about these questions. But it does tell us how to combine axiology with uncertainty.

We contend that consequentialism is currently in an analogous position with respect to generalizability. To see why, let us reflect on how defenders of rule consequentialism have developed the theory.

Those sympathetic to the generalization argument first realize that act consequentialism does not capture important pre-theoretic reasoning in sometimes allowing free-riding. To overcome the free-rider problem, the first solution that comes to mind might be to evaluate the expected utility of "everybody" following a given rule. However, some actions are morally permissible but would have bad consequences if everybody did them. So, the rule consequentialist changes the question from "what if everybody did that?" to a related question such as, "what if everybody felt free to do that?"

However, as we saw, that question still fails to produce the right results because we need guidance about how to respond to wrong-doers who do not follow the rules. In response to this, the rule consequentialist introduces an arbitrary cutoff, such as 90% of the population following the rule.

The rule consequentialist then needs to more precisely specify what sort of relation obtains between the agent and the rules. A natural first thought is that the agent simply "follows" the rules. However, rules can sometimes produce consequences independent of the actual actions of agents. And so rule consequentialists construct different relations; e.g., that the rules are "accepted" or "internalized."

But what happens when an agent can foresee that the act recommended by the rules she has internalized will result in a worse outcome than some other act? Rule consequentialists

[14]Brandt's "expectable utility" differs from expected utility in that expectable utility is a measure of an agent's preferences over sets of outcome-probability pairs, whereas expected utility is the probability-weighted average of the agent's preferences over outcomes. See Brandt (1959, 382-83).

respond to this problem by positing that the agent should perform the suboptimal act unless the consequences are particularly bad. They developed this thought into a "prevent disaster" clause that goes into effect when the consequences are "bad enough" (Brandt 1992, 87–88, 150–51, 156–57; Hooker 2000, 98–99).

At the end of this process, we are left with a theory that is only semi-adequate. It contains, at best, some good heuristics, but ultimately is unsystematic, ambiguous, and counterexample-prone. Further, it has become unhinged from how people actually reason when worried about generalizability. We claim that just as expected-utility theory helped act consequentialism avoid these problems when dealing with uncertainty, the right expected-utility theory will also help us systematically handle generalizability. To see how this might be accomplished, we first turn to the details of the kind of decision theory that supports non-generalized act consequentialism. We then present a new type of decision theory and explain how it provides a promising conceptual apparatus for a generalized act consequentialism.

4. Evidential and Causal Decision Theory

Evidential and Causal Decision Theory (EDT and CDT) are currently the two leading theories of instrumentally rational action. Both are concerned with achieving your ends regardless of what they might be (moral or immoral, selfish or altruistic), and both advise agents to maximize expected utility. The theories differ, however, in how they understand what expected utility amounts to.

According to EDT, you should choose the act that, in your estimation, is the best *indicator* of a good outcome, whereas according to CDT you should choose the act that, in your estimation, will *bring about* the best results.

We start with EDT. Consider:

> **Vitamins:** You would like to avoid kidney disease. You read a report that says people who take a daily multivitamin are much less likely to develop the disease than those who don't. Although vitamins are somewhat costly, you would much rather be healthy and a bit poorer than unhealthy and a bit richer. The report concludes that there's some genetic predisposition that both prevents people from acquiring the disease and causes them to take vitamins.[15]

If you believe the report, you know taking vitamins is correlated with good health but that it doesn't cause good health. We assume you have no independent way of discovering whether you have the predisposition that prevents kidney disease.[16]

According to EDT, it doesn't matter whether taking vitamins causes health or is instead just an effect of the predisposition. Taking vitamins is an indicator of health. So, you should take vitamins. Correlation is all that matters according to EDT.

Causal Decision Theory (CDT), on the other hand, thinks that what matters is whether taking vitamins actually causes good health, and not just whether taking vitamins is a good omen. So, if you learn that some predisposition that you can't change now causes both taking vitamins and good health, CDT says to save your money.

To spell out CDT and EDT precisely, we'll need a number of ingredients. First, there is the space of acts $\mathcal{A}$ that the agent can perform. In the case of Vitamins, $\mathcal{A}$ includes the acts of taking vitamins and not taking them.

Second, there is the space of outcomes. Outcomes are the bearers of ultimate value for the agent. In other words, a given outcome is a set of worlds that the agent is indifferent

[15] This case is a variant of the standard Medical Newcomb case from Gibbard and Harper (1978).

[16] Supporters of the "tickle defense" claim that in real life this assumption is unwarranted, and thus that EDT avoids any problem associated with cases like this. See Ahmed (2014b, 91–7) for a recent example and further references.

		State	
		s_1	s_2
Act	T	9	0
	$\neg T$	10	1

FIGURE 1. Payoff Matrix in VITAMINS.

between. Here, there are four outcomes: (o_1) worlds where you are free from kidney disease and don't spend money on vitamins, (o_2) worlds where you are free from kidney disease and spend money, (o_3) worlds where you have kidney disease and don't spend money, and (o_4) worlds where you have kidney disease and spend money.

Third, agents come equipped with a *utility function* u that quantifies how much she likes (or dislikes) each outcome. Let's assume $u(o_1) = 10, u(o_2) = 9, u(o_3) = 1,$ and $u(o_4) = 0$.

Fourth, there is the space of *states of the world* $\mathcal{S}$. It's a little harder intuitively to specify what a state of the world is, but the idea is that any given state of the world and any given action produce a unique outcome. So, states can be thought of as 'background conditions' that are not directly under the agent's control that combine with an action to produce an outcome. Formally speaking, actions map states of the world to outcomes. For simplicity, assume that in VITAMINS, there are only two states:

- s_1: You have the predisposition and will be healthy no matter what.
- s_2: You don't have the predisposition and won't be healthy no matter what.

Letting T refer to the act of taking vitamins, we then have the payoff matrix in Figure 1.

Finally, because the agent does not know which state of the world is actual, she also has a credence function Pr that encodes how strongly she believes various hypotheses.

According to EDT, you should consider how likely your actions render various states and how good the outcomes of those state-action pairs would be. That is, you should discount the *utility* of an act-state pair by the probability of that state *given* you perform the act.

So, EDT calculates the expected utility of an act as follows:

$$\mathrm{EU}_{\text{EDT}}(a) = \sum_{s \in \mathcal{S}} \Pr(s \mid a) u(o(a, s))$$

The agent should then pick the act that has the highest expected utility according to this method of calculating expected utility. In our VITAMINS example, you should compare how likely it is that you'd be healthy on the indicative supposition that you do versus do not take vitamins.

You know that if you take vitamins, you'll end up in either the upper left or upper right cell in Figure 1. We now need to assign probabilities to s_1 and s_2 conditional on T and $\neg T$. To make things concrete, let's assume $\Pr(s_1 \mid T) = .6$ and $\Pr(s_1 \mid \neg T) = .4$.

Then,

$$\begin{aligned} \mathrm{EU}_{\text{EDT}}(T) &= .6 \cdot 9 + .4 \cdot 0 \\ &= 5.4 \\ \mathrm{EU}_{\text{EDT}}(\neg T) &= .4 \cdot 10 + .6 \cdot 1 \\ &= 4.6 \end{aligned}$$

So, EDT says to take vitamins.

CDT says to consider the *causal* effects your actions might have and to choose the action with the best causal effects in expectation. Because CDT cares only about causal relationships between acts and states (and not mere correlation between acts and states), it appeals to

causal counterfactuals: What *would occur* were I to perform act a? We denote this counterfactual $a \mathrel{\square\!\!\rightarrow} s$, where $\mathrel{\square\!\!\rightarrow}$ is interpreted causally. So, $\Pr(a \mathrel{\square\!\!\rightarrow} s) > \Pr(s)$ just in case the agent thinks performing a will causally promote s, $\Pr(a \mathrel{\square\!\!\rightarrow} s) < \Pr(s)$ just in case the agent thinks a will causally inhibit s, and $\Pr(a \mathrel{\square\!\!\rightarrow} s) = \Pr(s)$ just in case the agent thinks a is causally independent of s.

According to CDT, an agent should evaluate the expected utility of an action as follows:

$$\mathrm{EU}_{\textsc{cdt}}(a) = \sum_{s \in S} \Pr(a \mathrel{\square\!\!\rightarrow} s) u(o(a, s))$$

This formula weights the value of each potential outcome of her action $u(o(a, s))$ for each state of the world s by how likely that state would be *were* she to perform that act, i.e., by $\Pr(a \mathrel{\square\!\!\rightarrow} s)$. She should then perform the act with the highest expected causal utility.

In the case, vitamins neither cause nor inhibit good health. So, the probability that you *would* be free from kidney disease *were* you to take vitamins is equal to the probability you'll be free from kidney disease. I.e., $\Pr(T \mathrel{\square\!\!\rightarrow} s_1) = \Pr(\neg T \mathrel{\square\!\!\rightarrow} s_1) = \Pr(s_1)$ and $\Pr(T \mathrel{\square\!\!\rightarrow} s_2) = \Pr(\neg T \mathrel{\square\!\!\rightarrow} s_2) = \Pr(s_2)$.

So, according to CDT:

$$\mathrm{EU}_{\textsc{cdt}}(T) = 9 \cdot \Pr(s_1) + 0 \cdot \Pr(s_2)$$
$$\mathrm{EU}_{\textsc{cdt}}(\neg T) = 10 \cdot \Pr(s_1) + 1 \cdot \Pr(s_2)$$

No matter what $\Pr(s_1)$ and $\Pr(s_2)$ are, CDT will recommend not taking vitamins.

Most philosophers agree with the verdict of CDT in VITAMINS. However, as we will see, CDT is not a good engine for generalization.

Consider the following problem:

> **TWIN PRISONER'S DILEMMA:** Aomame and her psychological twin are put in separate rooms and cannot communicate. If they both cooperate, they each get $5. If they both defect, they each get $1. If one cooperates and the other defects, then one gets $10, and the other gets $0. Assuming Aomame only cares about her individual payout, what should she do?

Aomame can no longer have any causal influence over what her twin does, as they're in separate rooms. I.e., $\Pr(c \mathrel{\square\!\!\rightarrow} C) = \Pr(C)$ and $\Pr(d \mathrel{\square\!\!\rightarrow} D) = \Pr(D)$ (where c and d denote Aomame defecting, and C and D denote her twin cooperating and defecting, respectively). Therefore, CDT reasons that Aomame's choice does not affect her twin's choice in any decision-relevant way: either her twin cooperates or her twin defects, and there is nothing Aomame can do about what her twin decides. In either case, Aomame is better off if she defects, so she ought to defect (Joyce, 1999, 148–50).

Of course, since Aomame and her twin are so alike, they will almost surely each wind up with $1. Had they both cooperated, however, they would have both ended up with $5. The problem, from the point of view of those who find the generalization argument plausible, is that Aomame knows that if *everyone* in the problem defects, she'll get less than if *everyone* cooperates. And she also knows that her twin will reason like she does. Nonetheless, CDT recommends that she defect since she can't causally influence her twin.

It may seem, then, that EDT, despite whatever flaws it has, does handle generalization correctly because it tells Aomame to cooperate. After all, the probability that her twin cooperates given that she does is much higher than the probability that her twin cooperates given that she defects.

We agree that EDT does do a better job with generalization overall than CDT. CDT only pays attention to a very special type of correlation between act and state—namely when that correlation reflects the fact that the act causally promotes or inhibits that state. So, if your action does not causally promote others behaving in the way you want, CDT will suggest free-riding. EDT, on the other hand, thinks *every* type of evidential correlation between act and state is potentially decision-relevant. In VOTING and TWIN PRISONER'S DILEMMA, evidential correlations that do not involve causation will help rationalize cooperative behavior.

However, while CDT is too spartan in its account of what counts as a relevant type of correlation, EDT is too generous. It is this generosity that, in practical decision-making, leads to taking vitamins even when vitamins are known not to prevent kidney disease. In the case of generalization, EDT will lead to the wrong verdicts as well.

To see why, let's return to Molly's voting predicament, introduced in Section 1, and add a few more details. Molly thinks Stephen will do a better job than Leopold. However, she also thinks that Stephen will likely only be elected if it's already inevitable that a major economic depression will strike her country in the next few years regardless of what anybody does. If Leopold is elected, then she thinks such a depression is unlikely. The adverse effects of a depression will outweigh any good or bad caused by either Stephen or Leopold. So, she thinks, conditional on her voting for Stephen, it's likely that fellow agents like her will vote for Stephen, in which case it's likely that there will be an economic depression. On the other hand, if she doesn't vote (or votes for Leopold), then it's likely similar agents will do the same, and more likely that no depression will occur. EDT then (wrongly) recommends not voting for Stephen.

Another reason EDT is insufficient for modeling the intuition behind the generalization argument is that EDT only supports generalization when the agent is ignorant of what others will do. For example, if Molly is certain that her fellow Stephen-supporters will vote, then her voting will not provide evidence that they will vote. EDT therefore recommends that she stay home. However, generalizers want to do their part even when (or *especially* when) they know that others will do the same. EDT fails to capture this intuition.

So, if the act consequentialist uses EDT, then they will be left with a theory that engages in generalization in a haphazard manner: sometimes generalizing inappropriately and sometimes failing to generalize when appropriate. If CDT is used instead, then they will be left with a theory that does not engage in generalization at all. Since CDT considers only the causal consequences of actions, it ignores the non-causal dependency that the generalization argument targets.

5. FUNCTIONAL DECISION THEORY

Functional decision theory (FDT) is a relatively new decision theory that can be used to model generalization in a principled way.[17] The critical feature of FDT is that it recognizes more than just causal dependencies as decision-relevant, but, unlike EDT, it does not see every evidential dependency as relevant.

To see how FDT works, we first return to the TWIN PRISONER'S DILEMMA. (This is not the place for a full-throated defense of the claim that it is instrumentally rational to cooperate in TWIN PRISONER'S DILEMMA.[18] We argue only that FDT is a good engine for creating a

[17]For recent philosophically accessible formulations of FDT, see Levinstein and Soares (2020) and Soares and Yudkowsky (2018).

[18]Arguments for cooperation in one-shot TWIN PRISONER'S DILEMMA include Hofstadter (2008, Ch. 29) and Horgan (1981, 1985). For metanormative arguments in favor of functional decision theory as a theory of instrumental rationality, see Greene (2018) and Yudkowsky (2010). Greene argues that functional decision theory is motivated

generalizing consequentialist moral theory.) Like CDT, FDT appeals to counterfactuals about what would occur were a certain action chosen. If some feature of the state of the world is merely evidentially correlated with your act, then FDT agrees such a correlation is irrelevant for decision-making purposes. Nonetheless, according to FDT, Aomame's twin's action *does* depend in a relevant way on what Aomame chooses, even though they're causally isolated.

Since they're psychological twins, both Aomame and her opponent are using the same procedure (or algorithm) to decide how to act. That is, they are both reasoning in the same way about the problem, and their reasoning process, whatever it is, will almost surely yield the same result.

Aomame and her twin are thus both tokens of the same logical type, like two physical computers running the same program. The function that governs how they act is the same, and it will generate the same output whether it is instantiated by Aomame or her twin. Because of their (nearly) identical decision procedures, Aomame knows that were her algorithm to output *cooperate*, her twin's would as well, and were it to output *defect*, her twin's would as well. So, according to FDT, Aomame's algorithm should counterfactually co-vary with her twin's. Since it's better for Aomame if her algorithm outputs *cooperate*, and Aomame can choose whether her algorithm results in cooperation or defection, FDT advises her to cooperate.

FDT thus takes the *algorithmic* dependence between Aomame and her twin into account. They use the same decision procedure, which is multiply instantiated in their cognitive architectures. Because of this dependence, FDT imagines that her twin's decision would change if hers did, despite their physical isolation. (There is a budding similarity here to the way in which a consequentialist generalizer imagines their decision co-varying with that of others).

By now, perceptive readers may have noticed that FDT— as well as the consequentialist generalizer that we aim to model — must use a different kind of counterfactual from the ones that feature in CDT. Instead of considering counterfactuals of the form $a \mathrel{\square\!\!\rightarrow} s$, FDT considers counterfactuals of the form $\texttt{MyAlg(input)} = a \mathrel{\square\!\!\rightarrow} s$.[19]

Let's unpack this. FDT considers what would be the case if your algorithm itself on a particular input were to output a. Such an output, on this reading of the counterfactual, would occur at every instantiation of your algorithm given the same input.[20] So, if another agent instantiated your decision algorithm, then according to FDT, if your algorithm were to output a for you, it would also output a for them.

The input to the algorithm includes a description of the decision problem. The problem contains relevant probability and utility functions along with information about what the deciding agent thinks about dependencies, both of algorithmic and causal form. Because the FDT agent cares about the relationship between her decision procedure and the decision procedures of other agents, the input will include all information about such connections.

---

by a "success-first" approach to decision theory that focuses on the causal consequences of decision theories instead of acts. He claims that the desire to create a success-first decision theory has motivated several influential decision theories in philosophy, including those of Gauthier (1986, Ch. 6), McClennen (1990), and Meacham (2010, 68–69). Functional decision theory is, in our view, the most complete, compelling, and well-motivated of these theories to date.

[19]A more complicated version of FDT searches over functions from the set of possible inputs to actions. It then calculates which function has highest expected utility overall, and chooses the action recommended by that function on the current input. The more complicated version is important for solving certain coordination problems. The formulation we use here is easier for expositional purposes, however, and makes the structural relationship with CDT more perspicuous. See Soares and Yudkowsky (2018).

[20]For simplicity, we assume the algorithm is deterministic. If it's indeterministic, then various outputs could occur with different probability, but this is a wrinkle best not addressed here.

In other words, the input includes all relevant information about what the agent believes and what she cares about.

FDT, then, takes a new position on the question of what relationships between acts and states are decision-relevant. EDT thinks all evidential dependencies are decision-relevant. CDT thinks only causal dependencies are decision-relevant. FDT cares about causal *and algorithmic* dependencies, where algorithmic dependencies are those that result from considering the agent as an instantiation of a procedure.

Nonetheless, in many ways, FDT is structurally like CDT. FDT's notion of expected utility is:

$$\mathrm{EU}_{\text{FDT}}(a) = \sum_{s \in \S} \Pr(\texttt{MyAlg(input)} = a \mathbin{\square\!\!\rightarrow} s) u(o(a, s))$$

FDT thus considers the counterfactual probability of what *would* happen were your algorithm to output a given the input. This is a subjunctive supposition: it holds fixed factors that do not depend on your action, just as CDT does. However, it has a broader notion of dependence, as it holds that Aomame's twin's action depends on how Aomame's own algorithm behaves.

Unlike EDT, FDT does not consider the mere evidence for a state that your action might provide to be decision-relevant. The same is true of the consequentialist generalizer. In the case above, where Stephen's election provides evidence that there will be an economic depression, presumably such a depression will happen independently of how your algorithm behaves (as the background conditions are either in place for a depression or not regardless of what your decision algorithm outputs). FDT thus explains why the consequentialist generalizer doesn't care about the evidential dependence between Stephen's election and the economic depression. What matters to both FDT and the generalizer (in addition to the causal dependencies) are potential between-agent algorithmic dependencies in voting behavior.

The main difference between FDT and CDT, then, is that CDT produces recommendations using physical counterfactuals about the agent's behavior. It asks, essentially: what would be the case if my body moved in this way rather than that way? In contrast, FDT produces recommendations using logical counterfactuals about the output of one's decision procedure. It asks, essentially: what would be the case if my decision algorithm output this rather than that? Each type of counterfactual involves impossibilities in some sense. CDT's counterfactuals involve small 'miracles' that break the laws of nature. The important question is not which type of counterfactual is more 'real' but rather which type is relevant in evaluating decisions.

5.1. **Similarity.** Unlike in toy examples such as TWIN PRISONER'S DILEMMA, agents rarely use *exactly* the same algorithm to make decisions. So, the FDT counterfactuals must not require identity between algorithms for behavior to co-vary counterfactually. Instead, FDT merely requires agents to be similar in their decision-making process for there to be a subjunctive connection between them.

For example, suppose Aomame is an FDT agent and is in a one-shot Prisoner's Dilemma against Randolf, who randomizes between CDT and FDT. CDT defects in the one-shot Prisoner's Dilemma, but FDT will cooperate against certain opponents. So, for Aomame, $1 > \Pr(\texttt{MyAlg(input)} = c \mathbin{\square\!\!\rightarrow} C) > \Pr(C)$. That is, Aomame thinks cooperating makes it more likely Randolf will cooperate, but there's still some chance he'll defect. Therefore, whether Aomame chooses to cooperate will depend both on how likely she thinks it is that Randolf will use FDT and on what the exact payoffs are.

More generally, however, connections between algorithms are hard to discern. The technical details are not yet worked out, and in any case are far beyond the scope of this paper.

(We note that the same is true of the technical details of causal decision theory.[21]) We will assume below that the relevant notion of similarity is not an absolute one: that is, how tightly one thinks different algorithms are connected will depend in part on the verdicts one wants the theory to deliver.

More formally, we can differentiate between versions of FDT by adding a parameter sim. The sim parameter determines how algorithms counterfactually vary with one another. Different theorists will want different versions of sim depending on their goals for the decision theory and their views on rationality and generalizability.

There are two natural extremes. At one end is the option to vary only other instantiations of the agent's exact decision procedure itself. At the other end, there is the option to connect the agent's procedure to every other decision procedure that is instantiated anywhere (since, at a high level of abstraction, they bear a salient mathematical relationship to one another). How this is cashed out will vary based on the desiderata of the decision theorist (or ethicist using FDT). When using FDT in ethical contexts, we can adjust the sim parameter to capture the level of generalization that matches our moral convictions (more on this below).

5.2. **Generalizability.** The most significant benefit of FDT is that it incorporates generalizing into the notion of instrumental rationality itself. FDT agents take into account the connection between their decision procedure and the decision procedure of other agents. For them, the relevant question isn't just what happens if they alone perform the act, but rather what happens if their decision procedure (or similar procedure) behaves in a certain way wherever it is instantiated.

Thus, even from a purely self-interested point of view, FDT recommends against certain types of free-riding. With opponents sufficiently like her, Aomame will cooperate in a Prisoner's Dilemma, not because she cares about overall welfare, but because she wants a better payoff herself. Likewise, if she thinks there are many agents sufficiently like her, Molly will vote in an election even if one vote can't make a difference and it's raining outside. This behavior need not stem from any sense of moral obligation but merely from the desire to have her favored candidate in charge.

The fact that FDT includes a free parameter to determine which algorithms count as *similar* is one of the main reasons the theory is a promising way to model the intuition behind the generalization argument. Consider that the pre-theoretic generalization question "What if everyone did that?" is ambiguous about who "everyone" refers to. Almost always, we do not have literally everyone in mind. For example, in the voting case, Molly's free-riding is relativized to her fellow Stephen-supporters. She doesn't mind if Leopold supporters stay home. Instead, it seems that Molly's question concerns the actions of people like her: supporters of Stephen who are trying to decide whether to vote and facing roughly the same constraints. FDT models Molly's reasoning by defining sim so that it picks out all and only the agents over whom Molly wishes to generalize her decision. If, instead, Molly wishes to generalize her decision over all potential voters, then sim can be defined in that way as well. (Indeed, another version of the free-rider problem asks the non-voter, "wouldn't it be a disaster if everyone was like you and didn't vote?")

FDT thus has the power to model Molly's generalization reasoning no matter how inclusive or exclusive she wants to be. In Molly's case, FDT interprets the question "What if everybody did that?" to (roughly) "What if everybody like me were to decide like that?," or better, "What if all instantiations of an algorithm like mine were to decide this way in this

[21]For example, theorists have found it difficult to construct formulations of CDT that adequately handle cases involving decision instability, such as those presented by Egan (2007) and Ahmed (2014a). See Arntzenius (2008) and Joyce (2012). Also see Armendt (2019) for a recent overview of the problem and further references.

kind of situation?" What counts as "like me" or "like mine" varies depending on the sim parameter, but for each notion of similarity, we have a theory that cares about more than just the causal consequences of the agent's action. In her free-riding problem, Molly ends up voting if she thinks sufficiently many people are sufficiently like her because she wants Stephen in charge.

5.3. **From FDT to Generalized Act Consequentialism.** We saw above that we can use CDT as an engine for a consequentialist moral theory. The ethicist supplies an axiology, and CDT accommodates uncertainty. Unfortunately, CDT-style act consequentialism cannot accommodate the generalizability argument.

Instead of changing the locus of evaluation to rules (or motives or virtues), we can now use FDT to achieve generalization while retaining focus on acts themselves. Generalized act consequentialism (or GAC) agrees with standard act consequentialism about the theory of the good. It agrees with the act consequentialist that acts are right when they have maximally good (or sufficiently good) expected consequences. However, it uses FDT instead of CDT to calculate the expected utility of acts.

By calculating expected utility with FDT, GAC directly incorporates generalizing into its act evaluations. GAC recognizes that algorithms are type-like entities. So, suppose GAC (`input`) returns action *a* as the uniquely permissible action on this particular occasion. Then, just like FDT, GAC will recognize that the decisions of other agents with related algorithms depend (non-causally) on what it outputs in this instance.[22]

FDT's notion of algorithmic dependence admits a similarity parameter (in addition to the axiology) that the ethicist has leeway to determine. As we noted in Section 5.2, it's not obvious how different algorithms should counterfactually co-vary. We see this as a virtue insofar as FDT can serve as the engine for a generalizable form of consequentialism. The ethicist can specify how inclusive these dependencies should be, and can thereby determine the precise type of generalization reasoning that is captured by the theory.

5.4. **Informal Reasoning of GAC Agents.** To get a sense of how a GAC agent would reason compared to agents following a CDT-driven act consequentialism, consider another variant of the free-rider problem:

> **BOYCOTT:** LexCorp makes DB Cola, which is a soft drink that Jimmy enjoys occassionally. Jimmy finds out that LexCorp has been mining in a remote part of the world to power their factories. The mining operation has exposed locals to dangerous chemicals and has made them sick, but LexCorp has continued operations unabated. LexCorp would only stop if enough people boycotted DB Cola so as to render the mining unprofitable, but this is extremely unlikely to occur as the operation is much more profitable than DB Cola's other options. Furthermore, the number of people made sick and the severity of the illness is caused by the presence of LexCorp's digging operation and not by the minerals they're after. Therefore,

[22] In Section 2, we claimed that GAC has more in common with acceptance-based rule consequentialism than the compliance-based version. This is because FDT evaluates the consequences of *one's algorithm outputting a*, and not the consequences of *everyone doing a* (as compliance-based rule consequentialism would do). This is seen clearly in prediction cases like Newcomb's problem, where FDT views changes to the output of one's decision algorithm as sometimes necessitating changes to what has been predicted (where such an effect is due to the decision algorithm the agent "accepts" and not the act the agent performs). (See Soares and Yudkowsky (2018) and Levinstein and Soares (2020) for more on this aspect of FDT.) Nevertheless, as we have emphasized in the text, GAC also differs significantly from acceptance-based rule consequentialism. The main difference between acceptance-based rule consequentialism and GAC is that the former focuses on procedures for evaluating sets of rules while GAC focuses on algorithmic dependencies between decision makers.

merely reducing the amount of minerals that are mined would not result in any health benefits to the locals.

Jimmy is looking at a can of DB Cola displayed in a convenience store and wondering whether it is permissible for him to buy it. Under a CDT-driven act consequentialism, he might conclude that the mining operation gives him little reason to refrain right now from buying a single can. Even though he enjoys DB Cola, it's exceedingly unlikely that any boycott powerful enough to stop LexCorp will occur, and virtually impossible that this decision alone would render their mining operation unprofitable. So, if Jimmy is a CDT-driven consequentialist, he may reason as follows: *I have the option either to buy this can of DB Cola or refrain. If I were to refrain, LexCorp would not stop mining operations. If I were to buy it, I'll gain utility from DB Cola's delicious flavor. Because the negative causal effects of buying it are effectively zero, I don't have a good reason to refrain.*

In contrast, GAC theory focuses on the consequences of its own recommendation counterfactually covarying with that of other agents, using algorithmic dependencies instead of CDT's causal dependencies.[23] Specifically, the GAC agent imagines a world in which all similar agents reason like this in a similar situation, and asks whether that would have good or bad consequences.

The GAC agent reasons as follows: *If* GAC *were to allow purchasing DB Cola in this case, then* GAC*-like agents would continue to support companies like LexCorp in similar situations, and such companies will continue to cause gratuitous harm for the sake of profit. On the other hand, if* GAC *were to mandate refraining here, then* GAC*-like agents will collectively avoid purchasing from LexCorp. There is a sufficiently high chance that collectively this would amount to a successful boycott and thus compel LexCorp to cease mining operations. Such an outcome is far superior to the utility gained from DB Cola's deliciousness, so it's better if* GAC *mandates refraining, and I therefore refrain.*[24]

Thus, GAC captures the intuition behind the "what if everyone did that" question by focusing on the agent's own decision algorithm and the algorithmic dependencies between themselves and others. It avoids the calculation of inculcation costs and instead directly generalizes its recommendation over other agents as they exist now. In doing so, it hews closely to the spirit of the generalization argument.

6. Generalized Act Consequentialism and Changing Conditions

We now return to the problem of rapidly changing environmental and sociological conditions. First, let's look at how GAC and popular forms of rule consequentialism (discussed in Section 2) manage common resources:

---

[23] GAC also takes into account how its recommendation covaries with its recommendations to the same agent at different times. This aspect of the theory has interesting applications to moral theory as well, but in this paper we focus on between-agent dependencies because those are the most relevant to the generalization argument.

[24] If sim is defined narrowly so that at most very few agents are like the GAC agent, then they reason: *I have the option either to refrain or not. However, since there are very few agents that face a situation similar enough to mine, the effects of* GAC *telling me to refrain will not influence the business decisions of LexCorp. Since this output would also cause myself and the few agents who face a situation similar enough to mine to lose out on the deliciousness of DB Cola, I think it is better if* GAC *says to buy DB Cola in this instance.* In delivering this result, GAC diverges from the non-consequentialist reasons to not buy DB Cola in a way familiar to consequentialist generalizers. It is equivalent to a situation in which a consequentialist generalizer asks a potential free-rider "what if everyone did that?" and the correct answer is, "it would still have the best consequences." The *consequentialist* generalizer, like the consequentialist non-generalizer, is committed to resisting moral indignation if the action cannot be criticized based on consequences.

PEBBLES: You are walking on a majestic beach that has beautiful pebbles. If you take a pebble as a memento, then you cause no harm to the ecology of the beach. However, if a large number of the pebbles were taken, then the beach would be much uglier than it is, and the ecosystem would suffer.

The first question a rule consequentialist might ask is: "What if everybody felt free to take a pebble?" The answer could be good or bad. Suppose at first there aren't that many people around who will ever visit the beach. The consequences of letting people take a pebble are then good, since the resources won't be depleted, and some people will be made happier after acquiring pebbles. So, we may want to inculcate the rule that you should feel free to take a pebble.

Now suppose an unexpected event occurs, and many people have access to the beach two generations later. Unfortunately, they have been inculcated with the rule that they should feel free to take a pebble, and the beach ends up ruined.

PEBBLES is a toy example, but it is often the case that environmental conditions will change what it is good for people to feel free to do. If there weren't that many people, it would be perfectly fine if everybody felt free to release gases into the atmosphere by flying in private jets for convenience. Because there are lots of people, it may not be good for everybody to feel free to do this. Over time, environmental conditions can unexpectedly change.

Perhaps the rule consequentialist will say that this sort of rule is too simple. Instead, the rule should have additional clauses that tell people under what conditions they should feel free to take a pebble or fly in their private jet. Perhaps a sufficiently complicated rule could indeed handle these sorts of cases. The problem is that as rules become more complicated they become more challenging to inculcate (thereby incurring higher transition costs). We find ourselves presented with a dilemma between complicated rules that have high transition costs and simple rules that can't respond nimbly enough to changing conditions.

In focusing on this dilemma, rule-consequentialist theory becomes unhinged from the intuition behind the generalization argument. Generalizers don't care about inculcation costs when wondering whether they should take a pebble or fly in their jet. Instead, they reason more along the lines of GAC, by asking what the best recommendation would be for all agents who who are similar enough to them, given what they know about the world. If you think that few people like you will want to take a pebble from the beach, then GAC allows you to take a pebble. But once you learn that many people will have beach access and want a pebble, then GAC will return a different verdict. That's because GAC takes into account your current views about what the world is like, which are encoded in your probability function.

This reasoning more closely mirrors how generalizers reason in these types of situations: they desire to do their part by behaving in a way that is best for situations like the one they face, and not in behaving in a way that would be best to teach the next generation. If "doing their part" requires complicated reasoning that is difficult to teach others, that doesn't matter — they want to do their part anyways.

Next, consider a more complex case involving the decision to have children. We tend to think that it is good if (most people) feel free but not morally compelled to have kids. However, in certain alternative environments that wouldn't be the case. In some situations, we might have a duty to try to curtail the size of the next generation. Consequently, if the population became too low, we may want to switch policies and instead encourage people to have children. So again, it's hard to prescribe a generational rule that we should inculcate in the population.

GAC, in contrast, takes multiple variables into account. Suppose Umberto is a GAC agent who is deciding whether to have kids. GAC will take into account, inter alia, Umberto's beliefs about the environment, how much Umberto personally values having a child (i.e., what his personal utility function is independent of moral considerations), and what Umberto thinks about how much other agents personally value having a child.

To showcase GAC's flexibility, let's start by assuming that Umberto generalizes his decision over about half of the population. This is because Umberto thinks only about half of people are similar enough to him (in terms of the situation they face or the reasoning process they employ) for their decision to algorithmically depend on what he chooses. Let's suppose that Umberto believes that the dissimilar half will have, on average, one kid per person. He further assumes that all new children will have roughly equally good lives in expectation. Finally, and crucially, he thinks that if more than 1/3 of the agents he is generalizing over have kids, there will be an extreme environmental catastrophe.

Umberto looks to GAC to deliver a verdict given this input. Because half the population is not similar enough to Umberto, GAC holds constant the results of their decisions. However, because the other agents are similar enough, GAC does change what they do in calculating a recommendation. Let's see what GAC will recommend.

First, suppose GAC tells Umberto to have kids regardless of other facts about his situation. When generalized over the others, this means that more than 1/3 have children. That would, according to Umberto's views, result in catastrophe. Because GAC recommends actions based on expected consequences, it does not recommend this. Indeed, GAC must not recommend having kids to more than 1/3 of the population Umberto is generalizing over.

Which part of the population will GAC recommend having kids to? Each agent might consider how much expected utility they associate with their having children. One heuristic for doing this would be to focus on personal desire. GAC would then recommend that Umberto have kids if he is in the top 1/3 in terms of personal desire to have children. When generalized over the relevant population, this recommendation would result in only 1/3 of people having children.

So, Umberto will reason as follows. *If I take myself to be in the bottom 2/3rds of the population in terms of personal desire to have kids, and if* GAC *were to recommend that I have kids in this instance, then there would be a catastrophe from too many people having kids. So, in that case,* GAC *must tell me not to have kids. If I am in the top 1/3rd of the population, then there would be no catastrophe if* GAC *recommended that I have kids.*

Here, again, GAC does a better job than rule consequentialism at capturing the reasoning of the generalizer. The generalizer does not care about rule inculcation and transition costs. Instead, they care whether their behavior would lead to good consequences when generalized over others given what they believe about the current situation and the realities people face right now.

In contrast, the non-generalizing act consequentialist only cares about the causal consequences of their having kids, and is happy to free-ride. Thus, for those who care about the generality argument, the free-rider dispute seems to exist between generalized and non-generalized act consequentialism.

7. Conclusion

As Singer (1955) noted at the start of his influential work on generalization, we are all familiar with the question "What would happen if everyone did that?" Singer took this question to be an exhortation against free-riding, and elliptical for the generalization argument:

"If everyone were to do that, the consequences would be disastrous (or undesirable); therefore, no one ought to do that." The generalization argument identified a common source of dissatisfaction with act consequentialism, but it left much unspecified.

This essay has explored the possibility of using advances in decision theory to create a version of act consequentialism that respects the generalization argument. Expected-utility versions of act consequentialism include two components: an expected-utility theory, such as causal or evidential decision theory, and an account of the good. Generalized act consequentialism captures generalizing intuitions by substituting the decision theory with one appropriate for generalizing: functional decision theory. The result is a view that is both principled and nuanced in its generalizing. It observes the environment as it actually is and tries to allocate its resources to create the most good given a particular way of cashing out the notion of similarity between agents. Unlike a standard act consequentialist, the GAC agent takes into account algorithmic dependencies between her choice and that of others, where these dependencies are sometimes non-causal and non-evidential. Such dependencies are precisely the connections that feature in common-sense generalizing intuitions.

If the preceding is correct, then the characteristic manifestation of the motivation for generalization within consequentialism — the free-rider problem — is best viewed as a disagreement between generalized and non-generalized act consequentialism. These theories do not disagree on fundamental issues like the nature of the good or whether acts should be the sole locus of evaluation, but instead on what type of dependency between acts and states is decision-relevant.

REFERENCES

Adams, R. (1976). Motive utilitarianism. *The Journal of Philosophy 73*(14), 467–481.

Ahmed, A. (2014a, October). Dicing with death. *Analysis 74*(4), 587–92.

Ahmed, A. (2014b). *Evidence, Decision and Causality*. Cambridge University Press.

Armendt, B. (2019). Causal decision theory and decision instability. *The Journal of Philosophy 116*(5), 263–277.

Arntzenius, F. (2008). No regrets, or: Edith Piaf revamps decision theory. *Erkenntnis 68*(2), 277–97.

Bradley, B. (2005). Virtue consequentialism. *Utilitas 17*(3), 282–298.

Brandt, R. (1963). Toward a credible from of utilitarianism. In H.-N. Castañeda and G. Nakhnikian (Eds.), *Morality and the Language of Conduct*, pp. 107–143. Detroit: Wayne State University Press.

Brandt, R. (1967). Some merits of one form of rule-utilitarianism. *University of Colorado Studies in Philosophy*, 39–65.

Brandt, R. B. (1959). *Ethical Theory*. Prentice-Hall.

Brandt, R. B. (1992). *Morality, Utilitarianism, and Rights*. Cambridge University Press.

Driver, J. (2001). *Uneasy Virtue*. Cambridge University Press.

Egan, A. (2007). Some counterexamples to causal decision theory. *The Philosophical Review 116*(1), 93–114.

Feldman, F. (2006, May). Actual utility, the objection from impracticality, and the move to expected utility. *Philosophical Studies 129*(1), 49–79.

Foot, P. (1985). Utilitarianism and the virtues. *Mind 94*(374), 196–209.

Gauthier, D. (1986). *Morals by Agreement*. Oxford University Press.

Gibbard, A. and W. L. Harper (1978). Counterfactuals and two kinds of expected utility. In C. A. Hooker, J. J. Leach, and E. F. McClennen (Eds.), *Foundations and Applications of Decision Theory*, pp. 125–62. D. Reidel.

Greene, P. (2018). Success-first decision theories. In A. Ahmed (Ed.), *Newcomb's Problem*. Cambridge University Press.

Gruzalski, B. (1981). Foreseeable consequence utilitarianism. *Australasian Journal of Philosophy 59*, 163–76.

Hare, R. M. (1954-1955). Universalisability. *Proceedings of the Aristotelian Society 55*, 295–312.

Hare, R. M. (1963). *Freedom and Reason*. Oxford University Press.

Hare, R. M. (1981). *Moral Thinking: Its Levels, Method, and Point*. Clarendon Press.

Harrison, J. (1952-1953). Utilitarianism, universalisation, and our duty to be just. *Proceedings of the Aristotelian Society 53*, 105–34.

Harrod, R. F. (1936, April). Utilitarianism revisited. *Mind 45*(178), 137–56.

Hofstadter, D. (2008). *Metamagical Themas: Questing for the Essence of Mind and Pattern*. Basic Books.

Hooker, B. (2000). *Ideal Code, Real World*. Oxford University Press.

Horgan, T. (1981). Counterfactuals and Newcomb's problem. *The Journal of Philosophy 78*(6), 331–356.

Horgan, T. (1985). Newcomb's problem: A stalemate. In R. Campbell and L. Sowden (Eds.), *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem*. University of British Columbia Press.

Jackson, F. (1991, April). Decision-theoretic consequentialism and the nearest and dearest objection. *Ethics 101*(3), 461–82.

Joyce, J. M. (1999). *The Foundations of Causal Decision Theory*. Cambridge University Press.

Joyce, J. M. (2012). Regret and instability in causal decision theory. *Synthese 187*(1), 123–45.

Levinstein, B. A. and N. Soares (2020, Forthcoming). Cheating death in Damascus. *The Journal of Philosophy*.

Mason, E. (2003). Consequentialism and the "ought implies can" principle. *American Philosophical Quarterly 40*, 319–31.

McClennen, E. F. (1990). *Rationality and Dynamic Choice*. Cambridge University Press.

Meacham, C. J. G. (2010). Binding and its consequences. *Philosophical Studies 149*(1), 49–71.

Miller, D. E. (2020). Moral education and rule consequentialism. *The Philosophical Quarterly doi: 10.1093/pq/pqaa023*.

Oddie, G. and P. Menzies (1992). An objectivist's guide to subjective value. *Ethics 102*, 512–33.

Potter, N. T. and M. Timmons (1985). Introduction. In N. T. Potter and M. Timmons (Eds.), *Morality and Universality*, pp. ix–xxxii. D. Reidel.

Ridge, M. (2006). Introducing variable-rate rule-utilitarianism. *The Philosophical Quarterly 56*, 242–53.

Sidgwick, H. (1884). *The Methods of Ethics* (Third ed.). London: Macmillan and Co.

Singer, M. G. (1955). Generalization in ethics. *Mind 64*(255), 361–75.

Singer, M. G. (1961). *Generalization in Ethics: An Essay in the Logic of Ethics, with the Rudiments of a System of Moral Philosophy*. Alfred A. Knopf.

Singer, M. G. (1985). Universalizability and the generalization principle. In N. T. Potter and M. Timmons (Eds.), *Morality and Universality*, pp. 47–74. D. Reidel.

Smart, J. (1956). Extreme and restricted utilitarianism. *The Philosophical Quarterly 6*(25), 344–354.

Smart, J. J. C. (1973). An outline of a system of ethics. In J. J. C. Smart and B. Williams (Eds.), *Utilitarianism: For and Against*. Cambridge University Press.

Smith, H. M. (2010). Measuring the consequences of rules. *Utilitas 22*, 413–33.

Soares, N. and E. Yudkowsky (2018). Functional decision theory: A new theory of instrumental rationality. *arXiv: 1710.05060*.

Thomson, J. (1997). The right and the good. *The Journal of Philosophy 94*(6), 273–98.

Timmons, M. (2002). *Moral Theory: An Introduction*. Rowman & Littlefield.

Yudkowsky, E. (2010). Timeless decision theory. Technical report, The Singularity Institute.

Zagzebski, L. (1996). *Virtues of the Mind: An Inquiry into the Nature of Virtue and the Ethical Foundations of Knowledge*. Cambridge University Press.

Zimmerman, M. J. (2005). The relevance of risk to wrongdoing. In R. F. Kris McDaniel, Jason R. Raibley and M. J. Zimmerman (Eds.), *The Good, the Right, Life And Death: Essays in Honor of Fred Feldman*. Ashgate.