Epistemic Normativity & Epistemic Autonomy: The True Belief Machine

Interviewer: How often do you engage in critical reflection?

CEO: I'm very busy, so I have my secretary critically reflect on my behalf once a week and type up a report.[1]

The veritic epistemic consequentialist argues that true belief is the only source of non-derivative epistemic value and it is to be promoted (see Goldman 1999; David 2005; Ahlstrom-Vij & Dunn 2014).[2] Reflection, inference, inquiry, etc. have derivative value just insofar as they promote true belief. If something else is equally conducive to true belief, it is of equal epistemic value. This is concerning. It entails that a totally passive subject receiving true beliefs through a mechanism analogous to Nozick's (1974) experience machine is just as well off, epistemically speaking, as one who reflects, infers, inquires and engages in characteristically human mental activity and forms the same beliefs. This result seems wrong, however. After setting out the problem in a bit more detail, I will respond to objections.

It seems clear that the subject who goes into the true belief machine ceases to be autonomous and this is where the problem lies. However, I will consider recent accounts of autonomy and argue that they provide no comfort for the veritic consequentialist. I will draw on some recent work from Adam Carter (2020) and Jonathan Matheson (2022) that helps us understand how the true belief machine interferes with intellectual autonomy. However, I will argue that neither approach provides a way to solve the problem that is in keeping with the order of explanation distinctive of veritic consequentialism.

[1] The joke is from Loader (2012).

[2] Epistemic utility theory (e.g. Joyce 1998, 2009, Leitgeb & Pettigrew 2010, Easwaran 2013, Easwaran & Fitelson 2015, Pettigrew 2013; 2016) is a form of epistemic consequentialism that focuses on credential states rather than full belief. They hold that accuracy (as determined by their preferred scoring system) is what is fundamentally valuable. Then Bayesian updating is derivatively valuable insofar as it is conducive to greater accuracy scores. I will focus on full belief in this paper, although my argument could be extended to deal with other states with a mind-to-world direction of fit.

In section (I) I discuss consequentialism generally. I will then briefly discuss the experience machine objection to hedonistic consequentialism and what I take the upshot of that to be. In section (II) I briefly discuss the core commitments of epistemic consequentialism and then lay out my parallel argument. I will spend most of this section anticipating objections and responding to them. In section (III) I consider recent work on intellectual autonomy. Here I argue that it may help us understand what is wrong with going into the machine, but it won't save veritic consequentialism.

(I) Consequentialism

Consequentialists maintain that the good is prior to the right. Goodness and badness are properties had by states of affairs. The consequentialist must first tell us which states of affairs are good and which bad non-derivatively. In the domain of ethics, common answers are pleasure, happiness, and desire-satisfaction. The contraries of these things (pain, unhappiness, desire-frustration, respectively) are non-derivatively bad. Suppose we pick pleasure as the only thing that is non-derivatively good. Then a state of affairs with a more favorable pleasure/pain ratio than another state of affairs is better than that second state of affairs.

Once the consequentialist has given us her account of the good, she then defines the right in terms of it. A maximizing consequentialist says that the action that produces a state of affairs at least as good as any state of affairs produced by any other action available to the agent at the time of acting is right. A satisficing consequentialist says that any action that produces a state of affairs that meets or exceeds a certain goodness-threshold is right. Going forward, I will focus on maximizing consequentialism. I do this for ease of exposition only, nothing substantive turns on this decision.

Let us turn to hedonistic ethical consequentialism. The hedonist says that pleasure is the only thing that is non-derivatively ethically good and pain the only thing that is non-derivatively ethically bad. Since the good is prior to the right, they must tell us what pleasure is before telling us which

actions are right. A natural way to go is to think of pleasure as a phenomenological state. This seems promising since phenomenological states are individuated by their intrinsic experiential properties rather than how they are brought about. So, we can first understand what pleasure is and why it is valuable and then use that account to explain why certain actions are good. This seems easy enough, since pleasure is naturally thought of as a kind of "buzz" feeling that could be produced in any number of ways. Similarly, pain is a kind of "anti-buzz" that could also be produced in any number of ways. Right actions, then, are those that produce a ratio of buzz-to-anti-buzz at least as favorable as the ratio produced by any alternative course of action.

Nozick's (1974) experience machine objection to hedonism can be reframed as an objection to hedonistic consequentialism. The objection is that it is possible, in principle at least, for life in an isolated, virtual world to result in a more favorable buzz-to-anti-buzz ratio than engaging in the kinds of activities that are intuitively worthwhile. For example, suppose a matrix-like machine could produce more of that buzz than living a characteristically human life without producing any anti-buzz. Each person connected to the machine receives neural stimulation that makes them think they are living an especially fulfilling life when in fact they are envatted. We can imagine a version of the machine where the experienced worlds of different people connected to the machine are unrelated and there is consequently no virtual community. What is important is that the machine produces a more favorable buzz-to-anti-buzz ratio than living a normal life.

It is a consequence of hedonistic consequentialism that the right thing to do would be to plug yourself into the machine. It might be objected that others could live better if the agent were to remain unplugged. Hedonistic consequentialism needn't be a form of egoism. Everyone's pleasure counts, not just that of the agent. That is true but it doesn't solve the problem. Rather, it just

requires that we plug everyone into the machine. This seems worse than just plugging one person in since now nobody is living a meaningful life.

There are several ways to respond to the argument.[3] Most people just reject hedonism and, by extension, hedonistic consequentialism.[4] We could also refine our account of pleasure, so that it isn't just a neural buzz.[5] We could identify the bearer of non-derivative value with something other than pleasure.[6] In that case, we are still consequentialists but no longer hedonists. Of course, we could bite the bullet.[7] Or we could maintain that even if going into the experience machine is wrong, we can't explain why except by reference to the value of pleasure.[8] Which way we go here doesn't matter for my purposes. My main concern is to show that there is an epistemic analogue and none of the obvious rejoinders address the problem.

(II) The True-Belief Machine

> **Veritic Epistemic Consequentialism:** The right action, epistemically speaking, is the one that results in at least as favorable a ratio of true to false belief as any other action available.

There are a number of ways this idea could be more fully developed, just as there are a number of ways of fleshing out the ethical analogue. There are maximizing and satisficing versions of the theory. Furthermore, we can consider both direct and indirect versions of the theory. Alvin Goldman (1986) uses an indirect version to give a theory of justification. Very briefly, he claims that

[3] Some involve saying that it confuses well-being with morality (e.g., Railton 1989; Goldsworthy 1992; Silverstein 2000). Even if this is true, I am considering it as an objection to hedonistic consequentialism as a moral theory, so they could agree with what I say here.
[4] At the very least, this is the story told in most introductory ethics textbooks. See Haber (1993:7), Sher (1996: 612), and Carson & Moser (1997: 7).
[5] See Donner (1991), Feldman (1997; 2002), and Heathwood (2006).
[6] See Brandt (1979).
[7] Neil Sinhababu told me in conversation this is his preferred response.
[8] Kawall (1999) makes a similar move defending mental-state theories of well-being from Nozick's thought experiment.

rather than evaluating beliefs directly, we evaluate them indirectly by first evaluating the truth-conduciveness of the belief-forming processes that produce them. The justificatory status of a belief is a function of the truth-conduciveness of the process that produced it.

However, veritic consequentialism has implications beyond the theory of justification. It is a theory about what is of fundamental epistemic value and how epistemic deontic statuses are explained in terms of that. The core of veritic consequentialism as such is that true belief is the only thing of non-derivative epistemic value, and it is to be promoted. Epistemically right acts are such because of how they promote true belief. Proponents of this order of explanation include David (2001), Ahlstrom-Vij (2013), Copp (2013), Goldman (2015), Ahlstrom-Vij & Dunn (2017) as well as many others working in epistemic utility theory (see fn. 4). It is this order of explanation generally that interests me here, rather than process reliabilism specifically.

Here is the analogous problem: suppose the matrix-like machine feeds you true beliefs rather than pleasure. That is, you no longer reflect on your epistemic credentials. You no longer inquire, and you no longer draw inferences. You have a better ratio of true to false beliefs (covering a large swath of areas of interest, let us suppose) than even the most scrupulous and dedicated human inquirer, but you are cognitively dormant. Your native cognitive faculties are replaced by a microchip that receives updates much in the same way computers and smartphones receive software updates. These updates cause the chip to receive information and induce belief in true propositions. You do not draw inferences from or critically reflect upon information received at all. The task of updating your belief system has been outsourced wholesale to the machine. You still move and act in the world. From a third-person perspective, the subject who goes in for the true belief machine is indistinguishable from one who does not. When I talk about plugging into the machine or going into

the machine, we can imagine that this just involves implanting a chip that is connected to a supercomputer elsewhere.

I submit that you are epistemically worse off than the normal human inquirer if you plug in. However, all my argument requires is that you are epistemically worse off than you would be if you had the same beliefs, but you formed them by reflecting, inquiring and inferring. If the veritic consequentialist is right, these actions are significant only insofar as they produce epistemically valuable consequences. If something else can produce those consequences, it is just as good epistemically. Intuitively, the true belief machine is not just as good epistemically. Just as it is ethically better to have a favorable pleasure-to-pain ratio achieved by living a normal human life rather than to have that same ratio produced by plugging in to the experience machine, it is epistemically better to have a favorable true-to-false belief ratio produced by normal cognitive activity rather than by plugging in to the true belief machine. This remains the case no matter how counterfactually robust we make the favorable true/false belief ratio produced by the machine. This is not to say that the value of reflection, inquiry and inference is completely independent of the value of true belief. It is just to deny that equally truth-conducive activities are always of equal epistemic value.

That is the problem I want to raise. I will consider 4 defensive maneuvers apparently available to the veritic epistemic consequentialist and show that appearances are deceiving: some don't work, and some are not even maneuvers against the claim I am making.

Defensive Maneuver #1

Someone could object that the subject in my thought experiment doesn't even have beliefs because they don't engage in inference, reflection and inquiry. The functional role of belief involves figuring in inferences. An information-bearing state that a subject is not disposed to draw inferences

from is not a belief.[9] This is not to say that every belief must be formed via inference. It is just to say that figuring in inferences is part of the functional role of beliefs as such. Perhaps some beliefs never figure in inference as either premise or conclusion, but part of what makes them beliefs is their potential to figure into inferences in these ways. However, it could be argued that the subject who plugs into the machine no longer draws inferences and, consequently, none of her mental states have this potential.[10] So, the thought experiment involves false advertising. The machine doesn't give you beliefs, but rather some other information bearing state with a different functional role.

One worry is that the subject might have the relevant dispositions to draw inferences without actually drawing the inferences. Compare: I might be disposed to get myself a glass of water when I'm thirsty if nobody brings me one, but nonetheless be fortunate enough for someone to bring me one every time I get thirsty. Similarly, I might be disposed to draw inferences from the information uploaded by the machine if the machine doesn't also upload the inferential consequences of that information. However, the machine does in fact upload those inferential consequences. So, just as in the case of the glass of water, I remain idle.

A deeper worry is that the epistemic significance of inference (for example) seems to run deeper than the veritic consequentialist says it does, if we take this line of thought seriously. According to the line of thought that gives rise to this worry, inference is part of what makes belief the kind of thing it is. It is not just a useful way of bringing about true beliefs but rather partially constitutive of belief itself. However, the consequentialist order of explanation involves explaining the epistemic significance of activities and processes such as inference in terms of a prior understanding of the states of affairs they bring about.

[9] Thanks to Kathryn Pogin for raising this point.
[10] Thanks to an anonymous referee for pointing out the need to be clear about this.

Here is another way to see the point. Assume for the sake of argument that the subject in my thought experiment does not have beliefs because the functional profile of beliefs involves inference and my subject is not disposed to draw inferences. Even if this is right, the subject still has an information bearing state like belief except with a slightly different functional profile. The information bearing state still regulates action, for instance. It just doesn't involve inference. So, we can call the information state "belief*". Why are beliefs more epistemically valuable than belief*s? Belief*s can represent the world accurately and it is the value of accuracy in general that presumably motivates the consequentialist order of explanation. If belief*s are just as accurate as beliefs, why are beliefs better? A consequentialist would, it seems, have to say that inferring is better than not inferring. What reason could they offer? They could say that inferences are good for the sake of their consequences. That won't help here though because belief*s have even better consequences with respect to accuracy if we go into the machine. A tempting answer here is that inference is partially constitutive of the kind of cognitive agency we enjoy and is for this reason part of the functional role of one of our information bearing cognitive states. If this is right though, then the ultimate account of why inference matters is not because inferences are conducive to good cognitive consequences but rather because they are partially constitutive of the kind of mental life we have. Something similar applies, I think, to reflection and inquiry, though I won't belabor the point here.

Defensive Maneuver #2

Veritic epistemic consequentialists might concede that the subject is worse-off in some way but deny that they are worse off epistemically. In epistemology we care about the pursuit of truth and the subject is doing well with respect to that. So, the problem must lie elsewhere, they might argue. When you plug into the machine you are worse-off eudaimonistically perhaps, though not

epistemically.[11] If we go this way, then the set-up of the problem involves a misdiagnosis. By mistaking a non-epistemic problem for an epistemic problem, we mistake a problem that ought to be solved by ethicists for a problem that ought to be solved by the theory of epistemic normativity.

In response, we should note that consequentialism of any kind is a substantive theory about normativity in the domain to which it applies. The hedonistic ethical consequentialist, for instance, is offering a theory about ethical normativity. They aim to give a theory that will account for a weighted most of our considered judgments about ethical matters. This means we first need some way to fix the reference of "ethical" without begging the question either in favor of or against that theory or any plausible rival. Once we've done that, we test the theory by applying it to cases within the domain on which we've settled. The hedonistic consequentialist claims that we can either directly or indirectly explain the value of everything in that domain in terms of the value of pleasure. That claim is substantive. The hedonistic consequentialist is not just stipulating that everything that can't be accounted for by her theory is beyond the purview of ethics. This would make hedonistic consequentialism uninteresting.

Similarly, the veritic epistemic consequentialist offers a substantive theory of epistemic normativity. So, first we fix the reference of "epistemic". The obvious way of doing so is to say that epistemology is the study of knowledge. The etymology of "epistemology" suggests something along these lines, though the result is a bit too narrow. The theory of knowledge is an important part of epistemology but not the whole thing. The same can be said for the theory of justification. When fixing the reference, we want to do so in a way that includes the contributions to epistemology made by ancient and early modern philosophers, even though they weren't pursuing quite the same

[11] Thanks to Sandy Goldberg for pressing me on this point.

questions we are today. In the *Stanford Encyclopedia of Philosophy*, Ram Neta and Matthias Steup offer the following,

> The term "epistemology" comes from the Greek words "episteme" and "logos". "Episteme" can be translated as "knowledge" or "understanding" or "acquaintance", while "logos" can be translated as "account" or "argument" or "reason". Just as each of these different translations captures some facet of the meaning of these Greek terms, so too does each translation capture a different facet of epistemology itself. Although the term "epistemology" is no more than a couple of centuries old, the field of epistemology is at least as old as any in philosophy. In different parts of its extensive history, different facets of epistemology have attracted attention. Plato's epistemology was an attempt to understand what it was to know, and how knowledge (unlike mere true opinion) is good for the knower. Locke's epistemology was an attempt to understand the operations of human understanding, Kant's epistemology was an attempt to understand the conditions of the possibility of human understanding, and Russell's epistemology was an attempt to understand how modern science could be justified by appeal to sensory experience. Much recent work in formal epistemology is an attempt to understand how our degrees of confidence are rationally constrained by our evidence, and much recent work in feminist epistemology is an attempt to understand the ways in which interests affect our evidence and rational constraints more generally. In all these cases, epistemology seeks to understand one or another kind of *cognitive success* (or, correspondingly, cognitive *failure*). (Steup & Neta 2020, their italics)

According to their definition, epistemology is the study of cognitive success and failure. If their definition is right, then one is epistemically better off just in case one is cognitively more

successful and epistemically worse off just in case cognitively less successful. One virtue of their definition for our purposes is that it was not designed with the intention of saving any particular theory of epistemic normativity. Another is that, as they show, it covers paradigmatic epistemological theories from a variety of subdisciplines and historical periods.

We can now return to the claim that epistemology is about truth, so the subject of my thought experiment is not epistemically worse-off. One way of interpreting that claim is as a way of fixing the reference of "epistemology." Cognition is truth-directed. Perhaps they are just trying to fix the reference in roughly the same way I have. That is, they are interested in certain kinds of truth-directed process, practice, or method. If this is what the response comes to, then it is fine so far as it goes but it doesn't go far enough to get the consequentialist out of trouble. Cognition might be truth-directed, but it doesn't follow that anything that gets us truth is successful cognition. Truth-directed processes might need to satisfy further constraints to be fully successful instances of cognition. We might worry that the subject in the true belief machine is missing out on something other than truth that full cognitive success requires, perhaps virtue of some sort (more on this in section (III).

Furthermore, it isn't clear that the subject of the thought experiment is really cognizing, despite believing truly. The Oxford dictionary says that cognition is "the *mental action or process* of acquiring knowledge and understanding through thought, experience and the senses" (my italics). We can replace knowledge and understanding with truth here to make the definition more veritist-friendly. Still, the subject doesn't seem sufficiently active to really be engaging in cognition. Her case is at best a borderline case of cognition.

To be clear, the bar for genuine cognition is not very high.[12] Ordinary perception is genuine cognition and all you have to do is open your eyes. Something similar applies to receiving testimony. However, in these cases the subject's faculties are not dormant. Rather, their employment of those faculties is unremarkable but nonetheless present. It is genuine cognitive activity even if the level of activity is low relative to that of other less mundane cognitive activity.

The upshot so far is that epistemology is the study of cognitive success/failure and the subject of my thought experiment is less cognitively successful than a normal human inquirer despite believing truly.

Of course, the consequentialist reply might also be interpreted as giving the meaning of "epistemology," which goes beyond fixing the reference. If so, then it does rule out my counterexample, but only by stipulation. That is, they might be saying that, by definition, epistemology is the study of truth-conductivity as such and any two processes that are equally truth-conducive are epistemically alike. If that's what they mean, then they aren't offering an interesting and substantive theory of epistemic normativity but rather giving a definition of "epistemic" that guarantees the impossibility of counterexample.

Defensive Maneuver #3

Someone might respond by trying to show that their theory of justification does not entail that the subject in the machine has justified beliefs. If one distinguishes propositional and doxastic justification, the objection could be about either. The objection will have more force if we have doxastic justification in mind. The machine could make adequate justification available to the

[12] Thanks to Baron Reed for pressing me on this point.

subject, the problem is that the subject isn't basing her beliefs on that justification. The beliefs are caused by the chip instead.

My argument in this paper is not an argument against reliabilism or any theory of epistemic justification for that matter. Similar arguments have been made by BonJour (1980) and Reed (2016) against the view that a sufficient condition for a belief to be justified is that it be the output of a reliable belief forming process.

I am interested in epistemic normativity here rather than the theory of justification. It is sometimes thought that the argument for reliabilism depends on epistemic consequentialism.[13] This claim has recently been disputed by Sylvan (2020b).[14] It doesn't matter for my purposes whether he is right, as I am interested in veritic epistemic consequentialism as such. If you are a reliabilist and a veritic consequentialist, your theory of justification might rule that the subject in my example doesn't have justification. After BonJour, reliabilists don't always think of reliability as a sufficient condition for justification. Rather, they sometimes impose further constraints.[15] The question still arises why the subject, despite lacking justification, is any worse off epistemically than the justified believer. Why the fuss about justification? By the lights of veritic epistemic consequentialism, it seems they are better off than the justified believer if they have more true beliefs, regardless of their justificatory status. Justification only matters insofar as it is instrumental to true belief, according to the veritic consequentialist. So, I concede that the subject in my thought experiment lacks (at least) doxastic justification. However, the veritic consequentialist is unable to explain why she is worse off epistemically because of this.

[13] Goldman (1986: 97) gives this impression at times by comparing his view to rule consequentialism. Cf. Percival (2002: 121), Chase (2004: 124), Berker (2013b: 350).
[14] Cf. Ahlstrom-Vij & Dunn (2017) for a carefully qualified answer.
[15] For example, the subject can't have defeaters for the beliefs they produce for those beliefs to be justified (Goldman 1979).

Here is another way to see the point. The problem I raise here is similar to, but not the same as, the "swamping" problem from Zagzebski (2003).[16] If justified beliefs are reliably produced beliefs, then it isn't clear why justified true beliefs are more valuable than true beliefs that were not reliably produced. If I have an unreliable espresso maker, the cups of espresso it makes when it works properly are no less valuable than the cups of espresso a reliable espresso machine produces (Zagzebski 2003: 13). Similarly, if true belief is what matters most fundamentally, the ones formed through capricious processes seem to be just as good as the ones formed by reliable mechanisms. The espresso tastes the same either way. For my purposes, it is unimportant whether this is indeed a serious problem. What matters is that it is similar to the problem I just raised but the attempts to solve it, even if successful, can't be used to address my problem.

One response to the problem just raised for reliabilism is that a state of affairs in which you have a justified belief is more valuable than a state of affairs in which you don't because having a reliably formed belief raises the conditional probability that other beliefs of yours are true (Goldman & Olsson 2009: 28). It isn't clear this addresses the problem Zagzebski raises, since it doesn't show that the justified beliefs themselves are more valuable than mere true beliefs, but rather that you are more likely to have valuable things if you have a justified belief. Even if this concern can be addressed and the response solves the problem Zagzebski and others raise, it doesn't address the problem I raise. The probability of any given belief of yours being true conditional on you entering the true-belief machine is as high as we stipulate it to be.

Goldman & Olsson also offer another response. They claim that a process that tends to produce independently valuable results might in some cases inherit some of the value of those results, even in cases where it does not produce them. They call this process "value autonomization" (Goldman &

[16] Cf. Jones (1997: 426), Riggs (2002), Kvanvig (2003) to whom we owe the term "swamping", Zagzebski (2003).

Olsson 2009: 33) and offer the following example. Good intentions are good because they tend to produce good results. However, good intentions still have value even when they fail to produce good results. This, they claim, is because they are generally instrumental to a good outcome. As a result, they inherit some value from those outcomes and maintain it autonomously even in cases where the good outcome doesn't obtain.

It isn't clear to me that they have made value-autonomization any less mysterious with their example, which bears much of the burden of their argument. Even if I am wrong about this, it still isn't clear that the resulting account is truly consequentialist. Rather, they seem to be saying that "in the beginning" consequentialism was true but now it isn't because certain means have become autonomous sources of value.

The most serious concern, however, is that they give us no account of why the activity of receiving updates from the true-belief machine can't autonomize value. If means autonomize value *because* they are conducive to certain ends, then the true-belief machine could autonomize even more value than any actual belief-forming processes, so long as we stipulate it is more accurate than characteristically human mental activity. Of course, Goldman & Olsson weren't trying to address the concern I raise here, so this isn't a failure on their part. I am only trying to show here that it isn't clear how we can take a consequentialist-friendly solution to Zagzebski's problem and apply it to the similar problem I raise.[17] Even if it can be shown that the subject in the true-belief machine has no justification, I am asking why justified belief (or knowledge for that matter) is more epistemically valuable than whatever the subject I describe has. Veritic epistemic consequentialism makes this question hard to answer since it requires that justification is only valuable insofar as it is conducive

[17] Ahlstrom-Vij (2013) argues that the swamping problem isn't really a problem at all for veritists. For my purposes here, that is fine. I am offering a different but similar problem.

to true belief. My point is that the well-known responses to the related "swamping" problem, even if effective as a response to that problem, won't help here.

Similarly, none of the responses to Berker's "separateness of propositions" problem for consequentialism and process reliabilism will help the consequentialist deal with the issue raised in this paper. Utilitarianism is a form of consequentialism. According to the utilitarian, the right course of action is the one that produces the best ratio of aggregate positive utility to aggregate negative utility. Since the ratio is determined by pure aggregation, it doesn't matter how the utility is distributed over individuals. This leads to the possibility of "organ harvest" cases. Consider a doctor who has a patient coming in for a routine check-up (cf. Thomson 1976: 206). Let us suppose the patient is healthy and she is a nuisance to everyone she interacts with, so much so that she has a net negative impact on aggregate utility. The doctor could drug her, harvest her organs, and use those organs to save several good Samaritans at the local hospital who will die unless they receive healthy organs soon. If the doctor were to do this, the result would be a more favorable ratio of aggregate positive to negative utility. However, it nonetheless seems wrong. Simple versions of utilitarianism get the wrong answer here, arguably, because they ignore the separateness of persons by only paying attention to aggregate utility. As Berker puts it,

> The idea here is that teleological theories such as act-utilitarianism do not treat intrapersonal trade-offs—trade-offs that involve harming or hurting a given person in one way in order to benefit or advantage that same person in another way—as fundamentally any different from interpersonal trade-offs when determining what an agent should do, but whereas intrapersonal trade-offs are morally defensible when the benefit to the one person outweighs the harm, interpersonal trade-offs are not morally defensible in the same way, or at least not always morally defensible in the same way. (Berker 2013a: 358).

He goes on to present a series of cases meant to establish that epistemic consequentialism faces a similar problem: the separateness of propositions. The reason is that considerations of aggregate epistemic utility will favor "cutting up the one to save the many" in epistemology and this will be even more obviously bad in the epistemic case than the ethical case. For instance, if one were to believe against the evidence now to secure a large number of true beliefs later, that would be epistemically amiss despite securing a favorable ratio of epistemic positive to negative utility. Berker argues that, because process reliabilist theories of epistemic justification are versions of consequentialism, they face the epistemic version of the separateness of persons problem faced by ethical consequentialists.

Consequentialists typically respond to Berker by claiming, in one way or another, that he is overstating the similarities between process reliabilism and rule utilitarianism (e.g., Goldman 2015, Ahlstrom-Vij & Dunn 2017). Even if this is true[18], it doesn't matter for my purposes. As I said above, I am not interested in the theory of justification here. I'm interested in epistemic normativity as such. It doesn't matter to me whether the process reliabilist thinks the true-belief machine dweller has justified beliefs. Even if process reliabilism is quite a bit different than rule utilitarianism, veritic epistemic consequentialism is a lot like hedonistic consequentialism.

I will here go a step further and argue that no version of rule consequentialism will help here, even if it has to do with epistemic normativity rather than the theory of justification. It might be thought that just as rule consequentialism in ethics helps us avoid counter-intuitive consequences of act consequentialism, the same could be made to work in the epistemic realm. It has been argued that accepting certain rules will have a sufficiently positive result even if conforming to those rules will sometimes fail to maximize the good (Hooker 2007). For instance, it might be that even though

[18] For an argument that it is not, see Berker (2015). It doesn't matter for my purposes if he is right.

organ harvests maximize utility, accepting a rule that permits organ harvests in cases where utility is maximized would have an overall negative effect. It would result in general fear of routine check-ups and that negative utility has a greater absolute value than the positive utility of the occasional harvest. What is the epistemic analogue of this? Why would it be epistemically worse to accept a rule that allows us to go into the true belief machine when the opportunity presents itself? It is not clear that accepting that rule incurs any epistemic costs at all.

Defensive Maneuver #4

One of the reasons I gave for thinking the subject is worse off than a normal agent with just as many true beliefs is that the latter reflects and the former doesn't. The philosophical significance of reflection is often taken for granted, but it has been called into question in recent years; notably by Kornblith (2012; 2017). However, Kornblith's main line of attack in his campaign against reflection will be of no use to the epistemic consequentialist here because it presupposes consequentialism. He argues, for example, that non-reflective processes are also quite truth-conducive and sometimes even more so than reflection in certain domains.[19] The argument works, perhaps, so long as we take for granted at the outset that two options are epistemically on a par just in case they are equally truth-conducive. But that is exactly what I am inviting the reader to reconsider. So, it would be viciously circular to make an argument that presupposes it here.

(III) Epistemic Autonomy

In this section I will consider the possibility that lack of autonomy is the problem with going into the true belief machine. Two ways of spelling this out will be considered: one involving autonomy understood as an achievement, the other involving autonomy understood as a character

[19] Cf. Proust (2013).

virtue. It will be shown that even if one or both are tenable, they are of no help to the veritic consequentialist.

First, I will consider autonomy as an achievement. In the success-from-ability tradition, knowledge is true belief achieved by the agent.[20] Proponents of this order of explanation sometimes claim that true beliefs need to be produced by processes that are suitably integrated into the subject's cognitive architecture for them (the true beliefs that is) to be an achievement creditable to the agent.[21] Carter (2022) goes further and argues that the processes don't just need to be integrated, the integration itself must be creditable to the agent for the beliefs produced by the ability to be autonomously held. Only autonomously held beliefs are candidates for knowledge. So, we are thinking of autonomy here as a property had by beliefs in virtue of how the processes that produce them are cognitively integrated by the agent.

Let us consider what happens when we apply this account to cognitive enhancement. By considering kinds of cognitive enhancement that already exist (e.g., Ritalin and neural implants) Carter (2020) argues persuasively that some pose a threat to autonomy and others don't. The difference is that some cognitive enhancements can be suitably integrated by the subject into their own cognitive architecture whereas others cannot. He points out that the problem with older versions of process reliabilism is that they are susceptible to counterexamples such as Lehrer's (1990) Tru-Temp case. In this case, a subject has a reliable thermometer implanted in their brain that causes them to form true beliefs about the temperature, although they don't know the thermometer is there. The beliefs seem to assail them from out of nowhere because they are not properly integrated into the subject's cognitive architecture.

[20] See Sosa (1991;2007), Greco (2003; 2010; 2013), Pritchard (2010; 2012); Turri (2011); Carter (2014).
[21] See Pritchard (2010), Greco (2010).

Carter's point can help us see what has gone wrong in the case of the true belief machine. Piecemeal outsourcing is compatible with autonomy because the result can, in the right circumstances, be integrated into the subject's cognitive architecture. We can imagine a variation of Tru-Temp where the subject knows that the thermometer is there and that it is reliable. In this case, they can understand where the beliefs are coming from, endorse the belief-forming process and integrate its products into their native cognitive architecture. In doing so they achieve epistemic self-regulation. In the true belief machine, the subject engages in wholesale outsourcing. There is nothing left of their cognitive architecture with which to integrate incoming beliefs. Their native faculties are dormant. They are consequently no longer autonomous.

However, this provides no comfort to the veritic consequentialist. We see this when we ask why autonomy matters, epistemically speaking. Carter says the following, "the acquisition of knowledge, true belief and understanding are epistemic goods and aspects of character…are typically explained as good to have… in virtue of their connection to such goods" (Carter 2020: 2939). How can the veritic consequentialist fill in the details? Carter mentions a number of epistemic goods here but the veritic consequentialist will have to start with just true belief since, according to her, that is the only one with non-derivative epistemic value. The value of intellectual autonomy is then explained in terms of its connection to true belief. The relevant connection is consequence. So, the derivative epistemic value of autonomy is determined by the non-derivative epistemic value of the true beliefs that result from it. The problem is that the machine can produce whatever true beliefs one can acquire through autonomous inquiry, or more for that matter. There would still be something epistemically amiss about plugging in; something unexplained by veritic consequentialism.

One could avoid this problem by identifying other epistemic goods of non-derivative epistemic value. Carter mentions knowledge and understanding. One could also add achievement to

the list perhaps. Arguably, these can't be produced by the machine. However, this is to abandon veritism.

We could also try to shed light on the issue by thinking of intellectual autonomy as a character virtue, like prudence or temperance. Jason Baehr defines an epistemic character virtue as, "a character trait that contributes to its possessor's personal intellectual worth on account of its involving a positive psychological orientation toward epistemic goods" (2011: 102). On this way of thinking, autonomy is a property of agents as a whole rather than their doxastic states individually. I will treat Jonathan Matheson (2022) as the representative of a family of views according to which the character virtue of autonomy involves cognitive, motivational and behavioral dispositions (cf. King 2020 & Battaly 2022). So far as I can tell, what I say about his view will be applicable to others in the family, despite there being differences in how they manage the details. According to Matheson (2022: 183) autonomy involves dispositions,

1) to make good judgments about how, and when, to rely on your own thinking as well as how, and when, to rely on the thinking of others.

2) to conduct inquiry in line with the judgments in (1), and

3) to do so because one loves the truth and appropriately cares about epistemic goods.

As Nathan King puts it, "Autonomy requires thinking *for* ourselves, but not *by* ourselves" (King 2020: 88).[22] Sometimes the autonomous person will defer to others because that is the most reasonable thing to do. It is no lapse of autonomy when I defer to medical experts about medicine,

[22] Thi Nguyen (2019) argues that a certain kind of autonomy, "delegational autonomy", involves the prudent outsourcing of some of our intellectual projects.

for example. So long as I outsource wisely, I might still do so autonomously. This is the reason Matheson includes (1).

Does the subject who enters the true belief machine have the virtue of autonomy? That depends on which epistemic goods are included in (3). Let us consider what happens when we only include true beliefs. (1) requires that the subject make good judgments about when to rely on her own thinking and when to rely on the thinking of others. If the acquisition of true belief is the only epistemic good we are after, then deciding in favor of plugging into the machine is an exercise of good judgment. The machine is more reliable at producing true belief and avoiding false belief than even the most dedicated and scrupulous human inquiry.

(2) requires that we conduct inquiry in line with that judgment. One way of reading (2) is that it can only be satisfied by someone who has engaged in a modicum of inquiry. That is, we could interpret (2) as saying that the subject must engage in at least some inquiry; how much is determined by the judgments in (1). If this is how we read (2), then the subject can't satisfy it by plugging in because machine dwellers don't inquire. However, the veritist is not out of the woods just yet. Even if the autonomous subject can't enter the machine, because doing so precludes the satisfaction of (2), the veritist is nonetheless committed to the value of autonomy being entirely derivative. According to the veritic consequentialist, autonomy earns its keep insofar as it is conducive to the production of true beliefs and the avoidance of false ones. However, the subject who plugs in does better in this respect than the autonomous subject, since the machine is stipulated to be better at producing true belief and avoiding false belief than any human unaided by the machine. So, the veritist can't explain the problem with entering the machine in terms of the autonomy one forfeits by doing so.

Alternatively, we could understand (2) as allowing for zero inquiry. Perhaps if the judgments in (1) have it that one should never rely on one's own thinking and only rely on the machine instead, then one needn't inquire at all. If we read (2) this way, then the subject could satisfy (2) despite entering the machine. Finally, (3) is easily satisfied by the subject who enters the machine if the only relevant epistemic good is true belief. She might enter because she loves the truth and continue to love it after plugging in.

The veritic consequentialist might recognize epistemic goods other than true belief. However, since they claim their value is entirely determined by how conducive they are to having true beliefs, the veritic consequentialist must recommend sacrificing these epistemic goods for a greater epistemic gain when one could get more true beliefs in some other way. The upshot is that the autonomy-as-character-virtue account might also be able to explain what is wrong with going into the machine, but not in a way that aids the veritist. If we include epistemic goods in (3) the epistemic value of which is not to be explained in terms of their conducing to true belief, then the absolute value of the goods sacrificed by entering the machine might be greater than the absolute value of the goods secured by entry.

It could be objected that the plugged-in subject, despite having a great many true beliefs and no false ones is nonetheless at the mercy of the machine because of her passivity.[23]This makes her worse off in one important respect than the autonomous subject who can pursue new information at her own discretion. She might very much like to know whether chocolate is toxic to cats, for example. If the machine doesn't upload a belief about this, she is out of luck. So long as the machine doesn't make her omniscient, it will always be possible that the subject finds herself in a situation

[23] Thanks to an anonymous referee for pressing me on this point.

where she wants to know something a normal inquirer could easily learn but can't because the machine doesn't upload any information on the topic.

What would the veritic consequentialist have to say about this possibility? As a consequentialist, she would have to view it as a trade-off and decide whether it is worth it by doing a cost-benefit analysis. The cost-benefit analysis, assuming veritism, tells in favor of it. The trade-off might inconvenience her, but it seems like a good trade-off from a purely epistemic point of view. This is especially clear if the only non-derivative epistemic value is true belief. Even if we weight some true beliefs differently than others and practical interests affect the weighting[24], on any plausible weighting it will be possible to compensate for the blind spots by giving the subject sufficiently many true beliefs about other important matters.

Of course, the subject not only lacks certain true beliefs, she also lacks the ability to inquire at her own discretion. However, the value of that ability is wholly derivative according to the veritic consequentialist. The value of the ability to inquire at your own discretion is wholly parasitic on the value of the true beliefs you secure by exercising this ability, on that view. So, the machine outperforms the ability by the veritist's criteria. Put another way, the inability itself has no negative epistemic utility. The negative epistemic utility caused by the inability is outweighed by the positive epistemic utility caused by the machine.

Furthermore, we could adjust the thought experiment so that the machine is sensitive to the subject's practical interests. My phone is sensitive to my habits and, without any prompting from me, provides information about traffic and the weather before I leave the house. Similarly, the

[24] Here I am assuming this is consistent with veritism just to imagine a best-case scenario for the veritist.

machine could exhibit sensitivity to the subject's interests and provide relevant information without any prompting from her.

(IV) Conclusion

I've shown that veritic epistemic consequentialism faces a version of the problem faced by hedonistic ethical consequentialism. It seems clear that much of the problem has to do with the subject's lack of autonomy. However, the veritic consequentialist has trouble explaining why the resulting lack of autonomy is epistemically problematic, rather than a prudent trade-off.

There are many ways we can go from here. To conclude, I will briefly consider three of them. We could reject veritism about epistemic value. We might opt for pluralism instead. Some epistemic goods, such as autonomy or understanding, arguably require activity on the part of the subject.[25] If so, then perhaps we can explain what has gone wrong with entering the true belief machine in terms of epistemic values that are necessarily forfeited by a cognitively dormant subject. Assuming that true belief is one of the values countenanced by the pluralist, she will need to explain why the understanding sacrificed by plugging in isn't outweighed by the value of the true beliefs the subject acquires by doing so.

Another possibility is that we hold on to veritism but reject consequentialism. It may well be that truth is the only thing with non-derivative epistemic value, but that the proper response to it is respect rather than promotion. Some values, arguably, call for a response other than promotion. For instance, humanity is valuable. However, it is far from clear that this means that suitably appreciating the value of humanity requires us to produce more humans (cf. Narveson 1976). Perhaps the proper response to the value of humanity is respect. In one sense of "respect", you respect something if

[25] See Hills (2016) for a defense of this claim about understanding.

that thing suitably constrains your deliberations (cf. Darwall 1977). Perhaps veritism is true, but the value of truth calls for respect (in this sense) rather than promotion. Kurt Sylvan (2020a) has developed a version of veritist non-consequentialism along these lines. Since the subject who plugs into the true belief machine doesn't deliberate at all, her deliberations are not suitably constrained by the value of truth. So, she does not respect the truth. Perhaps this is what is wrong with plugging in.

A final possibility is that we reject the value-first order of explanation shared by pluralism and veritic non-conseuentialism. Ross (1930) took an approach along these lines in the ethical domain. He identified several *prima facie* duties but declined to derive them from a prior account of ethical value. Berker (2013 a,b) appears to be sympathetic to the "duty first" order of explanation in epistemology. The problem animating this paper was that not all ways of promoting the value of true belief are equally epistemically valuable. If epistemic normativity is primarily a matter of discharging one's epistemic duties, rather than promoting epistemic value, then perhaps the problem is shut down at its source.

Works Cited:

Ahlstrom-Vij, K. (2013). In defense of veritistic value monism. *Pacific Philosophical Quarterly,* 94: 19-40.

Ahlstrom-Vij, K. & Dunn, J. (2014). A defence of epistemic consequentialism. *Pacific Philosophical Quarterly* 64: 541-51.

Ahlstrom-Vij, K. & Dunn, J. (2017). Is reliabilism a form of consequentialism?". *American Philosophical Quarterly,* 54(2):183-94.

Baehr, J. (2011). The inquiring mind. New York: Oxford University Press.

Battaly, H. (2022). Intellectual autonomy and intellectual interdependence. In Matheson, J. & Lougheed, K. (eds.) *Routledge Studies in Epistemology: Epistemic Autonomy.* N.Y: Routledge, pp. 153-73.

Berker, S. (2013a). Epistemic teleology and the separateness of propositions". *The Philosophical Review:* 337-93.

Berker, S. (2013b). The rejection of epistemic consequentialism. *Philosophical Issues* 23: 363-87.

Berker, S. (2015). Reply to Goldman: cutting up the one to save the five in epistemology". *Episteme*, 12: 145-53.

BonJour, L. (1980). Externalist theories of empirical knowledge". *Midwest Studies in Philosophy* 5:53-74.

Brandt, R. (1979). *A theory of the right and the good.* Oxford: Clarendon Press.

Carson, T. & Moser, P. (eds.) (1997). *Morality and the good life.* Oxford University Press: N.Y.

Carter, J. (2020). Intellectual autonomy, epistemic dependence and cognitive enhancement. *Synthese* 197: 2937-2961.

Carter, J. (2022). Epistemic autonomy and externalism. In Matheson, J. & Lougheed, K. (eds.) *Routledge Studies in Epistemology: Epistemic Autonomy.* N.Y: Routledge, pp. 21-41

Chase, J. (2004). Indicator reliabilism". *Philosophy and Phenomenological Research.* 69(1):115-37.

Copp, D. (2013). Indirect epistemic teleology explained and defended. in Fairweather, A. & Flanagan, O., ed., *Naturalizing epistemic virtue.* Cambridge: Cambridge University Press. pp. 70-91.

Darwall, Stephen. 1977. Two kinds of respect. *Ethics* 88: 36–49.

David, Marian. (2005). Truth as the primary epistemic goal: a working hypothesis. In Sosa, E. and Steup, M. *Contemporary Debates in Epistemology*, Wiley-Blackwell Publishing: Malden, MA. 296-312.

Donner, W. (1991). *The liberal self: John Stuart Mill's moral and political philosophy*. Cornell University Press: Ithaca, N.Y.

Easwaran, K. (2013). Expected accuracy supports conditionalization- and conglomerability and reflection". *Philosophy of Science* 80: 119-42.

Easwaran, K. & Fitelson, B. (2015). Accuracy, coherence and evidence. In Gendler, T. & Hawthorne, J. (eds.), *Oxford Studies in Epistemology, Volume 5*. Oxford: Oxford University Press: 61-96.

Elgin, C. (2021). The realm of epistemic ends. In *Epistemic autonomy*, eds. Matheson, J. & Lougheed, K. Routledge.

Feldman, F. (1997). *Utilitarianism, hedonism & desert*: *essays in moral philosophy*. Cambridge University Press: Cambridge, U.K.

Feldman, F. (2002). The good life: A defense of attitudinal hedonism. *Philosophy and Phenomenological Research*: 65(3). Pp. 604-28.

Goldman, A. (1979). What is justified belief? In Pappas, G. (ed.), *Justification and knowledge: new studies in epistemology*, Dordrecht: Reidel, pp. 1–25.

Goldman, A. (1986). *Epistemology and cognition.* Cambridge, MA: Harvard University Press.

Goldman, A. (1999). *Knowledge in a social world.* Oxford University Press: Oxford, UK.

Goldman, A. & Olsson, E. (2009). Reliabilism & the value of knowledge. in Haddock, A., Millar, A. & Pritchard, D. (eds.), *Epistemic value.* Oxford: Oxford University Press. pp. 19—41.

Goldman, A. (2015). Reliabilism, veritism, and epistemic consequentialism. *Episteme* 12: 131-43.

Haber, J.G. (1993). Introduction. In Haber, J.G. (ed.) *Doing and being: selected essays in moral philosophy*. New York: MacMillan.

Hills, A. (2016). Understanding Why. *Noûs* 49 (2):661-688.

Heathwood, C. (2006). Desire-satisfactionism & hedonism. *Philosophical Studies*: 133. Pp. 539-63.

Hooker, B. (2007). Rule-consequentialism and internal consistency: a reply to card. *Utilitas*, 19: 514–19.

Jones, W. (1997). Why do we value knowledge? *American Philosophical Quarterly,* 34.

Joyce, J. (1998). A non-pragmatic vindication of probabilism. *Philosophy of Science* 65: 575-603.

Joyce, J. (2009). Accuracy and coherence: prospects for an alethic epistemology of belief. In Huber, F. and Schmidt-Petri, C. (eds.) *Degrees of belief*. Dordrecht: Springer, 263-300.

Kawall, J. (1999). The experience machine and mental state theories of well-being". *Journal of Value Inquiry: 33(3):* 381-87.

King, N. (2020). *The excellent mind: intellectual virtue for the everyday life.* Oxford: Oxford University Press.

Kornblith, H. (2012*). On reflection*. Oxford: Oxford University Press

Kornblith, H. (2017). How central are judgment and agency to epistemology? *Philosophical Studies* 174: 2585-2697.

Kvanvig, J. (2003). *The value of knowledge and the pursuit of understanding.* Cambridge University Press: Cambridge, MA.

Lehrer, K. (1990). *Theory of knowledge*. London: Routledge.

Leitgeb, H. & Pettigrew, R. (2010). An objective justification of bayesianism I: measuring inaccuracy" *Philosophy of Science* 77: 236-72.

Loader, Paul (2013). Existential Times. in *The Onion and Philosophy* Ed. Sharon Kaye. Chicago: Carus Publishing Company.

Matheson, J. (2022). The Virtue of Epistemic Autonomy. In Matheson, J. & Lougheed, K. (eds.) *Routledge studies in epistemology: epistemic autonomy*. N.Y: Routledge, pp. 173-95.

Narveson, Jan. (1973). Moral problems of population. *The Monist.* 57: 62-86.

Nguyen, C.T. (2019). Expertise and the fragmentation of intellectual autonomy. *Philosophical Inquiries* 6 (2):107-124.

Nozick, R. (1974). *Anarchy, state, and utopia*, Oxford: Blackwell.

Percival, P. (2002). Epistemic consequentialism. *Proceedings of the Aristotleian Society, Supplementary Volumes,* Vol. 76, pp. 121-151.

Pettigrew, R. (2013). Epistemic utility and norms for credences. *Philosophy Compass* 8:897-908.

Pettigrew, R. (2016). *Accuracy and the Laws of Credence*. Oxford: Oxford University Press.

Pritchard, D. (2010). Cognitive ability and the extended cognition thesis. Synthese, 175(1), 133–151.

Proust, J. (2013). *The Philosophy of Meta-Cognition: Mental Agency and Self-Awareness*. Oxford: Oxford University Press.

Railton, P. (1989). Naturalism and prescriptivity. *Social Philosophy and Policy*: 151-74.

Riggs, W. (2002). Reliability and the value of knowledge. *Philosophy and Phenomenological Research*, 64(1): 79-96.

Reed, B. (2016). Who Knows? in *Performance epistemology*, (ed.) Fernández, M.A. Oxford University Press: 106-123.

Roberts, R. C., & Wood, W. J. (2007). *Intellectual virtues: An essay in regulative epistemology*. Oxford: OUP Oxford.

Ross, W.D. (1930). *The right and the good*. London: Oxford University Press.

Scanlon, T.M. (1998). *What we owe to each other*. Cambridge, MA: Belknap Press of Harvard University Press.

Silverstein, M. (2000). In defense of happiness: a response to the experience machine. *Social Theory and Practice*: 279-300.

Sher, G. (ed.) (1996). *Moral philosophy: selected readings*. 2nd Edition. Fort Worth: Harcourt Brace College Publishers.

Steup, M. &Neta, R. Epistemology. *The Stanford Encyclopedia of Philosophy* (Fall 2020 Edition), Edward N. Zalta (ed.), URL <https://plato.stanford.edu/archives/fall2020/entries/epistemology/>.

Sylvan, K. (2020a). An epistemic non-consequentialism. *Philosophical Review* 129(1): 1-51.

Sylvan, K. (2020b). "Reliabilism without epistemic consequentialism. *Philosophy and Phenomenological Research* 100: 525–555.

Zagzebski, L. (2003). The search for the source of epistemic good. *Metaphilosophy*, 34(1/2): 12-28.