Synthese (2023) 201:47
https://doi.org/10.1007/s11229-023-04043-3

ORIGINAL RESEARCH

Epistemic deontology, epistemic trade-offs, and Kant's formula of humanity

James Andow[1]

Received: 24 February 2022 / Accepted: 5 January 2023

Abstract
An epistemic deontology modelled on Kant's ethics—in particular the humanity formula of the categorical imperative—is a promising alternative to epistemic consequentialism because it can forbid intuitively impermissible epistemic trade-offs which epistemic consequentialism seems doomed to permit and, most importantly, it can do so in a way that is not ad hoc.

Keywords Epistemology · Consequentialism · Deontology · Epistemic normativity · Epistemic trade-offs

1 Introduction

Epistemic consequentialists struggle to forbid certain kinds of intuitively forbidden epistemic trade-offs due to their commitment to the idea that the epistemic right is fixed by facts about conduciveness to the epistemic good. This suggests a deep problem with epistemic consequentialism. So argues, most prominently, Berker (2013). This problem for consequentialism represents both an opportunity and a challenge for epistemic deontology.[1] The opportunity is that certain kinds of epistemic trade-offs seem to be epistemically forbidden regardless of the consequences, and epistemic deontology is well-placed to entertain absolute constraints that apply regardless of the consequences.[2] The challenge is to forbid those trade-offs in a non-ad-hoc way, i.e., provide a satisfactory story as to why there would be such absolute constraints.

[1] I use 'epistemic deontology' to refer to models of epistemic normativity analogous to deontology in ethics (rather than to include all theories concerning the epistemic right/duty).
[2] One might explore deontological epistemologies for other reasons (see, e.g. Alston, 1988 and Clifford, 1877).

✉ James Andow
j.andow@uea.ac.uk

[1] University of East Anglia, Norwich, UK

Published online: 26 January 2023

The above motivates exploring the prospects of an epistemic deontology on the model of Kant's ethics.[3] Why? Because Kant's ethics is the most prominent attempt to provide a solid grounding for absolute normative constraints.[4] It is natural to ask whether a structurally analogous approach to epistemic normativity can provide and ground epistemic norms that forbid the trade-offs that cause problems for epistemic consequentialism. The main point of this paper is to argue that an epistemic deontology modelled on Kant's formula of humanity—which tells us to always treat humanity as an end and never as a mere means—promises to be able to meet this challenge. But first, to understand why it is worth asking whether the challenge can be met, it is important to understand why the opportunity arises. What's the problem concerning consequentialism and epistemic trade-offs?

The paper proceeds as follows. Section 2 explains the nature of the challenge to epistemic consequentialism that motivates this exploration of accounts of epistemic normativity on the model of Kantian ethics. Importantly, it explains why consequentialists struggle to say the right thing about certain epistemic trade-off cases. Section 3 outlines some key aspects of Kantian ethics focussing on the categorical imperative in the FORMULA OF HUMANITY and demonstrates that an account of epistemic normativity on this model can provide a satisfactory treatment of all relevant trade-off cases. Section 4 explores an alternative approach taking the model of the FORMULA OF UNIVERSAL LAW and shows that it is less helpful for providing a successful treatment of the relevant trade-off cases.[5] Section 5 wraps up.

2 The problem for epistemic consequentialism

Epistemic consequentialism is a diverse family of views about epistemic normative properties. An epistemic consequentialist theory about some set of epistemic normative properties says the facts about those epistemic normative properties depend solely on facts about conduciveness to epistemic value. Epistemic consequentialists, beyond this point of agreement, can diverge with respect to many other issues. Which epistemic normative properties are they giving a theory of? Which terms pick out epistemic normative properties? What is the locus of evaluation of the relevant epistemic norms? What is of ultimate epistemic value? What makes one thing rank higher in terms of conduciveness to epistemic value than another? Is expected or actual value important?

[3] So my topic isn't Kant's ideas about epistemology. For explorations of Kant's ideas, see, e.g., Chignell (2007) and Hadisi (2022).

[4] The scope for saying sensible things using a Kantian framework should be clear. The idea that theoretical and practical rationality are unified under the Categorical Imperative is taken seriously in Kant scholarship, see, e.g., O'Neill (1990, chp.1) and Rescher (1999, chp.9). See Hadisi (2022) for a very different view and approach to a Kantian account of epistemic normativity. See Sylvan (2020) for another attempt to provide a rather different Kantian epistemology that tries to deal with trade-offs. See Cohen (2014) for a more similar attempt to articulate an epistemic deontology (my Section 4 argues her strategy doesn't help with trade-offs).

[5] This shouldn't be taken as a rejection of the idea that the various formulations are ultimately equivalent. Thinking all formulations are ultimately equivalent is compatible with thinking that the FORMULA OF HUMANITY is a more helpful model for thinking about epistemic normativity. The thought might be similar to the Kantian line that, while the FORMULA OF UNIVERSAL LAW expresses the structure of the categorical imperative, the FORMULA OF HUMANITY brings it closer to intuition.

Is the relevant conduciveness relation a direct one or an indirect one? And so on. This means that there are lots of different theories that fall within the consequentialist family.

The trade-off problem is that epistemic consequentialisms allow certain epistemic trade-offs which are intuitively not allowed.[6] The problem is not supposed to be one of mere counterexamples. The thought is that this systematic failing betrays a deep problem: that to understand epistemic normativity in terms of conduciveness to epistemic value is just a mistake. While all epistemic consequentialisms face a trade-off problem, not all trade-off cases are problematic and it is different trade-off cases that cause a problem for different epistemic consequentialisms. So the most helpful way to explain the trade-off problem in its general form is to look at a range of trade-offs that are problem cases for a range of different epistemic consequentialisms.

First, consider a direct epistemic consequentialist theory that operates with a veritist account of final epistemic value:[7]

DEC It is epistemically permissible for S to believe that p at t iff S's believing that p at t maximizes overall epistemic value for S, where overall epistemic value for S is determined by the number of true beliefs S ends up with minus the number of false beliefs.

Such a theory struggles with cases of the following form.[8]

TRUTH FAIRY Suppose Adriana starts with no reason to believe that p is true and no reason to believe that it is false. The Truth Fairy is a very powerful being, and she makes Adriana the following credible offer: if Adriana believes that p, the Truth Fairy will make Adriana's epistemic situation very, very good overall. The Truth Fairy will arrange for Adriana to have many, many true beliefs, and very, very few false ones. However, the Truth Fairy does not guarantee that Adriana's belief that p itself will have any particular epistemic status as a result of her actions.

Such cases represent a problem for accounts such as DEC. DEC is committed to treating Adriana's believing that p as epistemically permissible, but intuitively it is not epistemically permissible. Although the Truth Fairy's offer might make it, e.g., incredibly prudent for Adriana to believe that p, our intuitions are clear that nothing in the offer gives Adriana any *epistemic* reasons for believing that p.

Note that, while DEC assumes a veritist account of final epistemic value, and TRUTH FAIRY is constructed accordingly with a fairy who promises many true beliefs and few

[6] I bracket various relevant empirical questions here. However, there is some empirical work, e.g., I have empirically examined some relevant issues concerning permissibility judgments in epistemic trade-off cases in previous work (Roberts et al., 2018 and Andow, 2017a), and there is some empirical work on voluntarism about mental states such as belief (see note 15). A final theory of epistemic normativity will have to take such evidence into account.

[7] Two notes: (a) See Berker (2013) for the distinction between direct and indirect epistemic consequentialisms; (b) I use the terms '(im)permissible', 'right/wrong', etc., simply to pick out the main normative property of interest in our epistemic norms (the most basic sense of 'epistemically (not) okay') as doing so facilitates the structural analogy with ethical theories. It isn't straightforward which everyday epistemic terms/concepts, such as 'justified', 'warranted' and 'rational,' or their technical uses in epistemology, we should think of as picking out this property.

[8] Case based on Elstein and Jenkins (2019) via a previous paper of mine, Andow (2017b). For more cases, see, Berker (2013).

false ones, similar cases can be constructed that would cause problems for any simple direct consequentialist theory with a different account of final epistemic value.[9] In this section, I consider three further consequentialisms and three further cases. Each of these also assumes veritism, but, again, similar cases would cause similar problems for similar theories with different accounts of final epistemic value.

Second, consider a 'no cross-propositional trade-offs' epistemic consequentialism which is capable of forbidding the trade-off in TRUTH FAIRY because it forbids trading-off epistemic value with respect to one proposition for the sake of epistemic value with respect to other propositions.

NEC It is epistemically permissible for S to believe that p at t iff S's believing that p at t maximizes epistemic value *concerning p* for S, which, within a veritist framework might be something like *maximises the probability S has a true belief concerning whether p*.

While the TRUTH FAIRY causes no problems for NEC, NEC still faces problematic trade-offs. Consider cases of the following form.[10]

LAST FAIRY Every time a child stops believing that fairies exist, there is 50% chance that a fairy somewhere falls down dead. Tatiana is the last child on earth to believe that fairies exist but does so simply on a whim in the face of compelling evidence and testimony to the contrary. Unbeknownst to Tatiana, fairies do exist but are teetering on the edge of extinction. Indeed, Tinkerbell is the last remaining fairy.

Again, it seems *epistemically* impermissible for Tatiana to have formed this belief. It seems impermissible despite the fact that having the belief maximizes the chances that she has a true belief about the very proposition that is the content of the relevant belief.

Third, consider the following account whose locus of evaluation is belief formation processes.

FEC It is epistemically permissible for S to use belief formation process B iff S's using B maximizes epistemic value for S, where overall epistemic value for S is determined by the number of true beliefs S ends up with minus the number of false beliefs.

While the fairies we've encountered so far are irrelevant to FEC, WORSE FAIRY causes FEC trouble.[11]

[9] To help make the point: consider that UNDERSTANDING FAIRY offers an similarly unacceptable bargain if she promises great overall levels of understanding in the event that you come to 'understand' one particular thing. Her bargain can't make it an epistemically appropriate state to be in with respect to that one thing. What 'understanding one particular thing' amounts to will depend on your preferred account, of course, but it might include a lot (e.g., accepting clusters of propositions, treating them as standing in certain inferential/explanatory relations, regarding that relationship as constitutive of understanding, there being a particular kind of phenomenology related with doing so).

[10] There are two differences between TRUTH FAIRY and LAST FAIRY (likewise for the other pair). Each serves a different purpose in the paper. The difference in the relation between the relevant epistemic *cost* and *pay-off* in the trade-off (e.g., whether they concern the same proposition) helps make clear how different cases target different forms of consequentialism (e.g., DEC vs NEC). The difference in the agent's end (e.g., a deliberate attempt to maximise value vs the satisfaction of a whim) is not relevant to how consequentialisms handle the cases. But it is an important difference later when it comes to what strategies the Kantian requires to handle the full range of trade-offs.

[11] Case based on one I used in previous work, see Andow (2017b).

WORSE FAIRY Suppose Tinashe starts with no reason for using a belief forming process B. The Worse Fairy is a very powerful being, and she makes Tinashe the following credible offer: if Tinashe uses B, she will make Tinashe's epistemic situation very, very good overall. She will arrange for Tinashe to have many, many true beliefs, and very, very few false ones. There is no guarantee and no particular reason to think that using B would, without the Worse Fairy's bargain in play, have any particular positive consequences.

Accepting the offer in WORSE FAIRY is no better than accepting the offer in TRUTH FAIRY but is clearly allowed by FEC. Note that cases similar to WORSE FAIRY, in which the adoption of B will output p, a proposition which, without the fairy's offer in play, the agent would have no epistemic reason to accept, are going to cause problems for indirect epistemic consequentialisms whose locus of normative evaluation is *beliefs*.[12] Such indirect consequentialists include those who endorse IEC: It is epistemically permissible for S to believe that p at t iff S's believing that p is the result of belief formation process B that, if always used by S, would maximize overall epistemic value for S, where overall epistemic value for S is determined by the number of true beliefs S ends up with minus the number of false beliefs.[13]

Fourth, consider an account like FEC but which forbids trading off epistemic value with respect to one process for the sake of epistemic value with respect to other processes.

REC It is epistemically permissible for S to use belief formation process B iff S's using B maximizes epistemic value concerning B for S, which, within a veritist framework might be something like *leads to a high ratio of true to false beliefs output by B*.

And then consider cases of the following form:

FAIRY GENERATOR Unbeknownst to humanity, fairies exist and are born through spontaneous generation but only in the homes of children who have adopted the following belief forming process: whenever you see something sparkly or hear something tinkly, believe it was caused by a fairy. The rate of generation is such that, in the homes of such children, the number of sparkles and tinkles that are caused by fairies quickly exceeds the number with other causes. One child, Hope,

[12] We can run versions of this case in which the agent is assured that p will be output by the process ahead of their decision to adopt B, and others where they are unaware. In neither version would the fairy's offer redeem believing p in such a way.

[13] This is not Goldman (1979)'s process reliabilism (see Goldman, 2015, (a reply to Berker) and discussion in Berker, 2015 and Vahid, 2016). Process reliabilism doesn't make conduciveness sufficient for justification. So, it is not consequentialist in the relevant sense, being open to side-constraints that could forbid trade-offs (see also, Dunn and Ahlstrom-Vij, 2017), (such constraints being in need of, note, a nonconsequentialist normative foundation like the Kantian might provide). Moreover, Goldman's suggested approaches to *reliability* assess truth ratio with reference to processes' behaviour *across nearby possible worlds*. Since nearby worlds contain no fairies, the process in WORSE FAIRY isn't reliable in this sense. While one can bring intuitions about TRUTH FAIRY, etc., to bear on what we might call 'modal consequentialisms,' it requires reflecting on the intuitions' deeper nature (à la Elstein and Jenkins, 2019). That would take us beyond the scope of this paper. But the basic idea is: *what the epistemic norms are* intuitively doesn't depend on contingent facts about the preponderance of fairies; but to rely on a modal interpretation of conduciveness to the good to outlaw accepting fairy-bargains is to say it does.

> on a whim, and in the face of compelling evidence and testimony to the contrary, adopts such a belief forming process.

I take it that, intuitively, the effects of forming beliefs in this way do not redeem the irrationality of adopting a belief forming process on a whim.

As Berker (2013, p. 377) argues, this systematic failure of epistemic consequentialism to forbid forbidden trade-offs seems to be a symptom of a deeper problem.

> It is epistemic consequentialism's fixation on the promotion of epistemically valuable state of affairs, its reduction of beliefs to mere instruments serving our independent epistemic ends, that causes all of these problem cases to arise…[E]pistemic consequentialism tries to analyse intrinsic epistemic merit in terms of instrumental epistemic merit, and for this reason should be rejected.

We seem to be able to come up with problematic trade-offs for any version of epistemic consequentialism we pick. This suggests there is something wrong with the key commitment of epistemic consequentialism—that the right is a matter of conduciveness to the good—for it is that commitment which leads to inevitably sanctioning trade-offs along the lines of the fairy cases in the above.

3 Formula of humanity

As epistemic consequentialism can be understood by analogy with consequentialism in ethics, so epistemic deontology can be understood by analogy with deontology in ethics. Here's Alexander and Moore (2016, Sect. 2) on the difference in ethics

> In contrast to consequentialist theories, deontological theories judge the morality of choices by criteria different from the states of affairs those choices bring about. The most familiar forms of deontology, and also the forms presenting the greatest contrast to consequentialism, hold that some choices cannot be justified by their effects-that no matter how morally good their consequences, some choices are morally forbidden. On such familiar deontological accounts of morality, agents cannot make certain wrongful choices even if by doing so the number of those exact kinds of wrongful choices will be minimized…For such deontologists, what makes a choice right is its conformity with a moral norm. Such norms are to be simply obeyed by each moral agent; such norm-keepings are not to be maximized by each agent.

Epistemic deontology will thus be well-placed to deliver absolute constraints of the kind needed to forbid the relevant epistemic trade-offs. The challenge will not be to provide such constraints per se—one could do that in a completely ad hoc way—but to provide a satisfactory account as to why there would be absolute constraints against those trade-offs. Given the nature of the challenge, it is natural to hope that it might help to consider building an epistemic deontology on the model of Kant's deontological approach in ethics. Kant's ethics, as set out in *The Groundwork of the Metaphysics of Morals*, is the most famous attempt to provide a solid grounding for absolute normative constraints and, in particular, ones that rule out certain kinds of trade-off.

Since Kant's ethics is to be our model for an epistemic deontology we first need a quick outline of the relevant features of Kant's ethics.[14] Kant's key idea is that moral norms are supposed to articulate ways in which we have reasons to act regardless of the purposes we are pursuing and the ends we desire. What could ultimately be of intrinsic moral value, i.e., what could be of value independent of the ends we choose for ourselves, what could be an end in itself, an end that has a claim on all rational agents? The Kantian story is that we are rationally committed to act in certain ways (and not in others) simply in virtue of our rational agency, and it is that rational agency, or 'humanity,' that is to be regarded as ultimately valuable, i.e., the 'general capacity for choosing, desiring, or valuing ends' (Korsgaard 1996a, p. 114).[15] This is Kant's Categorical Imperative which he formulates in a few different ways. In this paper, I'm first going to focus on FORMULA OF HUMANITY. It is this formulation which I argue provides the model for an epistemic deontology that can give a satisfactory treatment of epistemic trade-offs.[16] In the next section, I'll argue that the same can't be said for the FORMULA OF UNIVERSAL LAW.[17]

Here Kant's FORMULA OF HUMANITY from *Groundwork for the Metaphysic of Morals* (29):

> FORMULA OF HUMANITY Act in such a way as to treat humanity, whether in your own person or in that of anyone else, always as an end and never merely as a means. [Ak 4:429]

We need to know how to derive duties from the Categorical Imperative. Kant talks about two different types of duty which are derived in slightly different ways.[18]

[14] My understanding of the project of Kantian ethics is informed by Korsgaard (particularly 1985). But I don't have space to get into interpretative issues here.

[15] An interesting question: Must such a Kantian story be committed to doxastic voluntarism (or indeed a broader epistemic voluntarism)? Must the relevant agency in the epistemic realm require direct rather than indirect voluntarism? Maybe not. But many will suspect so. So it is worth saying a few things about the issue. (A) The rich literatures on voluntarism and the related issue of epistemic akrasia will be valuable resources in developing a full Kantian account along the lines I sketch here, which requires working out what kind of agency is required and whether we have it (e.g., Heil, 1983; Alston, 1988; Greco, 2014; Owens, 2002; Horowitz, 2014; Borgoni, 2015; Borgoni and Luthra, 2017; Neta, 2018; Coates, 2012), as will discussions about akrasia within Kantian ethics (e.g., Korsgaard, 1996b; Wallace, 2001). (B) Most/all non-Kantian accounts of epistemic normativity may also be hostage to voluntarism – for 'ought implies can' reasons (Alston, 1988). (Although (i) the intuitive status of 'ought implies can' is in question Kurthy et al., 2017; Buckwalter and Turri, 2015; Chituc et al., 2016; Henne et al., 2016; Leben, 2018; Turri, 2017; Kissinger-Knox et al., 2018; Cohen, 2018; Phillips and Cushman, 2017; Buckwalter, 2019), and (ii) the Kantian account I sketch may depend on voluntarism in a way others don't. If agency is to be characterized by voluntarism, voluntarism is fundamental to the story the account I sketch tells about *the source* of epistemic normativity.) (C) Although the ordinary concept of belief has been claimed to be in tension over the issue of voluntarism (Heil, 1983), some recent studies suggest voluntarism enjoys better intuitive support than is often assumed (Turri et al., 2018; Cusimano and Goodwin, 2019).

[16] An alternative way to build a Kantian account might be to ground epistemic normativity in some characteristically epistemic agency (rather than in general practical agency as I do here). Thanks to an anonymous reviewer for suggesting that alternative strategy. Maybe someone taking that strategy could make use of some ideas sketched here. But I won't revisit the issue.

[17] But by saying that, I don't intend to commit myself either way on whether the formulas might ultimately be equivalent. See note 5.

PERFECT DUTIES One ought always (or never) to do x
IMPERFECT DUTIES One ought to do (or avoid) x sometimes, to some extent

We can derive *perfect* duties from the FORMULA OF HUMANITY in the following way. One has a perfect duty not to treat humanity, whether in your own person or in that of anyone else, merely as a means to an end, because to act in such a way *conflicts* with treating humanity as end in itself – exactly what it means to treat humanity as an end in itself will be explored a little later on. And one has an *imperfect* duty to act in ways that *harmonize* with humanity as an end in itself. And here's what Kant has to say about that notion of 'harmonizing' (30):

> For a positive harmony with humanity as an end in itself, what is required is that everyone positively tries to further the ends of others as far as he can. [Ak 4:430]

One reason to be optimistic that an epistemic deontology on this model might be able to forbid the necessary epistemic trade-offs is that the FORMULA OF HUMANITY gives the Kantian in ethics the resources to forbid moral trade-offs (whereas this isn't so straightforward using only, e.g., the FORMULA OF UNIVERSAL LAW). Consider cases like the following.[19]

> A gunman is about to fire a machine gun into a crowded train platform. Judy has climbed out of the line of fire into a precarious position balanced on a high ledge above the tracks. A nervous person has joined her. From their vantage point they can see some police officers who, as the gunman makes to pull the trigger, are still a few seconds away from capturing the gunman. If only a 10-second distraction could be contrived! Judy comes up with a clever plan. If she makes a sudden, loud noise she can …
>
> JUDY 1 …distract the gunman for long enough for police to disarm and arrest the gunman preventing a mass murder. The loud noise will unfortunately cause the nervous person to flinch, topple off the ledge onto the track and be killed by an express train which is about to pass through. Judy follows through with her plan and it works.
>
> JUDY 2 …make the nervous person flinch, topple off the ledge onto the track and be killed by an express train which is about to pass through. The nervous person's death will cause the gunman to be distracted for long enough for the police to disarm and arrest him preventing a mass murder. Judy follows through with her plan and it works.

Suppose we wanted to give different verdicts about Judy 1 (whose course of action involves no trade-off structure) and Judy 2 (whose course of action involves a clear trade-off) such that Judy 1 acts in a permissible way and Judy 2 acts in an impermissible way. This differential verdict is something which the FORMULA OF HUMANITY is well-placed to provide. Judy 2 treats the nervous person—a rational agent, their humanity—as a mere means to an end. This is not the case for Judy 1 as she doesn't use the nervous person as a means to an end.

[18] I follow Hill (1971)'s articulation of the nature of perfect/imperfect duties.

[19] A riff on the Trolley Problem. For original context and cases see Foot (1967) and Thomson (1985).

How would an epistemic deontology on the model of the FORMULA OF HUMANITY handle the epistemic trade-off cases that troubled the consequentialist? It will only be able to forbid the relevant trade-offs if making those trade-offs involves acting in such a way as to treat humanity, whether in one's own person or in that of anyone else, as a mere means rather than as an end in itself. Remember that what we are looking for, if we are to derive a perfect duty, is a potential conflict between (i) an action and (ii) treating humanity always as an end in itself and never as a mere means. And, if we are to derive an imperfect duty, what we are looking for is an opportunity to positively try to further the ends of others (or ourselves) as far as one can.

The first point to make here is that a successful Kantian strategy for forbidding epistemic trade-offs will focus on (a) the duty not to treat *one's own humanity* as a mere means, rather than (b) duties concerning the humanity of third parties. It might have seemed that focusing on third parties could be a promising strategy for the Kantian. After all, in all the epistemic trade-off cases we've looked at the agent's end is something other than the humanity of others. In some cases the agent makes the trade-off deliberately in order to improve their own epistemic position. In some cases the agent allows their epistemic lives to be driven by the satisfaction of a whim. As a result, we might have been tempted to say, performing such trade-offs fails to properly respect the humanity of third parties as ends in themselves and, indeed, uses their humanity as a mere means. However, unfortunately, we can at best derive *imperfect* duties to others not to perform the relevant trade-offs. Why? Because, despite the fact that performing an epistemic manoeuvrer to improve one's own epistemic position or to satisfy a whim doesn't give explicit consideration to third parties, it also doesn't actually *use third parties as a means*. Contrast this with the standard derivation of a perfect duty not to lie for the sake of personal gain. In that case, the perfect duty can be derived because the general truthfulness of others is a necessary means to your lie being successful. But nothing analogous applies in these epistemic trade-off cases. By believing on a whim, for example, as in LAST FAIRY, one's success in satisfying one's whim by believing isn't dependent on the rational lives of others (it simply fails to harmonize with humanity as an end in itself, i.e., one could do better by others' humanity). The same seems to apply to adopting *a belief-forming process* on a whim, as in FAIRY GENERATOR. Similarly, your ability to maximise epistemic value for yourself by taking TRUTH FAIRY's offer (or WORSE FAIRY's), in no way depends on other people not having similar policies to yours. So, it seems we can't use the FORMULA OF HUMANITY to derive a perfect duty against the kind of epistemic trade-offs that cause trouble for the consequentialist. At least, we can't do it *if we focus only on humanity in third parties*.

The prospects are better for making the case that, in making epistemic trade-offs—either on a whim or because it is conducive to epistemic value—one fails to properly respect *one's own humanity* as an end in itself. Nonetheless, the Kantian will need two different strategies for dealing with two types epistemic trade-offs we've encountered.[20] On the one hand, there are cases, such as LAST FAIRY and FAIRY GENERATOR, in which the agent makes the trade-off to satisfy a whim. At this point, I should note

[20] As noted above, this difference is only significant for the Kantian. A consequentialist account doesn't need to handle cases differently depending on the adopted end of the epistemic agent.

that the 'on a whim' in these cases serves to make them a 'stalking horse' for all trade-offs in which the agent's end is something other than humanity. The Kantian will be able to use the same strategy for all epistemic trade-offs whose end is anything other than humanity—the general capacity for choosing, desiring, or valuing ends—and not just to those made on a whim. On the other hand, there are cases, such as TRUTH FAIRY and WORSE FAIRY, where the end of the trade-off is plausibly to treat the agent's own humanity as an end in itself; the deliberate purpose of these trade-offs is to attain a very good epistemic position – something that will clearly aid the agent in exercising their agency. Deliberately attempting to inordinately improve one's general epistemic state does seem compatible with taking (one's own) humanity as one's end. As a result, such cases pose a slightly more difficult problem for the Kantian, and the strategy required to forbid them will be different. I'll now take each of these types of trade-off in turn.

How does making an epistemic trade-off on a whim fail to properly respect *one's own humanity* as an end in itself? Here, the Kantian seems to be able to derive a perfect duty relatively easily. We might say: when one believes in order to satisfy a whim, there is a failure to treat one's rational agency as an end in itself and one uses one's rational agency as a means to that end. How does one use one's rational agency as a means? The thought is that to satisfy one's whim one does something that puts in place a likely barrier to achieving one's ends, e.g., a false belief may lead one to select courses of action that cannot achieve one's intended goal. This strategy for deriving a perfect duty not to make epistemic manoeuvrers on a whim is isomorphic to the Kantian derivation of the perfect duty not to commit suicide. Suicide to avoid unbearable pain is off limits because it treats one's humanity as a means—it destroys one's humanity in order to avoid pain—without simultaneously treating one's rational agency as an end in itself. While somewhat different in scale and significance, this model does seem applicable to the case of believing on a whim. In believing on a whim, one doesn't treat one's humanity as an end in itself. The end of believing on a whim is the satisfaction of the whim. One does use one's humanity to obtain that end insofar as one accepts some potential impediment to navigating the world in a rational fashion. This derivation will work in the case of any epistemic trade-offs whose end is anything other than humanity, the 'general capacity for choosing, desiring, or valuing ends' and not just to making manoeuvrers on a whim.

However, the strategy just sketched won't work for deriving all the duties the Kantian needs to forbid all the problematic kinds of trade-off. Consider cases like TRUTH FAIRY and WORSE FAIRY. In these cases, the end of the trade-off *does* plausibly treat the agent's own humanity as an end in itself. So we can't derive a duty not to perform these trade-offs using the same resources as we just used to derive a duty not to, e.g., believe on a whim. To see this, it is helpful to think about TRUTH FAIRY and WORSE FAIRY on the model of self-improvement. Typically, so goes the Kantian line, one has an imperfect duty toward self-improvement. In a typical case, call it Case A, one starts off with a portfolio of skills $S_{1...n}$, at the end of the process each of $S_{1...n}$ has been enhanced, and the enhancement of skills was one of linear improvement over time. Now consider Case B and Case C. In Case B, to improve one's overall portfolio of skills one must embark on a course of training which will lead to a slight erosion of one of the skills one currently has with vast improvements to $S_{2...n}$ but some degra-

dation of S_1. In Case C, the path to self-improvement is non-linear over time, it is not a process of constant moment-on-moment improvement but involves some stages of inevitable backsliding, although the final portfolio is one in which each skill has been enhanced relative to its state at the beginning of the process. I think that in the case of self-improvement, it seems very strange to think that the difference between A on one hand and B or C on the other is such that in A one treats humanity as an end in itself and in B and C one does not. We shouldn't say that. Likewise, it seems wrong to say that by accepting the offer in TRUTH FAIRY and WORSE FAIRY one fails to treat one's humanity as an end in itself or even fails to harmonize with the end of humanity simply because of the fact that the path to epistemic improvement isn't straightforward and involves accepting some potential impediments to navigating the world in a rational fashion. In other words, we can't use quite the same resources to forbid trade-offs in TRUTH FAIRY and WORSE FAIRY as we applied to the case of believing on a whim.[21]

So, how can an epistemic deontology on the model of the FORMULA OF HUMANITY forbid trade-offs in cases like TRUTH FAIRY and WORSE FAIRY? How does making an epistemic trade-off because it is conducive to epistemic value fail to properly respect *one's own humanity* as an end in itself? To get to an answer we need to unpack a little further what it really means to treat one's own humanity as an end in itself and not a mere means. It helps to think of accepting the offer in TRUTH FAIRY and WORSE FAIRY as a form of paternalistic self-deception. For what is required of the protagonist in the TRUTH FAIRY, if she is to accept the offer, is to bring herself to accept a belief (which she has no independent reason to accept) in order to maximize her overall epistemic state.

How can the FORMULA OF HUMANITY be used to derive a duty not to engage in paternalistic self-deception? Start by thinking about how the FORMULA OF HUMANITY plays out in cases of deception. The Kantian generally thinks that deception of others is problematic. The reason is that to deceive someone impinges on their rational agency. Deception doesn't leave the deceived person free to make up her own mind. Korsgaard (1986, pp. 139–140) puts that in a helpful way:

> If you give a lying promise to get some money, the other person is invited to think that the end she is contributing to is your temporary possession of the money: in fact, it is your permanent possession of it. It doesn't matter whether that would be all right with her if she knew about it. What matters is that she never gets a chance to choose the end, not knowing that it is to be the consequence of her action.

Now note that this same analysis applies even in the case of *paternalistic* deception, e.g., a paternalistic lie. A paternalistic lie is a lie told to an individual motivated by a paternalistic care and the judgment that, on balance, it is better for that individual to have a false belief than a true belief (or indeed no belief). Paternalistic lies remain problematic, indeed wrong, on the whole, because they do not allow the individual

[21] Indeed, plausibly one has an imperfect duty to perform epistemic trade-offs, in these cases, for similar reasons as in the case of self-improvement.

lied to the freedom of making up her own mind, to choose or not choose the end that the liar supposes to be in their best interest.[22]

Now, note further that paternalistic deception remains problematic even in similar cases where the liar's plan is oriented towards the chosen end of the individual who is lied to. Consider the case in which you are highly invested in trying a particularly steep and scary water slide but struggle with a fear of heights which means you are worried you won't be able to realise your goal. I advise you to just sit at the top of the slide as it will "give you the time to overcome your fears and get yourself into the right mental state of readiness to take on the scary water slide." This is just a ruse, however, because I know that the water slide is so slippery that once you sit at the top there is no way to avoid sliding all the way down. Or, consider the case in which you want to perform better in your job and I advise you to perform a bunch of spurious magical rituals at lunchtime to "appease the dark angels of data entry who are currently impeding your performance." This is just a ruse, however, because I'm expecting the rituals to have the non-magical effect of simply making your chatty and distracting colleagues less inclined to engage you in conversation leaving you with more time to complete your tasks. The paternalistic deception in this case impinges on your rational agency because it doesn't allow you rational freedom in the pursuit of your ends. You don't get to make your mind up about the means of realising your goal. Your rationality is bypassed (even if you are not exactly denied the chance to choose your goal).

And finally, note that a similar analysis applies even in the case of paternalistic *self*-deception (as we might think of TRUTH FAIRY and WORSE FAIRY).[23] For similar reasons, I should not dupe *myself* into accepting the efficacy of superstitious nonsense, or into thinking that sitting on the top of a slide will give me time to overcome my fears, even if such beliefs happen to be conducive to my ends.[24] To make the issue with the relevant kind of self-deception clear, we need to appreciate the kind of mental duplicity required on the part of the self-deceiving individual. We need to distinguish between myself as the deceiver (with the plan to pursue a given end by a given means) and myself as the deceived (the deception of whom is the means). We could call these parts of me (it doesn't matter for this discussion whether we think of them as

[22] Except perhaps lying 'to someone who lacks autonomy if our end is to restore or preserve her autonomy, or to restore or preserve things which are necessary conditions of it' (Korsgaard, 1988, 350, see also O'Neill 1985).

[23] More generally, it seems doubtful that a Kantian line on deception can be imported over to all phenomena typically thought of as self-deception. Self-deception as a more general phenomenon is typically understood for the Kantian to be concerning primarily because of the effect of *habitual* self-deception on rational agency. However, the issue of how best to understand the Kant(ian) line on self-deception as a more general phenomenon is quite complicated for various reasons (see Darwall, 1988; Baron, 1988; Wood, 2002 Martin, 1988 and Papish 2018, for some helpful discussion). However, in the kinds of cases that concern us, i.e., epistemic trade-offs, the nature of the "deception" required makes the cases relevantly analogous to the deception of others.

[24] Such cases feel different if one envisages versions which are closer to everyday instances of "self-deception" in which I am "in on the trick" or "playing along" with the deception—not really accepting the lie and nonetheless somehow gaining the benefit. But those cases aren't relevant to the discussion here. The kind of self-deception that concerns us here, as it is the kind involved in epistemic trade-offs, requires a more straightforward deception of the self (in the sense that, e.g., getting the Truth Fairy's reward requires a genuine acceptance of p as true).

temporal parts or as psychological parts). It is wrong for me (the deceiver) to deceive myself (the deceived) because I (the deceiver) bypass my (the deceived's) rationality by denying myself (the deceived) rational freedom. I (the deceived) can't be in on the plan (otherwise I'm not deceived and the plan won't work). It doesn't matter whether I (the deceived) would be all right with the plan if I (the deceived) knew about it—and of course it is essential to the plan that I (the deceived) don't. What matters is that I (the deceived) never get a chance to be in on the plan. It is in this sense that in such cases I do not allow myself the relevant freedom to make up my mind for myself about the deceptive plan. Instead, I allow my rationality to be bypassed and thereby fail to properly respect my humanity.[25]

It is this sense in which an agent fails to allow themselves the freedom to make up their mind when they make an epistemic trade-off on the basis that doing so is conducive to epistemic value.[26]

Consider the agent who takes the TRUTH FAIRY's offer. To take the TRUTH FAIRY's offer, the epistemic agent needs to achieve the kind of mental duplicity we've just been talking about. As someone trying to maximise epistemic value, the epistemic agent taking the TRUTH FAIRY's offer needs to bring themselves to believe that p; they must invite themselves to accept p as true (and not just permissible); they must present p to themselves as true (not just permissible). To take the TRUTH FAIRY's offer, the epistemic agent needs to bring themselves into a state with respect to p that they (or some part of them) understands to contribute to the end of having true beliefs by being true – but of course that's not what's going on.[27] To take the TRUTH FAIRY's offer, you (the "doing that which is conducive to epistemic value" part of you) need to deny yourself (the "accepting that p" part of you) the freedom to make up your mind as to whether and how believing that p contributes to your chosen ends.[28] This is how deliberate epistemic trade-offs involve a troubling kind of paternalistic self-deception, denying oneself freedom in pursuing one's own ends. To truly treat one's humanity—one's rational agency, one's capacity for choosing, desiring, or valuing ends—as an end in itself is not compatible with allowing it to be bypassed in this way.

[25] The strategy developing here to tackle cases such as TRUTH FAIRY and WORSE FAIRY would also generate a duty not to believe/adopt a belief-forming process in cases with similar structures to LAST FAIRY and FAIRY GENERATOR but in which the relevant epistemic manoeuvrers are part of deliberate consequentialist strategies rather than simply to satisfy a whim. In such cases there's just the same kind of peculiar luck involved which results in the end you invite yourself to adopt, in the effort towards self-deception, happening to align with relevant consequentialist end. In this respect, such trade-offs are similar to a case of paternalistic self-manipulation in which you persuade yourself you will get rich in order to bring it about that you will get rich. Such self-manipulation fails to properly respect your own humanity as an end in itself, you use your rational agency as a mere means.

[26] The concern is not that the agent denies themselves the freedom to make up their mind over the normative issue (e.g., whether the belief is *permissible* or *justified*). To insist agents bracket off consequentialist considerations from their deliberations about this *normative* issue would clearly beg the question.

[27] Thanks to an anonymous reviewer for this journal for pushing me to express things in this way.

[28] That thought will be more complicated for consequentialisms whose locus of normative evaluation (e.g., belief formation processes) is distinct from its bearer of ultimate value (e.g., beliefs) (e.g., FEC) and I can't develop it in full here. But the basic thought will be similar: in adopting a belief formation process, one commits to accepting its outputs as true on the basis that, for example, its outputs will be generally true, and it is that evaluatively-relevant feature of a belief formation process which one is not allowing oneself to make up one's mind about when one accepts worse fairy's offer.

So, can the deontologist provide a system of norms that forbids these trade-offs that cause problems for epistemic consequentialism? Of course, they could have done so by fiat. Can they provide a satisfactory account as to why trade-offs would be forbidden? The above represents a first step towards doing that. The reason it is only a first step is, in part, that there's a bit more work to be done by the Kantian here to demonstrate that the considerations just sketched above – concerning deliberate trade-offs – will give a satisfactory treatment of problem cases across the full range of possible consequentialisms. The reason it is only a first step is also, in part, that, for all I have said above, the Kantian approach to grounding normativity in rational agency, and the Categorical Imperative, might be completely misguided – and I don't attempt anything like a defence of Kant's categorical imperative here. While neither of those tasks is something I can develop at length here, I hope to have articulated grounds for optimism that a fully-developed epistemic deontology on the model of Kant's moral philosophy may be able to provide a firm foundation for norms that forbid epistemic trade-offs. The great promise of Kant's categorical imperative is to provide a firm grounding for normativity, and an epistemic deontology modelled on Kant's categorical imperative as articulated by the FORMULA OF HUMANITY seems able to produce intuitive verdicts about epistemic trade-off cases. As such, an account of epistemic normativity modelled on Kant's formula of humanity has been shown to be a promising alternative to epistemic consequentialism.

4 Universal law

Cohen (2014) explores the prospects for an epistemic deontology on the model of Kant's categorical imperative focused on the FORMULA OF UNIVERSAL LAW (and the associated method of deriving duties via a universalizability test). Her hope is to show 'it has the potential to provide a robust Kantian account' that is 'capable of contributing to current debates about the ethics of belief' (318). In this section, I suggest that it is the FORMULA OF HUMANITY, rather than Kant's ethics in general, that provides a promising model for an epistemic deontology.[29] The reason is that taking Kant's FORMULA OF UNIVERSAL LAW as a model for an epistemic deontology doesn't produce an account that can easily handle trade-offs in a satisfactory way.[30]

Why do I think taking the FORMULA OF UNIVERSAL LAW as a model for an epistemic deontology doesn't produce an account that can handle trade-offs in a satisfactory way? Before we get to that, we must first understand the relevant aspects of Kant's ethics. Remember that for Kant, morality is supposed to bind one regardless of whether one wants or chooses to be moral; morality binds all rational agents simply in virtue of their rationality. To work out the nature of such a categorical imperative, we can reflect on questions like the following: What kind of norm could one be rationally committed to following? How could a particular course of action chosen by an agent be irrational simply in virtue of that agent's being rational? This helps us see why what is going to make for a rational or an irrational action, for the Kantian, is doing

[29] The formulas might be equivalent even if one of them is a more helpful model. See note 5.

[30] The strategy I consider isn't exactly Cohen's. Some footnotes highlight and explain differences.

such-and-such *for a particular reason*. This is why the Kantian story often centres on the notion of a *maxim* which, as Cohen puts it, 'formulates an agent's policy or intention' in acting. A classic example of a maxim might be

LIE I will lie to make it easier for me to achieve my goals whenever I am in a position to do so

It is acting on certain maxims, such as LIE, that is irrational, in the Kantian story. What could make acting on a maxim categorically irrational? The Kantian thought is as follows: by acting on a maxim you treat that maxim as encoding sufficient reason for *your* performing that action; by treating yourself as having sufficient reason for acting when you act on that maxim you are committed to *anyone* acting on that maxim having sufficient reason for acting; so, you are committed to it still being the case that you would have sufficient reason for acting when acting on that maxim in a world in which that maxim was a *universal law of action*.

This is the reasoning behind Kant's first attempt to articulate the nature of the Categorical Imperative (Kant, 24).

FORMULA OF UNIVERSAL LAW Act only on that maxim through which you can at the same time will that it should become a universal law. [Ak 4:421]

Which maxims is it is impermissible to act upon? Which is it irrational to act upon simply in virtue of being a rational agent? The ultimate test of maxims is to be a universalizability test. We are to try to imagine a world in which our maxim was a universal law of action, and to consider whether acting on such a maxim would be justified in such a world – would it encode sufficient reason for acting. Maxims that fail the universalizability test are maxims that we have a duty not to act upon.

Kant claims there are two ways a maxim might fail a universalizability test.[31]

FAILURE BY CONTRADICTION IN CONCEPTION A maxim fails because acting on that maxim could not achieve the purpose it encodes in a world in which that maxim was a universal law of action. There would be no advantage to be gained, for example, by lying in a world in which everyone always tried to do so, and so lying to gain advantage is irrational.[32] An action whose maxim fails by contradiction in conception fails because it could not achieve its purpose in a world in which its maxim was universalised. By acting on such a maxim in this world, one is treating oneself as exceptional and exploiting the rational agency of others, because your act can only achieve its purpose because others don't act on this maxim.

FAILURE BY CONTRADICTION IN WILL A maxim fails because a world in which your maxim were a universal law of action would be a world in which your rational agency—your ability to pursue your own ends—would be impaired. By acting on the maxim 'avoid helping those who ask for help in order to avoid expending energy,' for example, you are treating this as being a justified course of action

[31] It's controversial how best to understand the distinction between contradiction in conception and will (see Korsgaard, 1985). The interpretation here is the 'Practical Contradiction Interpretation' which Korsgaard defends.

[32] As Korsgaard (1985) puts it, "What the test shows to be forbidden are just those actions whose efficacy in achieving their purposes depends upon their being exceptional. If the action no longer works as a way of achieving the purpose in question when it is universalized, then it is an action of this kind."

> for anyone, but in a world in which this was a universal law of action it would generally be a lot more difficult for you to exercise your will.[33]

What would an epistemic deontology look like which was constructed on the model of the FORMULA OF UNIVERSAL LAW and an epistemic universalizability test? Cohen provides a helpful illustration of how to test epistemic maxims for universalizability. In the example, the maxim fails the test because we end up with a contradiction in will.[34] Here's the example (Cohen, 2014, p. 323):

> …I am in the process of determining whether I should believe that p. As I do so, I encounter a piece of evidence that falsifies it. If I ignore this evidence and believe p anyway because it suits my desires, I am effectively thinking under the maxim:
>
> ¬EM I will ignore evidence in cases when it falsifies a belief I desire to be true.

And Cohen (2014, p. 324)'s explanation for why this maxim fails to be universalisable is as follows:

> …I am a cognitively dependent being who needs epistemic help from others. Yet if the maxim ¬EM were universalized, others' beliefs would be unreliable. I could never be sure whether any given belief they hold is based on their wishes or on objective grounds. On this basis, I could never rely on their cognitive contribution, which, as an epistemically dependent being, I cannot possibly will. Therefore, the maxim ¬EM leads to a contradiction in the will: I cannot consistently will it to be a universal law.

[33] It's common to associate perfect duties with contradiction in conception and imperfect duties with contradiction in will. But perfect duties could be detected via contradiction in will (Korsgaard 1985). A contradiction in will could be such that by willing the relevant maxim to be universal law one would will something that would genuinely incapacitate you as a rational agent, i.e., genuinely thwart one's agency. In cases of this kind (and only in such cases) a maxim failing universalizability via contradiction in will indicates a perfect duty not to act on it.

[34] The best case for an epistemic maxim that generates contradiction in conception I can think of is 'believe only falsehoods', as a world in which this was universalized might well be a world in which the mental state of belief died out. But this generates only trivial duties. Cohen gives one example of substantive epistemic maxim she thinks fail: 'I will not believe testimony.' She claims it fails because, "the universalization of the maxim not to believe testimony would entail the disappearance of the practice of testimony, since in a world in which no one believed testimony, giving it would become a pointless exercise" (Cohen, 2014, p. 327). Cohen thinks this leads to contradiction in conception just like the classic case of promise-breaking: if no one could be relied upon to keep promises, there would be no practice of promise-making, and thus you couldn't advance your interests by making a promise and breaking it. I don't find this comparison compelling. The reason you shouldn't make false promises for personal gain, on the Kantian account, is because in doing so you would rely on the practice of legitimate promise making in others and this can be detected by thinking about whether false promising would be a coherent policy in a world of universalised false promising—and it isn't. But that is not what's going on in the case of refusing to believe testimony. The coherence of refusing to believe testimony doesn't rest on wider practices either of giving or believing testimony. The coherence of a policy to not eat turnips doesn't depend on the existence of turnips or the practice of eating them. The coherence of a policy to not commit terrorist atrocities doesn't depend on their being an extant practice of terrorist atrocities not to commit. And that can be shown by the fact that such policies can be pursued and coherent—albeit admittedly quite empty in some cases—in a world in which there is no testimony, turnips, terrorist atrocities, or whatever. So any perfect duty we have not to be testimony disbelievers can't be detected through contradiction in conception.

Can such an epistemic deontology on the model of the FORMULA OF UNIVERSAL LAW give a satisfactory treatment of epistemic trade-offs such as TRUTH FAIRY, LAST FAIRY, WORSE FAIRY, and FAIRY GENERATOR? I don't think it can.[35]

Consider TRUTH FAIRY. The truth fairy promises significant epistemic value if you believe that p but makes no guarantees about the final value of that belief. The maxim underlying the acceptance of the belief would be something like *I will adopt this belief in order to secure significant epistemic value for myself*. Now suppose it were a universal law of belief formation that people always accepted offers like those in TRUTH FAIRY (and indeed accepted any belief whose formation resulted in a significant epistemic pay-off). Would it be impossible or incoherent to act on such a maxim in such a world? I see no reason to suppose so.[36] Would willing such a world be somehow willing the frustration of one's own ability to pursue one's ends? Again, I see no reason to suppose so. If our world were populated with epistemic agents who always took opportunities to maximize epistemic value for them, that could only make it easier to pursue one's ends simply in virtue of the world being populated by better sources of testimony. And yet, the intuitive verdict with respect to the truth fairy case is pretty clear: it is not epistemically okay to accept her offer and believe. There is no practical irrationality involved in choosing to believe in such a way. (Or rather, there *is* practical irrationality involved, it is just that you can only detect it through the FORMULA OF HUMANITY and not a universalizability test.)

A similar story applies to WORSE FAIRY. The nature of the offer in the WORSE FAIRY is that to secure significant epistemic value one must adopt a particular belief-forming process. Would it be impossible or incoherent to act on the maxim 'when in a position to do so, adopt a belief-forming process to secure significant epistemic value' in such a world? I see no reason to suppose so. Would willing such a world be somehow willing the frustration of one's own ability to pursue one's ends? Again, I see no reason to suppose so. Although a world in which everyone always acted on this maxim would be strange, and would perhaps contain agents with some surprising beliefs, it would nonetheless be a world in which the information available from the average epistemic agent would be (if any different) higher quality rather than lower – and thus there is no contradiction in willing such a world. Again, a universalizability test can detect no practical irrationality to accepting offers like those in the WORSE FAIRY. We need the FORMULA OF HUMANITY for that.

[35] Cohen (2014) would likely not agree. But see n. 34 & 38 for some defence of my interpretation of how an epistemic deontology on the model of the formula of universal law would pan out.

[36] It has been put to me that the practice of taking such offers would destroy the practice of believing—perhaps people would no longer typically *believe* when they accepted a proposition as true but do something different—and that this means the practice of taking such offers fails to universalise due to contradiction in conception. Although the line is similar to the one I pushed in n.34 (in relation to believing only falsehoods), I don't buy it in this case for reasons I don't have space to get into. However, and more importantly, even if this line of thought were to go through, such a contradiction in conception wouldn't indicate the kind of substantive epistemic duties the Kantian needs to ground. *Believing* on the basis of such offers might be out, but only on a technicality. There's no similar line of reasoning that can get you a prohibition on *accepting propositions as true, and allowing them to feature in your life in exactly the same way as if you believed them* on the basis of such offers, but the Kantian needs to be able to provide those kinds of prohibitions too as none of the intuitions about trade-off cases seem to rest on a technicality about what counts as a belief. A similar point can be made about similar moves in relation to the other cases too.

In LAST FAIRY, Tatiana believes on a whim, in the face of compelling evidence, and is unaware of the consequences of her belief including the impact on final epistemic value for her. Here Tatiana's maxim is something along the lines of *I will adopt this belief because I feel like it* and that can clearly be universalized without contradiction in conception. Everyone always believing on a whim when they felt like it would in no way undermine your ability to believe what you feel like believing. However, in this case, we do encounter contradiction in will and so it is worth asking whether this is the kind of contradiction in will that would indicate a perfect duty.[37] If everyone believed stuff on a whim whenever the whim took them, it could be more difficult to pursue ends in such a world due to a relative paucity of sources of good information. Does this fact betray a perfect or an imperfect duty not to act in this way?[38] What matters is whether the universalizability test is failed because exercising agency, pursuing ends, is rendered impossible in such a world (detecting a perfect duty), or simply impaired (detecting an imperfect duty). My diagnosis is that one's agency is simply impaired. Indeed, it wouldn't even be such a very difficult world to navigate, it would just be one in which folks were slightly less reliable sources of information than we might desire. So, the failure of this maxim to universalize shows only that there is an imperfect duty not to believe things on a whim. But such an imperfect duty isn't enough to do justice to our intuition that it is never epistemically permissible to believe on a whim regardless of the consequences; the intuition is not just that we have a duty to adopt not believing things on whims as an end (and thus to sometimes and to some extent not believe on a whim).

Similarly, in FAIRY GENERATOR, Hope adopts a belief-forming process on a whim and the same point applies: willing that this maxim were a universal law of belief formation process adoption would not be willing a world in which one's ability to achieve one's ends was completely destroyed, but only severely diminished. So again the failed universalizability test betrays only an imperfect duty. However, again, an *imperfect* duty not to adopt belief formation processes just doesn't do justice to our intuitions. It is never *epistemically* okay to do this; intuitively, it is not just that we have a duty to adopt not adopting belief forming processes on whims as an end.

In short, taking Kant's FORMULA OF UNIVERSAL LAW as our model for an epistemic deontology doesn't look promising in the same way as taking FORMULA OF HUMANITY as our model. Working just with the FORMULA OF UNIVERSAL LAW, it is less obvious how we can handle the kind of trade-offs that cause problems for the

[37] See n.33. Some perfect duties show up in contradiction in conception, such as to not make false promises, but others might show up in contradiction in the will if universalization of the relevant maxim involves willing one's complete compromise as a rational agent.

[38] Cohen's line is likely to be different. Concerning a maxim not to believe testimony, she says "we cannot possibly renounce others' cognitive contribution. As far as cognition is concerned, no one can get by alone" and "Knowledge is by nature a collaborative task, and renouncing others' cognitive contribution would amount to renouncing the whole of human knowledge all together, which I cannot possibly will to do" (325). But I disagree. It is surely not *impossible* to will the renouncing of the whole of human knowledge. More pertinently, it is surely not the case that by willing the renouncing of all human knowledge one would be willing one's complete compromise as an agent. I don't deny that our rational lives are deeply social in the sense that Cohen is emphasizing. Our ability to pursue our purposes *would* be severely impaired were we to limit our epistemic inputs to those that originate solely in ourselves. But the compromise is not complete and so the duty derived not perfect.

epistemic consequentialist in a satisfactory way. This is why I think it is the FORMULA OF HUMANITY, that provides a promising model for an epistemic deontology that can step in to solve those problems.

5 Conclusion

Epistemic deontology has been presented with an opportunity. Recent arguments suggest a critical problem with one of epistemic deontology's key competitors. Berker, in particular, argues that epistemic consequentialism fails to deal appropriately with certain trade-offs and, more importantly, this failure has been argued to signal a deeper failure of a consequentialist analysis of epistemic normativity. This is an opportunity for epistemic deontology because the deontologist is well-placed to entertain absolute constraints against particular kinds of epistemic manoeuvrer, and perhaps the relevant kinds of epistemic trade-off. The main challenge, if an epistemic deontology is to be defended as a serious alternative to epistemic consequentialism, is not only to say the right things about the relevant trade-off cases, but to do so in a well-grounded way rather than a cheap ad hoc fashion. This paper has argued that an epistemic deontology on the model of the humanity formula of Kant's categorical imperative is well-placed to do this. The great promise of Kant's approach to normativity is to provide a compelling story as to why epistemic norms would be binding and it seems that using the formula of humanity we can derive epistemic duties not to perform the relevant kinds of trade-off. I've suggested that this promise is not replicated by the formula of universal law; the formula of universal law doesn't seem to be useful model for epistemic deontology since consequentialist policies don't fail universalizability tests in any obvious way. This means that, regardless of whether the two formulations are ultimately equivalent at some deeper level, the formula of humanity is the more useful model. Obviously, the promise of any epistemic deontology on the model of Kant's categorical imperative is dependent on the promise of the Kantian attempt to ground facts about normativity in the nature of rational agency or humanity. But it is beyond the scope of this paper to provide a full defence of this project.

Beyond the main conclusions of this paper, there are a couple of more general points that might be worth taking away from the discussion in this paper. First, future development of Kantian ethics of belief shouldn't neglect the FORMULA OF HUMANITY. We've seen that the FORMULA OF UNIVERSAL LAW isn't a helpful model for providing a successful treatment of epistemic trade-offs.[39] And second, epistemic trade-offs are an important testing ground for further attempts to develop a Kantian account of epistemic normativity. The potential shortcomings of the model of the FORMULA OF UNIVERSAL LAW and universalizability tests were not apparent before we considered trade-off cases.[40]

[39] Although, see n.5.

[40] An audience member encouraged me to say something on another matter. Can't the consequentialist take some comfort from the fact that so much wrangling is required from an epistemic deontology on the model of Kant's ethics to accommodate intuitions about epistemic trade-offs? Another response in the same vein asks a different question: Shouldn't the wrangling required be taken as a sign that it is wrong to treat being able to accommodate such intuitions as a constraint on satisfactory theory in this domain? Perhaps we should

Acknowledgements Acknowledgements due to all those whose questions and comments have helped shape this paper, particularly Alix Cohen, Aimie Hope, and Rupert Read, as well as various audience members at a talk in Bristol, and students in my epistemology classes. Thanks also to the reviewers whose comments and suggestions were invaluable.

References

Alexander, L. and M. Moore (2016). Deontological ethics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2016 ed.). Metaphysics Research Lab, Stanford University.

Alston, W. P. (1988). The deontological conception of epistemic justification. *Philosophical Perspectives, 2*, 257–299.

Andow, J. (2017). Do non-philosophers think epistemic consequentialism is counterintuitive? *Synthese, 194*(7), 2631–2643.

Andow, J. (2017). Epistemic consequentialism, truth fairies and worse fairies. *Philosophia, 45*(3), 987–993.

Baron, M. (1988). What is wrong with self-deception? In B. P. McLaughlin & A. O. Rorty (Eds.), *Perspectives on self-deception.* University of California Press.

Berker, S. (2013). The rejection of epistemic consequentialism. *Philosophical Issues, 23*(1), 363–387.

Berker, S. (2015). Reply to Goldman: Cutting up the one to save the five in epistemology. *Episteme, 12*(2), 145–153.

Borgoni, C. (2015). Epistemic Akrasia and mental agency. *Review of Philosophy and Psychology, 6*(4), 827–842.

Borgoni, C., & Luthra, Y. (2017). Epistemic Akrasia and the fallibility of critical reasoning. *Philosophical Studies, 174*(4), 877–886.

Buckwalter, W. (2019). Theoretical motivation of ought implies can. *Philosophia, 48*(1), 83–94.

Buckwalter, W., & Turri, J. (2015). Inability and obligation in moral judgment. *PLoS ONE, 10*(8), e0136589.

Chignell, A. (2007). Kant's concepts of justification. *Noûs, 41*(1), 33–63.

Chituc, V., Henne, P., Sinnott-Armstrong, W., & De Brigard, F. (2016). Blame, not ability, impacts moral ought judgments for impossible actions: Toward an empirical refutation of ought implies can. *Cognition, 150*, 20–25.

Clifford, W. K. (1877). The ethics of belief. In T. Madigan (Ed.), *The ethics of belief and other essays* (pp. 70–97). Prometheus Books.

Coates, A. (2012). Rational epistemic Akrasia. *American Philosophical Quarterly, 49*(2), 113–124.

Cohen, A. (2014). Xiv-Kant on the ethics of belief. *Proceedings of the Aristotelian Society, 114*(3), 317–334.

Cohen, Y. (2018). An analysis of recent empirical data on ought implies can. *Philosophia, 46*(1), 57–67.

Cusimano, C., & Goodwin, G. P. (2019). Lay beliefs about the controllability of everyday mental states. *Journal of Experimental Psychology: General, 148*(10), 1701–1732.

Darwall, S. L. (1988). Self-deception, autonomy, and moral constitution. In B. P. McLaughlin & A. O. Rorty (Eds.), *Perspectives on self-deception.* University of California Press.

Dunn, J., & Ahlstrom-Vij, K. (2017). Is reliabilism a form of consequentialism? *American Philosophical Quarterly, 54*(2), 183–194.

Footnote 40 continued
focus on *explaining away* intuitions about trade-offs rather than treating them a symptom of a major problem with accounts of epistemic norms that can't forbid the right trade-offs. If I were a consequentialist, I'd press the Kantian in the latter way. For considerable wrangling seems to be required on the consequentialist side of things in order to say sensible things about trade-offs too. But, in any case, I agree that the fact one's account involves wrangling is a theoretical cost to be borne in mind.

Elstein, D., & Jenkins, C. I. (2019). The truth fairyand the indirect epistemic consequentialist. In N. Pedersen & P. Graham (Eds.), *Epistemic Entitlement.* Oxford University Press.
Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review, 5*, 5–15.
Goldman, A. I. (1979). What is justified belief? In G. S. Pappas (Ed.), *Justification and knowledge: New studies in epistemology* (pp. 1–23). Springer.
Goldman, A. I. (2015). Reliabilism, veritism, and epistemic consequentialism. *Episteme, 12*(2), 131–143.
Greco, D. (2014). A puzzle about epistemic Akrasia. *Philosophical Studies, 167*(2), 201–219.
Hadisi, R. (2022). Kant's account of epistemic normativity. *Archiv für Geschichte der Philosophie Online first*.
Heil, J. (1983). Doxastic agency. *Philosophical Studies, 43*(3), 355–364.
Henne, P., Chituc, V., De Brigard, F., & Sinnott-Armstrong, W. (2016). An empirical refutation of ought implies can. *Analysis, 76*(3), 283–290.
Hill, T. E. (1971). Kant on imperfect duty and supererogation. *Kant-Studien, 62*(1–4), 55–76.
Horowitz, S. (2014). Epistemic Akrasia. *Noùs, 48*(4), 718–744.
Kant, I. (1785/2005). *Groundwork for the Metaphysic of Morals*. www.earlymoderntexts.com.
Kissinger-Knox, A., Aragon, P., & Mizrahi, M. (2018). "Ought implies can, framing effects, and empirical refutations. *Philosophia, 46*(1), 165–182.
Korsgaard, C. M. (1985). Kant's formula of universal law. *Pacific Philosophical Quarterly, 66*(1–2), 24–47.
Korsgaard, C. M. (1986). The right to lie: Kant on dealing with evil. *Philosophy & Public Affairs, 15*(4), 325–349.
Korsgaard, C. M. (1988). Two arguments against lying. *Argumentation, 2*(1), 27–49.
Korsgaard, C. M. (1996). *Kant's formula of humanity. In creating the kingdom of ends*. Cambridge University Press.
Korsgaard, C. M. (1996). *The sources of normativity*. Cambridge University Press.
Kurthy, M., Lawford-Smith, H., & Sousa, P. (2017). Does ought imply can? *PLoS ONE, 12*(4), e0175206.
Leben, D. (2018). In defense of 'ought implies can'. In T. Lombrozo, J. Knobe, and S. Nichols (Eds.), *Oxford studies in experimental philosophy* (Vol. 2).
Martin, M. W. (1988). Self-deception and morality. *Philosophical Review, 97*(3).
Neta, R. (2018). Evidence, coherence and epistemic Akrasia. *Episteme, 15*(3), 313–328.
O'Neill, O. (1985). Between consenting adults. *Philosophy and Public Affairs, 14*(3).
O'Neill, O. (1990). *Constructions of reason: explorations of Kant's practical philosophy*. Cambridge University Press.
Owens, D. (2002). Epistemic Akrasia. *The Monist, 85*(3), 381–397.
Papish, L. (2018). *Kant on evil, self-deception, and moral reform*. Oxford University Press.
Phillips, J., & Cushman, F. (2017). Morality constrains the default representation of what is possible. *Proceedings of the National Academy of Sciences, 114*(18), 4649.
Rescher, N. (1999). *Kant and the reach of reason: Studies in Kant's theory of rational systematization*. Cambridge University Press.
Roberts, P., Andow, J., & Schmidtke, K. A. (2018). Lay intuitions about epistemic normativity. *Synthese, 195*(7), 3267–3287.
Sylvan, K. (2020). An epistemic non-consequentialism. *The Philosophical Review, 129*(1), 1–51.
Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal, 94*(6), 1395–1415.
Turri, J. (2017). How ought exceeds but implies can: Description and encouragement in moral judgment. *Cognition, 168*, 267–275.
Turri, J., Rose, D., & Buckwalter, W. (2018). Choosing and refusing: Doxastic voluntarism and folk psychology. *Philosophical Studies, 175*(10), 2507–2537.
Vahid, H. (2016). Epistemic normativity: From direct to indirect epistemic consequentialism. In M. Grajner & P. Schmechtig (Eds.), *Epistemic reasons, norms and goals* (pp. 227–248). De Gruyter.
Wallace, R. J. (2001). Normativity, commitment, and instrumental reason. *Philosophers Imprint, 1*(3), 1–26.
Wood, A. W. (2002). *Unsettling obligations: Essays on reason, reality, and the ethics of belief*. CSLI Lecture Notes (CSLI- CHUP) Series: CSLI Publications.