NYU | LAW

New York University School of Law

October 12th, from 4-7 pm
Lester Pollock Room, FH, 9th Floor

Colloquium in Legal, Political, and Social Philosophy

Conducted by
Jeremy Waldron and Liam Murphy

Speaker: **David Enoch, University of Oxford & The Hebrew University of Jerusalem**

Paper: **Nudging and Autonomy**

Colloquium Website: http://www.law.nyu.edu/node/22315

A note for NYU Colloquium participants, Sep 2023:

I'm attaching two (related) papers, which together are too long. So:

The nudging paper is a real, forthcoming paper. The epistemic autonomy paper is very much work in progress.

Please read the nudging paper (though you should feel free to pick and choose among the different subsections of the long section 6).

As for the epistemic autonomy draft: The most important parts to read (including for a discussion that will focus on the nudging paper) are section 1 (setting out the problem), section 5 (the pluralism and incommensurability explanation), and section 6 (on autonomy as a state that is essentially a by-product). The rest of the paper is probably less of interest to those not interested specifically in the epistemic case.

Thanks,

David

How Nudging Upsets Autonomy

David Enoch*

Everyone suspects – perhaps knows, but at least suspects – that nudging offends against the nudged's autonomy[1]. But it has proved rather difficult to say why. In this paper I offer a new diagnosis of the tension between even the best cases of nudging and the value of autonomy. If true, this diagnosis improves our understanding of nudging, of course, but it also improves our understanding of the value of autonomy. And while this diagnosis on its own falls short of identifying the moral status of different instances of nudges – which are wrong and which aren't – it goes some way towards doing so, and also shows what more by way of input is needed for determining the moral status of particular nudges.

After quick reminders about the value of personal autonomy (in section 1) and nudging (in section 2), I present (in section 3) a distinction between two values in the vicinity of autonomy – that of non-alienation and that of sovereignty. This distinction – motivated by considerations having nothing to do with nudging in particular – is then employed (in section 5) in order to suggest my diagnosis: Nudging offends against the ideal of personal autonomy not because it offends against either sovereignty or non-alienation, but because it severs the appropriate connection between the two. I present this account first, in section 5, somewhat roughly, and then discuss some more details in section 6. Thus, it emerges that the full ideal of personal autonomy includes sovereignty, non-alienation, *and* something about how the two are related in specific cases. In this respect, the ideal

* For comments on earlier versions I thank Mitch Berman, Monika Betzler, Daniel Brudney, Hanoch Dagan, Hasan Dindjer, Kim Ferzan, Tweedy Flanigan, Chaim Gans, Jonathan Gingerich, Kate Greasley, Till Grüne-Yanoff, Scott Hershovitz, Matt Kramer, Shai Lavi, Dani Levitan, George Letsas, Christian Löw, Ofer Malcai, Eliot Michaelson, Ittay Nissan-Rozen, David Plunket, Assaf Sharon, Saul Smilansky, Levi Spectre, Pär Sundström, Doron Teichman, Laura Valentini, Alec Walen, Benjamin Young, Eyal Zamir and two referees for *The Journal of Philosophy*. I presented a very early version of this paper at the Hebrew University Faculty of Law colloquium, and then later versions at LMU, Tel Aviv, the Analytic Legal Philosophy Conference, and as one of the Burman Lectures at Umeå. I thank the participants for these discussions.
The research for this paper was supported by the ISF grant 1236/21.

[1] Well, *pretty much* everyone. Sunstein seems to deny this (see, for instance, his very brief reply to Waldron (2014b)). For a survey of objections to nudging, and many references, see Hansen and Jespersen (2013), 4-5.

of personal autonomy structurally resembles that of knowledge, according to some (virtue-epistemological) accounts. I find it helpful to quickly present the relevant structure – utilizing the epistemic analogue – in a separate section (4), just preceding the presentation of my positive account. The account of the tension between nudging and autonomy (in sections 5 and 6) leaves the question of wrongness open. In the concluding section (7) I briefly comment on the conditions in which nudging is pro-tanto wrong.

1. The Ideal of Personal Autonomy

An autonomous life is, other things being equal, a better life, or so I here assume, together with many, many others[2]. That is, a life which is shaped, to a considerable extent, by the values and choices of the person whose life it is is, other things being equal, for this reason better than a life that lacks such self-directedness. This ideal—metaphorically, the ideal of being a part-author of one's life-story, rather than merely the passive protagonist in it[3]—is powerful and important, both ethically and (in particular) politically. And it is, of course, absolutely central to the liberal tradition.

The ideal of personal autonomy applies also locally – not as a way of evaluating whole lives, but more specific segments or parts of lives, or even specific actions, decisions or choices[4]. And again metaphorically – before details are filled in and intuitions precisified – there is value in, say, being the author of one's career decisions, or of one's choices about relationships, rather than blindly following authority about such things, or merely drifting into relationships and careers, without anything worth thinking of as self-authorship here.

[2] I borrow here a few lines from my "Autonomy as Non-Alienation, Autonomy as Sovereignty, and Politics" (2022).

[3] This metaphor comes from Raz (1986, p. 369). But see also—in a related context—Christman's (2009, p. 9) literary critic metaphor. The word "part" in "part-author" is important. All of our lives and life-stories are shaped in numerous, deep ways by circumstances that are not under our control. Even given this obvious fact, though, there's a difference between a part-author and a mere protagonist.

[4] There is some discussion in the literature about the relation between global and local autonomy (see, for instance, Oshana (2006, Chapter 1)). We need not worry about it here.

More precise accounts of the value of autonomy differ about the details, and below there will be more on this. For now, though, we can settle for the pretty standard intuitive presentation above, together with the following three observations.

First, autonomy is a graded concept. Lives, parts of lives, specific decisions or actions can be more or less autonomous. It's not as if all of these can either be autonomous or fail to be autonomous. They can suffer from autonomy-deficits to different degrees. Although this feature of the value of autonomy is sometimes neglected, the observation that autonomy comes in degrees is in no way new, and I think (and hope) that it's becoming rather standard in the literature[5].

Second, autonomy is typically thought of both as a value, that is, as partly constitutive of the good life (as above), and as generating constraints on intervention – from the state and from other agents. Our discussion here will for the most part be restricted to autonomy as a value.

Third, in thinking, even initially, about autonomy, it may be helpful to think of the clearest examples of offenses against the value of autonomy. What, in other words, do autonomous lives and autonomous actions and choices stand in contrast to? I can choose, say, to stay at my job rather than quit to do something else entirely perfectly autonomously. But if I so choose because I am coerced to do so (with a threat of physical violence, say, to me or my loved ones), then my staying at my job suffers from a very serious autonomy deficit. Similarly, if my employer, wanting me to stay at my job, makes sure that other options are unavailable, this makes my decision to stay at my job much less autonomous, perhaps at times simply non-autonomous. If I am misled about the nature of the decision to stay (or about the nature of alternative options), this too compromises my autonomy here. And perhaps also – though this is tricky, and will be relevant below – if I am manipulated into staying at my current job this compromises my autonomy. Thus, coercion (and threat thereof),

[5] See, for instance, Meyers (1987, 625), Oshana (1998, 93), Stoljar (2014, throughout). For Killmister (2018) it's a central claim that autonomy comes in degrees along (four) different dimensions.

narrowing down options, deception, and (perhaps some kinds of) manipulation are the paradigmatic ways in which autonomy may be compromised. There may be others as well[6].

2. Nudging[7]

Nudges are non-coercive interventions in choices, by way of shaping the circumstances in ways that are known – from behavioral psychology – to affect people's behavior. The interesting cases for our purposes here are *paternalistic* nudges[8], that is, nudges used to benefit the one being nudged. The standard examples are, by now, well, pretty standard: If we have empirical evidence showing that people overwhelmingly tend to stick with the default option, and also that people tend to under-save for retirement, we may make it the case that a decent retirement savings scheme is the default – so that they don't have to opt in to get it, but rather have to opt out if they choose not to. Such an intervention is in no way coercive, nor does it restrict people's options – they can still opt out, they merely need to check a box on the relevant form (so that the cost of exercising the choice, we can safely assume, is negligible). And while we can't predict with much confidence whether some specific person will stay with the default or choose to opt out, we *can* predict, with *considerable* confidence, that merely changing the default in this way will result in many more people saving adequately for their retirement[9]. And if we know that people tend to choose those dishes in the cafeteria that are placed roughly at eye-level, and that people often choose food that is not good for them, perhaps food that they themselves would agree is less good for them, we may nudge them in the prudent direction by placing the salads at eye-level, and the fat-rich carbs elsewhere. This too

[6] Worries about autonomy arise also upstream from the relevant preferences, in the literature on adaptive preferences, and indeed, on false consciousness. See my "False Consciousness for Liberals: Part I" (2020), and the many references there.

[7] There is now a huge literature on nudging. The wave starts (pretty much) with Sunstein and Thaler *Nudge* (2008). For a helpful recent survey of the normative issues surrounding nudging, see (Schmidt and Engelen, 2020).

[8] They are the interesting cases for our purposes here because in them, the value of autonomy is especially central. In non-paternalistic nudges other considerations may take center stage. See Zamir and Teichman (2018, 177-178) for the claim that the relation between nudging and paternalism is not in general that strong.

[9] Or so, at least, I am here assuming. Reality may be more complicated. See Bubb and Pildes (2014).

will not restrict their liberty or amount to coercion – they can, after all, choose the fries, at negligible further cost (looking a bit down). And we cannot predict with any confidence how this will affect the choice of a specific diner at a specific lunch. But we *can* predict rather confidently that it will significantly increase the number of salads chosen.

Nudges, it is safe to assume, work[10]. Sure, neither always nor necessarily, and it may be a good idea to use more fine-grained information in order to use them well. But when used well, I shall assume, they work, and as even just the examples above show, they can bring about importantly positive results for all involved, all without restriction on liberty. And yet, they come with a strong sense of liberal discomfort. The important point is not about knee-jerk objections from the right to anything too government-looking (going so far as to declare Sunstein, at some point, "the most dangerous man in America"[11]) – the discomfort runs much deeper than this. Perhaps a part of it is due to worries about abuse of power, or perhaps to the unpleasantness of the thought that some people know what's good for me better than me[12]. But even this can't be the whole story, because, first, the intuitive discomfort survives assuming away worries about abuse of power; and second, because sometimes other people do know better than me what's good for me, and nudges seem problematic even when they do, and furthermore, because nudges seem problematic even in cases when I myself confess to be, say, weak-willed, so that knowing what's good for me is not an issue at all (I don't deny that the salad is better for me than the fries)[13].

It seems clear that at least a part of what explains the intuitive discomfort nudges give rise to – perhaps especially in liberals – is due to the sense that they typically offend against the nudged's autonomy. You can choose a retirement savings plan autonomously – perhaps by getting

[10] Although that too – certainly at such a high level of generality – is controversial. See, for instance, Maier et al (2022) and Della Vigna and Linos (2022).
[11] Waldron (2014) quotes Glenn Beck from a back cover of one of Sunstein's books.
[12] This is a line emphasized, for instance, by Waldron (2014).
[13] In the general (not necessarily libertarian) paternalism literature there is a view that ties the distinct wrongness of paternalism to problematic beliefs about the paternalized' s competence. See Quong (2011, 80). I reject such views in my "What's Wrong with Paternalism?" (2016). And for a broader critique of currently fashionable views that tie the epistemic and the moral too closely together, see our "There Is No Such Thing as Doxastic Wrongdoing" (forthcoming).

all the information and rationally taking it into account, or even by seeking the advice of experts and then making your own decision partly based on this advice. But if you end up with a specific – even advisable – savings plan merely because someone at nudging-central-command made sure that's the default option, you may be better off in terms of resources after retirement, but you are not here the poster child of the value of personal autonomy. And the nudger, while they have not restricted your liberty, have nonetheless fallen short of an ideal of fully treating you as an autonomous person[14].

If you don't see the initial force of this intuition, think about nudging not in political or institutional contexts, but in close interpersonal relationships. Suppose you and your partner need to decide about a vacation destination. And suppose you know that your partner tends to be more concessive on such things after a sufficiently good meal. So you make sure the topic only comes up after a really good dinner, and she agrees to your favorite vacation option. And suppose further that it's much less likely that she would have agreed had the topic come up before dinner. Now, there are delicate things to be said about the example, and I revisit some of them later on. But for now what's important to see is that regardless of other complicating factors[15], and even regardless of the overall moral permissibility or impermissibility of your behavior here, the interaction falls far short of the ideal of respecting her personal autonomy. If you're not yet sure you see a flaw here, just think about a relationship in which almost all interactions are of this kind – are you still not sure such a relationship falls short of the ideal of respecting each other's autonomy[16]?

[14] Here, for instance, is Luc Bovens (2008) in one of the earliest philosophical discussions of nudging: "There is something less than fully autonomous about the patterns of decision-making that Nudge taps into. When we are subject to the mechanisms that are studied in 'the science of choice', then we are not fully in control of our actions."

[15] For one thing, in the example as stated your intervention is not paternalistic. So let's add the assumption that your partner too will better enjoy the vacation you are suggesting, more so than the options she may have preferred pre-dinner. Also, as always with nudging, there are questions about the default – after all, it's not as if there's something special about pre-dinner discussions. And it's not as if you offend against your partner's autonomy if you fail to bring up issues at the moments where they are least likely to be concessive.

[16] Closely related here is also a rationality ideal – in particular, the relational ideal of interacting with others as rational, at least in the sense of capable of responding to reasons. I'm not sure what exactly the relations are between autonomy and rationality, but I'm sure there are some such close relations. I thank Oren Bar-Gill for relevant discussion.

All of this, though, remains on the intuitive level, where it does seem clear that nudging offends against (something in the vicinity of) the value of autonomy. But it has proved hard to back this up on a more reflective level. Recall the paradigmatic ways of violating autonomy: First, there's coercion. But at least in the cleanest cases of nudging, an accusation of coercion cannot stick. By making the retirement-saving-scheme the default, no one is coercing you in any way – you are perfectly free to check the "opt-out" box, at no additional costs to you. No valuable option for you is taken off the table, no threat is issued. And while some cases of nudging involve deception, many do not: Again, in the retirement-savings-scheme case, the employer (or the State) may be fully clear and explicit about shifting the default in this way, as can the cafeteria owner in that cafeteria case. But the air of an offense against autonomy remains even in these cleaner, transparent cases of nudging (to which I return below). So it's not about deception or anything of the kind. Manipulation – another standard violation of autonomy – is a harder case, because there does seem to be something manipulative about nudging. The problem, though, is that an account of how it is that manipulation offends against autonomy is not much easier to find than one about nudging[17] (and arguably, the account I'm going to end up suggesting may be applied to manipulation more generally).

So thinking about the paradigmatic ways in which autonomy is sometimes violated does not help in vindicating and explaining the sense that nudging offends against the value of autonomy. And a helpful recent survey (Schmidt and Engelen 2020) also shows how even on several different understandings of autonomy, it's not clear that nudges – certainly not all nudges – are problematic from the point of view of the value of autonomy. Before we give up on the intuition that nudging does come with an autonomy-deficit, then, a further diagnostic effort is called for.

3. Autonomy as Sovereignty and Autonomy as Non-Alienation

[17] For a helpful overview, see Noggle (2022).

Let me introduce a distinction between two autonomy-values. The distinction will be relevant, first, in showing just how hard it is to come up with a diagnosis of how it is that nudging offends against autonomy, but ultimately, also in offering such a diagnosis. The distinction is between autonomy as sovereignty and autonomy as non-alienation[18]. Perhaps the best way of introducing the distinction is by examples.

Think of paradigmatic weakness-of-will cases. Suppose I have a dieting policy to which I am deeply committed. It is motivated, say, by health concerns, by my deep desire to stay alive and reasonably healthy, by my love of the people I care most about and in whose lives I want to continue playing a role for a while, by commitment to my on-going intellectual projects, etc. But suppose that at the presence of some fancy dessert, I once again succumb to temptation. This choice of mine – while not coerced, of course, and while being at least in some senses perfectly my own – does not manifest the value of autonomy to its full extent. *Why* this is so is a question we don't need an answer to right now. But *that* this is so is clear on intuitive grounds. A conception of autonomy that fails to respect this intuition will be deeply flawed for this very reason. The problem here seems to be that – despite me being in control (it's not as if anyone is forcing the dessert down my throat or threatens me with unwelcome consequences if I don't have it) – still there is a (local) tension here between how my life goes and my deep commitments. One value in the vicinity of autonomy, then, is the value I call *non-alienation*, that is, the value of shaping one's life according to one's deep commitments[19].

---

[18] As far as I know, the distinction first appears explicitly in the literature in Brudney and Lantos (2011). I re-introduce it in my "Hypothetical Consent and the Value(s) of Autonomy" (2017) and "False Consciousness for Liberals, Part I" (2020), and develop it more explicitly in "Autonomy as Non-Alienation, Autonomy as Sovereignty, and Politics" (2022). When I wrote those papers, I had not been familiar with Brudney and Lantos (2011). I want to thank Ben Schwan for bringing their paper to my attention, and to take this opportunity to set the record straight – while Brudney and Lantos use a different terminology, write in the bioethical context, and do not develop the distinction in detail, still that distinction is very much present in their paper. (The distinction – or something close to it – appears also in Valdman (2010), but Valdman is concerned to argue that (his analogue of) sovereignty is not of value at all. I differ, of course – and I challenge anyone siding with Valdman to watch the following video, and insist that the hostess's autonomy is in no sense compromised here: https://www.youtube.com/watch?app=desktop&v=sCX_TcKDr4w. I thank Doron Teichman for drawing my intention to this *Curb Your Enthusiasm* clip.

[19] I don't need – and don't want – to commit here to a specific view of what it is for something to be a deep commitment of mine. I like thinking about such things in terms of commitments that are endorsed by higher-

Now suppose that my son – knowing how weak-willed I tend to be in such situations – takes the dessert away. Clearly, he is now increasing the extent to which my life goes according to my deep commitments. But – at least if I proceed to insist, indeed, to assert my autonomy – there is a clear sense in which he is offending against my autonomy by taking this option, well, off the table. The problem is that he deprives me of my ability to control the situation. It is no longer me who has the final word here, it is no longer my say that determines how things proceed (with regard to whether or not I have that dessert). So while my son here doesn't offend against my non-alienation – indeed, he is actively promoting it – still he offends against my ability to control how my life goes. He offends against the value I call *sovereignty*.

As even just these examples show, a full account of the value(s) of autonomy must accommodate both sovereignty and non-alienation – none is eliminable without loss in our understanding of the self-authorship intuitions that underlie talk of autonomy. Of course, this observation leaves a lot more to be said – how are sovereignty and non-alienation related? Is one of them somehow more basic than the other? Is one of them reducible to the other? Or perhaps both are reducible to some third value?[20] But for our purposes here we can safely ignore these further questions, and focus on the two values themselves. Notice that this distinction comes up in the context of giving an account of informed consent to medical treatment, of offering an account of the normative status of hypothetical consent, of understanding something resembling false consciousness[21]. If this distinction – motivated independently of a discussion of nudging – can help with an understanding of how nudging upsets autonomy, this will of course be especially nice. In fact, it will also serve as some independent confirmation of the significance of the distinction.

---

order ones, and not alienated by even higher-order ones, along roughly Frankfurtean lines, but there's nothing necessary about this picture.

One topic that is relevant – and that I can't discuss here – has to do with the fact that even our deepest commitments are not stable, and that some shifts in them, but not all, may suffer from an autonomy deficit. Some of what I say in "False Consciousness for Liberals" (2020) is relevant here. I thank Jean Thomas for relevant discussion here.

[20] I address these questions in detail in my "Autonomy as Non-Alienation, Autonomy as Sovereignty, and Politics" (2022).

[21] Again see the references in footnote 18 above.

But *does* this distinction help in such a way? If there are really two values in the vicinity of autonomy, then for nudging to offend against the value of autonomy it must, it seems, offend against either sovereignty, or non-alienation, or both. But at least the best cases of nudging do no such thing. To see this, consider the following variation on the cafeteria theme: Suppose that I keep kosher, and that keeping kosher is important to me, but also that I am sometimes weak-willed, especially in the presence of bacon[22]. If you're the cafeteria owner where I have my lunch every day, and you know one or two things about behavioral psychology, you know that people tend to choose items that are in roughly their eye-level. If you then make sure to always place the bacon higher or lower than eye-level, you are nudging me into keeping kosher[23]. But you are not in any significant way offending against my sovereignty – I can still take the bacon if I want to, at no extra cost (in money or anything else). It's not as if – like my son in a previous example – you take the bacon off the shelves, preventing me from pursuing that option if I so choose. It remains entirely up to me whether or not I choose the bacon. Nor do you offend against my non-alienation: In fact, you nudge me in the direction of loyalty to my deep commitments, you intervene (in the nudging kind of way) for, not against, my life being harmonious with my deep commitments. But in this version of the cafeteria case, it remains true that the nudging intuitively offends against the value of autonomy. At the very least, the way the cafeteria owner interacts with me – while possibly benevolent and helpful – does not manifest the full ideal of interaction between autonomous creatures.

So far, then, the distinction between non-alienation and sovereignty makes the need for a diagnosis – in what way does nudging offend against the value of autonomy – more pressing. In order to see how it can nevertheless help with the needed diagnosis, it will be useful to take a page out of virtue epistemologists' book.

[22] I briefly mention this example in my (2022, footnote 47).

[23] Assume I'm strict enough to care about the food being kosher and to not want (to want) to have bacon, but not that I'm so strict as to not eat at a cafeteria that serves bacon. Or, if you find this hard, just pick another example.

4. Interlude: Archers, Achievements, and Virtue Epistemology

An archer succeeds, in at least one sense, when and only when they hit the bullseye. Some archers, of course, are better than others – that is, roughly, they possess whatever it is by way of skill and perhaps other properties that make an archer a good archer. Hitting the bullseye is important. Being a good archer – possessing the archery skills to a sufficiently high level – is important. But these are not the only two important things here. For if the archer, while manifesting good archery skills, hits the bullseye (on this occasion) by fluke, this performance still falls short of the full archery achievement. That achievement requires not just hitting the bullseye *and* manifesting good archery skills, but also hitting the bullseye *because* having manifested good archery skills[24]. The full archery achievement consists of hitting the bullseye by, or because, or in virtue of, manifesting good archery skills.

For many years now, virtue epistemologists have been relying on such analogies in order to offer a (post-Gettier) understanding of knowledge[25]. A belief is in at least one sense successful when and only when it is true. And just as with archery, there are standards of good epistemic housekeeping, or of epistemic justification. And indeed, there are epistemic virtues – the epistemic analogues of archery skills. When one forms a belief, the belief being true is important, as is forming it in a justifiable way, in a way that manifests the epistemic virtues. But these are not the only important things here. For if the believer, while manifesting epistemic virtues, hits on the truth (on that occasion) by fluke, this performance still falls short of the full epistemic ideal – perhaps, of

[24] This is a point about the relevant achievement, not about the motivations of the good archer. It is consistent with the point in the text – and quite plausible, it seems to me – that the only aim the good archer has upon shooting is hitting the bullseye; while shooting, that is, perhaps the archer should not have the full achievement in mind, and should accord absolute priority to the aim of hitting the bullseye. The thought that they should be willing to go for, say, a lower probability of hitting the bullseye just in order to increase the probability of the full achievement described in the text sounds to me objectionably fetishistic, and the point in the text is not committed to it. But we don't need to decide this issue here.

[25] The archer example originally comes from Sosa, but pretty much everyone uses it. For an overview, and many references, see Turri et al (2021).
In virtue epistemology, this kind of point is often tied to talk of the believer deserving credit for their belief (even to the point of suggesting an analysis of knowledge in terms of true belief for which the believer deserves credit). I'm not sure how plausible this is in the epistemological case (see, for instance, Lackey's (2007) influential critique). Anyway, in what follows I don't rely on the credit point at all.

knowledge. That achievement requires not just hitting the truth *and* manifesting the epistemic virtues, but also hitting the truth *because* having manifested the epistemic virtues[26]. Knowledge – the full epistemic achievement, arguably – consists of hitting the truth by, or because, or in virtue of, manifesting epistemic virtues. And you can see how this line of thought is supposed to deal with Gettier cases – the Gettierized believer has indeed hit on the truth, and has indeed manifested epistemic virtues (their belief is justified), but they haven't hit on the truth *because* they manifested the epistemic virtues. The Gettier circumstances do not undermine truth, nor do they undermine justification. They undermine the relation between them that is needed for knowledge.

Naturally, much more needs to be said here. Details have to be filled in – for instance, how should we understand the "because" in the condition that the believer hits the truth *because* having exercised the epistemic virtues; and what *are* the epistemic virtues? And it's not as if anything resembling this virtue epistemological account is consensual – powerful objections have been raised[27], replies offered, revisions suggested, and so on[28]. But for our purposes here we can proceed – for now, at least – without further details: After all, the virtue epistemology case is brought here merely as an analogy to the nudging case (to which we are about to return). So I will mention more details below, only when they are needed or helpful for discussing the nudging case.

Before concluding this epistemic interlude and returning to nudging, then, I want to make just the following two points[29].

First, as many virtue epistemologists emphasize, this structure – where the full achievement requires some objective condition (truth), something about faculties or skills or virtues (justification, perhaps), and the right relation between them, the former because of the latter – this structure is very common. Even if it is not adequate as a fully general account of achievement, it surely fits many cases of achievements (succeeding in fixing the car because of one's excellent mechanical skills

[26] This is a point about the relevant epistemic achievement, not about the motivations of the good believer. The point from footnote 24 applies, mutatis mutandis, to the epistemic case as well.
[27] For one influential criticism, see Lackey (2007).
[28] For a very good survey, and for many references, again see Turri at al (2021).
[29] For both, see the relevant discussion and references in Turri et al (2021).

rather than as a matter of luck; doing the right thing in virtue of being motivated by the morally relevant features of the circumstances; making an excellent contribution to philosophy in virtue of exercising one's philosophical abilities; making a scientific discovery in virtue of exercising one's scientific virtues well; …). And this means that it won't be too surprising if we find this structure elsewhere as well – including, as I'm about to argue in the next section, in the case of understanding the relation between nudging and the value of autonomy.

Second, many virtue epistemologists point out – as one of the main advantages of an account of the kind mentioned – that it seems to give a natural answer to the problem of the value of knowledge[30]: An account of knowledge should explain why knowledge is especially valuable – compared, for instance, to merely true belief. And the problem is that not with all attempted solutions of Gettier's problem, for instance, is this condition satisfied. But the virtue epistemological account sketched above seems to satisfy it especially naturally: There does seem to be something especially valuable in such full achievement, where success is achieved not as a matter of luck but as a function of the skills of the relevant person – and this remains so when the relevant person is a believer, the relevant skills are epistemic, the success is truth, and the achievement knowledge. Just as we see more value in the success of the archer who hits the bullseye because having exercised their excellent archery skills than in that of the skillful archer who nevertheless flukes their way to the bullseye, so too we see more value in the success of the believer who hits the truth because having exercised excellent epistemic skills. So we shouldn't be surprised to find this structure underlying another full value – indeed, perhaps even the full value of autonomy.

5. Nudge and Autonomy Again: The Diagnosis

It is high time to present my diagnosis, then – my suggested explanation of how it is that nudging offends against the value of autonomy. I first present, in this section, the main idea (that at this

[30] Perhaps Zagzebski is especially influential on this. See Pritchard et al. (2022), section 3, and the references there.

point the reader may anticipate) somewhat roughly, then proceeding to some further details and implications (in the next section).

I'm standing there, then, at the cafeteria, considering the (very much non-kosher) bacon. Suppose that I see it, I consider it, and I find the strength to resist temptation and stay loyal to my deep commitments. In this happy case, my choice manifests both sovereignty (it's entirely my call), and non-alienation (I make the decision that is in line with my deep commitments). Furthermore, my decision manifests non-alienation *precisely in virtue of* it manifesting sovereignty. It is because it was my call (and the fact that I made the right call) that the decision also manifested the value of non-alienation. But now suppose that the cafeteria owner makes sure the bacon is not at eye level, and that this fact plays a significant role in explaining why it is that I don't choose the bacon. On this occasion, my choice still manifests sovereignty (it remains entirely my call), and it still manifests non-alienation as well. But it can no longer be truly said that the decision manifests non-alienation *because* it manifests sovereignty. There will be much more about the relevant notion of "because" or "in virtue of" below, but for now: In the happy, self-control case, if you wonder "Why didn't he choose the bacon?" a good answer will be something like "because it was his choice, and he keeps kosher" (perhaps together with "and he showed strength of will this time"). But this will *not* be a good answer in the nudging case. In that case, a good answer will have to refer to the nudging, to the cafeteria owner's (mild) intervention. In the nudging case, what explains why I didn't choose the bacon is – to a considerable extent – the fact that it was placed below eye level by the owner in order to decrease the likelihood of my choosing it. An explanation that neglects to mention this fact is a worse explanation for this neglect[31].

The fact, then, that nudging seems so clearly to offend against the value of autonomy even when it doesn't have the features of paradigmatic violations of autonomy (like coercion or deception), and even when it offends neither against sovereignty nor against non-alienation, lends plausibility to the thought that there's more to the value of autonomy than merely sovereignty *and*

[31] I find somewhat related lines of thoughts in Hausman and Welch (2010, 128) and in Schwan (2022).

non-alienation. And the analysis above – certainly together with the arching and virtue epistemological analogies – renders plausible the thought that what is also needed for the full manifestation of the value of autonomy is the right relation between sovereignty and non-alienation. The value of autonomy is fully manifested in a choice only if it manifests non-alienation *precisely because* it manifests sovereignty.

This diagnosis has the right generality to it. It applies quite naturally to other cases of nudging: In the happy nudge-free case in which I choose an adequate retirement-savings-rate, my choice can manifest non-alienation and sovereignty, and furthermore, it can manifest the former precisely because of the latter. In the case, though, in which I am nudged into an adequate level of savings by an intentional engineering of the default option, I may yet show sovereignty and non-alienation, but the connection between the two will be severed: If you wonder "Why did he save this much for his retirement?" an answer that will not mention the nudge will be inadequate. Similarly, it seems to me, for any other nudging case.

How about the opposite direction? Does this diagnosis overgeneralize? Does it apply to any non-nudging cases? Though it will take too much space to argue for this here, I think that the answer is "no". In cases of coercion, there's loss of sovereignty. In cases of false consciousness (and the like) there's arguably loss of non-alienation[32]. In cases of boosting[33] (done well) the value of autonomy may be fully present (though whether this is so may depend on more details regarding the relevant "because". See below.) Manipulation cases do, I think, often manifest the same problem as nudging cases. But this is not a problem – in fact, it reinforces the initially plausible thought that nudge-cases constitute an interesting sub-set of manipulation cases.

The good extensional fit (something is a nudge iff the above diagnosis applies to it), together with the analogies in the previous section, make for a very strong case, I think, for

[32] See my "False Consciousness for Liberals, Part I" (2020).
[33] Boosting amounts – very roughly – to influencing decisions not by bypassing of rational decision making mechanisms, but rather by boosting them, making them more rational. See Hertwig and Grüne-Yanoff (2017).

this diagnosis. And let me remind you of the two points with which I concluded the previous section: First, I noted there how the structure virtue epistemologists use (an objective success condition, a virtue condition, and a "because" relation between them) arguably applies more widely to many other achievements. Similarly, then, the full autonomous achievement is only present when one acts in accordance with one's deep commitments, when one exercises control or sovereignty, and furthermore, when one does the former in virtue of doing the latter. Second, I noted there that this virtue epistemological story nicely explains the special value of knowledge. As you can imagine, then, I now want to suggest that the analogous structure explains the special value of autonomy fully understood. The full ideal of autonomy consists not just of non-alienation and sovereignty, but also of the special relation between them[34].

This concludes, then, my initial presentation of my explanation for how nudges are antagonistic to the value of autonomy. In light of the previous few sentences, this story also makes good, I think, on the hope that thinking about nudging will help us come up with a better understanding of the value of autonomy[35].

[34] Here's Sosa (2003, p. 174) "We prefer truth whose presence is the work of our intellect, truth that derives from our own virtuous performance. We do not want just truth that is given to us by happenstance, or by some alien agency, where we are given a belief that hits the mark of truth not through our own performance, not through any accomplishment creditable to us." This sounds plausible to me, and it remains plausible when applied to autonomy as in the text.
Let me suggest here – very tentatively – that the relation between the knowledge case and the autonomy case may run deeper than mere analogy. This quote may suggest that something about autonomy is *already* there as a part of an account of achievements in general, and if so, also of knowledge. But I cannot discuss this suggestion here. For some discussions of close themes (but I'm not committed to all the details), see Carter (2022).

[35] Some people (e.g. Waldron (2014)) think that the problem with nudging, or anyway one central problem with it, is that the nudger exploits the nudged's foibles or rational weaknesses. On my account, this is strictly speaking irrelevant. Even if the nudger exploits a way in which the nudged is actually reasoning extremely well, still the nudger's interventions may sever the "because" relation between the nudged's sovereignty and non-alienation. Suppose that your spouse usually tends not to be concessive enough (when it comes to agreeing about joint vacations). In that case, she may be reasoning better after dinner. Still, at least arguably, your nudging her (by making sure the topic only comes up after dinner) offends, to an extent, against the value of her autonomy (though whether this is so may depend on delicate issues about salience; see below.)

6. Details and Implications

But more details are needed. This section puts more flesh on the diagnosis from the previous section by filling in some of the required details and noting some relevant implications.

6.1 The Explanatory "Because"[36]

The full ideal of autonomy, I suggested, includes non-alienation in virtue, or because, of sovereignty. It is this "because" relation that is severed when nudging is in place. But more needs to be said about the nature of this "because".

While this "because" certainly has a causal element, causation cannot be the full story here. This is so, because there are always many causal factors that play a role in bringing about any of our decisions and actions, in any set of circumstances, whether or not nudging is in place. Thus, when I show strength of will in the cafeteria, it's possible that a part of what causally allows me to do so is a kind word from a co-worker earlier in the morning, or something about the temperature in the cafeteria, or indeed, the way the owner placed the dishes irrespectively of any intention to nudge me (perhaps motivated just by the desire to maximize profits). Similarly in the nudging case: Sure, the nudging plays a causal role. But so do many other things, including things about me of the kind that typically play a role in more autonomy-friendly explanations of actions. There's a point here that Sunstein and others often rightly emphasize[37], and about which they are surely right: It's not as if there's some privileged (nudge-free) baseline, one that allows for pure decisions that are not subject to all sorts of causal influences. Our choices are always made within a complex causal nexus, and there's no natural, default way for these to be arranged. It's not as if there's a naturally right place for bacon in the cafeteria, some clear pre-nudging, pre-normative-discussion answer to the question whether the retirement savings arrangement should be opt-in or opt-out. For similar

[36] Discussions with Ben Ohavi and Ofer Malcai were especially helpful on this point.
[37] This is a central claim in Sunstein's *Why Paternalism* (2014a), for instance.

reasons, it's going to be very hard to give an account of the "because" in my diagnosis in purely causal terms.

Return, though, to the underlying intuition. It's the one I put above using why-questions. The relevant distinction is between different answers to such questions as "Why did he not choose the bacon?". And why-questions are typically requests for an explanation (that may be, and often is, causal). So I suggest that we understand the "because" in the requirement that the non-alienation will be manifested because sovereignty or control is manifested as depicting an explanatory relation. In the happy, nudge-free case, what explains my action and the fact that it manifests non-alienation is my sovereignty. In the nudge case, what does the explaining is the nudge. Hence the difference.

Notice that in understanding the "because" as a (causal-)explanatory because, I remain loyal to the virtue-epistemological analogy[38], and indeed to the general view of achievement that comes along with it, for there too the common understanding of the needed relation (between hitting the truth and epistemic virtue) is explanatory.

Of course, once it's clear that the relevant relation is explanatory, it brings with it all the features of explanatory relations. Chief among those is context-dependence. What the appropriate answer is to a why-question depends on the context in which the question is asked[39]. In a context in which the presence of oxygen is taken for granted, an adequate

[38] Indeed, in his discussion of the "because" relation Greco (2003) heavily draws on Feinberg's view of blame. So perhaps the intuitions I use the analogy with virtue epistemology to strengthen have just as strong an origin in the practical domain after all.

[39] It may also depend on a host of epistemic features. Consider the following important complication (I thank Daniel Brudney for drawing my attention to it): The information brought in from behavioral psychology is general and statistical, it doesn't in itself say anything about me, and whether (for instance) I would have resisted the bacon even had it been placed at eye level. At times, then, we cannot know whether the nudges played an indispensable causal role in bringing about the relevant action, and the effect of the nudge is that nor can we know that it didn't. This may suffice to deprive me of the *knowable* achievement (of resisting temptation without the help of a nudge). As long as the salience of explanations is sensitive to such things as well – so that an explanation that fails to refer to the nudge is less good for this fact, even when the information about the nudge's effectiveness is general and statistical – the points in the text here stand. And when we are trying to evaluate more general nudges – as general policies, not as aimed at a specific individual – we can often know that the nudge will be causally efficacious in at least some, often many, cases, even if we can't know in which.

explanation of the fire will refer to the match lighting it. In a context in which the presence of sparks is taken for granted and the background assumption is that there is no oxygen present, a good explanation of fire will refer to the malfunctioning of the system supposed to keep the oxygen out[40]. (In purely causal terms, of course, both a spark and oxygen are necessary conditions for the fire[41]).

What this means in our context is that whether a choice is fully autonomous (in the sense requiring also the explanatory relation between sovereignty and non-alienation) will depend on such contextual features, on whether in that context the nudging-intervention is salient, and so on. Now, it will be convenient to return to this context-dependence at the end of the next subsection, but for now I just want to note that it seems very natural here, and that we should welcome it. Ours is a normative inquiry, about the value of autonomy, and what offends against it. Within such a normative inquiry, some context-dependence of the kind introduced by the explanatory requirement seems precisely the thing to expect[42]. But again, I return to discussing some examples at the end of the next subsection.

6.2 Autonomy, to repeat, Comes in Degrees

[40] This kind of example is common in the literature on explanations, and indeed, in the relevant literature in virtue epistemology. See Greco (2003) (who uses, following Feinberg, a version of the fire example), and the references there.

[41] But the discussion of causation here is also complicated. There are attempts to distinguish – within an account of causation – between normal and abnormal factors, so that in the example in the text, in common contexts, the spark will be abnormal and the oxygen normal. Some philosophers hope that we can distinguish between the causal role played by the abnormal factors, and the merely enabling role played by the normal ones. If so, perhaps what I try to capture in the text in terms of the explanatory because and salience can be alternatively captured in purely causal terms, together with an account of normalcy. See, for instance, Gallow (2022), section 1.2.3. For relevant discussion, and for this reference, I thank Christian Löw.

[42] Some context-dependence is to be expected, but perhaps not *any* context dependence. That is, perhaps we should restrict the relevant contexts, to just those that are somehow relevant for the relevant individual and their autonomy. I think that this can be done in a non-ad-hoc way, but I'm not sure exactly how. I thank Hasan Dindjer for relevant discussion.

Knowledge is arguably yes-no[43]. So it's not surprising that in the virtue-epistemological context, much ink has been spilled on such questions as whether in a specific case – some particular version of a Gettier case, for instance – the believer reached the truth because of their good exercise of their epistemic virtues or skills, or because of (say) luck. If the former, the belief may amount to knowledge. If the latter, not so. This need for a dichotomy creates problems, for often *both* luck *and* skills or the exercise of virtues play a partial role in explaining the fact that the believer has reached the truth, and it's not clear what – in the context of an attempt to give an account of knowledge, dichotomously understood – to say about such cases. Even when at attempt is made to understand things here in a more scalar way – so that it's not *either* luck *or* epistemic virtues that explain reaching the truth, but both, to different degrees – still the desired result is dichotomous, so that a dichotomous distinction is introduced based on the scalar one that does the more basic work, for instance, in terms of either luck or skill being the more salient factor, or the more dominant one[44].

But in this respect we can do here much better than the virtue epistemologists[45]. For autonomy, as already noted, comes in degrees, and we have no need for a dichotomous distinction at any level. Notice that this point – that choices and lives can be more or less autonomous, that autonomy is not *either-or* – is extremely plausible independently of what we end up saying about nudging. Much of the feminist discussion of adaptive preferences, for instance, talks of an "autonomy-deficit" – not declaring such choices *non*-autonomous, but rather *less* autonomous than paradigmatically autonomous ones[46]. The point arises in other contexts as well, and is really perfectly natural and intuitive even pre-theoretically: The thought that autonomy is not all-or-nothing, that there are choices that fall short of the full ideal of autonomy and yet manifest autonomy to a considerable degree – such thoughts are a commonplace. So we can certainly help

[43] But perhaps only arguably. One way of understanding Sosa's (e.g. 2009) distinction between kinds or levels of knowledge is precisely as challenging this point.
[44] See, for instance, Lackey (2007, 348); Carter (2016).
[45] Whether virtue epistemologists themselves can do better – opting for a more fully scalar view – is something I am not sure about. Perhaps they can, at the price of rendering knowledge much less central to their account. I am okay with this, but perhaps not all of them are.
[46] Again see the references in footnote 5 above.

ourselves – without any ad-hoc-ness worry – to a scalar view of the value of autonomy when discussing nudging as well.

And what this means is that we can say such plausible things as that the more salient the nudge is as a part of the explanation (of the achieved non-alienation), the more serious the autonomy deficit the choice suffers from. Similarly in the opposite direction: The more dominant the explanatory role played by the agent's sovereignty, the more autonomous the choice. And this allows us flexibility that renders the account even more plausible. For instance, intuitively it seems that a nudge is more autonomy-challenging the larger the effect shown by the behavioral psychology findings it relies on[47]. If, for instance, we have reason to think that hardly anyone will opt-out of the retirement savings plan, simply because almost everyone almost always goes with the default, then this nudge renders the savings rate rather strongly non-autonomous. And we can easily explain this: In such a case, the achieved non-alienation (saving in a rate appropriate to my deep commitments about my future) is not at all explained by my sovereignty, but almost entirely in terms of the nudge. If, however, the effect of the default bias is rather weak, then the nudge leaves more room for autonomy, and again we can explain this: For then, the explanation will have to invoke the nudge, but the agent's sovereignty will also play a significant explanatory role.

Going scalar can also help with interesting, more challenging cases. For instance, the literature often contrasts nudging with rational persuasion[48]. Thus – an employer can explain to her employees about the importance of a greater rate of retirement-savings, trying to engage their rational capacities and convince them to save more. Nudging, it is often noted, consists in bypassing the nudged's rational capacities, influencing their choices in non-persuasive ways. But the contrast – though often insightful and significant – is not remotely that clean. Suppose, for instance, that I inform my employees about all the reasons they have to save more for retirement, but I make sure

[47] See in this context Kiener's (2021) resistibility condition.
[48] Hausman and Welch (2010, 127) accuse Sunstein and Thaler of classifying cases of persuasion as instances of nudging.

to do so in a deep voice, knowing that people tend to trust more what is said in a deep voice[49]. Am I persuading my employees? Am I nudging them? If they then choose to save more, is their choice autonomous? Going scalar allows us to avoid such moot questions, and say the obvious: Their choice is nudged to an extent, but I also engage in persuading them. The choice they end up making is partly due to their sovereignty, and partly due to the nudging. Which means that their choice does manifest some autonomy, but not as much as it would have but for the deep-voice manipulation[50].

We can now combine the lesson about scalarity with the one about context-dependence and salience from the previous subsection, to show that while all of this apparatus allows us quite a bit by way of flexibility, it doesn't leave the account *too* flexible to be of any value. As I noted above, every choice and action is preceded by a very rich causal history, but the vast majority of its parts are in no way relevant to an explanation of how it is that the action went some way towards non-alienation. (If you ask "Why did he not choose the bacon?", you are unlikely to be happy with the answer "Because his parents met 55 years ago"). And autonomy does not require, of course, *ab initio* self-creation: That there's more to the causal history of our actions than our sovereignty is no threat to our autonomy (in the sense relevant here)[51], nor – as a result – is it a threat to the account suggested here. On the other hand, room is left for very many causal interventions that do result in an autonomy-deficit because they are salient, even if they do not undermine autonomy altogether. Think here of a generalization of the deep-voice-persuasion case above: Using a nice-looking presentation in order to convince your Dean to accommodate some need of the department, being amusing from time to time in class so that your students respond better to the substantive stuff

[49] Needless to say, I have no idea whether this is so.

[50] And there are other ways in which nudging and persuasion may interact: I may persuade my employees to choose the nudging cafeteria rather than the non-nudging cafeteria (by offering them data about akrasia, etc.). I may nudge someone into listening to the rational persuasion (suppose that on travel websites, I make viewing a brief video explaining that people should offset their carbon footprint the default option, which people can opt out of; but the video contains only rational persuasion). And so on. In all such cases – that are not, as far as I know, systematically discussed in the nudging literature – going scalar as in the text makes available very plausible analyses.

[51] This is hardly the place, obviously, for a discussion of free will. Let me just note that what I say here in the text is consistent, as far as I can see, with pretty much any remotely plausible compatibilist story.

that's going on, and so on: It's hard (and maybe also unpleasant) to imagine human life without all of these, and they are often quite salient (if you ask "Why did the Dean again accommodate that department's needs?" the answer "Did you see the Chair's presentation?" seems in place, even if it is never the full story). And what this means is that while autonomy is often manifested, it is rarely manifested to the maximum possible extent. I find this result highly plausible.

6.3 Intention[52]

Compare the nudge version of the bacon-in-the-cafeteria case, with another one, in which the cafeteria owner has no interest in helping me stay loyal to my deep religious commitments but nevertheless places the bacon far from eye level for some other reason, or for no reason at all. The causal influence on my choice remains the same, of course – what matters for that is where the bacon is placed, not what the owner had in mind in placing it there. But intuitively, the nudge case seems to offend against my autonomy in a way that the second case does not[53]. Can this be explained?

The fact that the cases are alike causally but differ in terms of the value of autonomy gives us yet another reason to prefer the explanatory "because" over the merely causal one. And asking about the appropriate responses to why-questions again can help here. In the nudge case, to repeat, if asked "Why did he not choose the bacon?", no answer that will neglect to mention the nudge will be adequate[54]. In the no-nudging case, though, very often the placement of the bacon will not be salient. In those cases, then, there will be no significant autonomy-deficit.

---

[52] I thank Alon Harel, Eliot Michaelson, Hasan Dindjer, and David Plunkett for pressing me on this issue. Hansen and Jespersen (2013) also emphasize the role of intention, but they do things in a very different way, which I do not accept.

[53] Raz (1986, 377) makes a similar point about coercion – that it affects autonomy more than a similar narrowing down of options that is not intentional. His suggested explanation is different from the one I give below in the case of nudges.

[54] Unless, that is, the context is a very unusual one. Perhaps, for instance, in a context in which everyone takes for granted that a lot of nudging is going on, and we're only interested in explaining the difference between the cases in which the nudge works and cases in which it doesn't, the situation is different.

I don't want to pretend that salience or context-dependence is simple. It's not clear – nor is it uncontroversial – how best to fill in the details. And the suggested account incorporates whatever problems come along with such salience talk. But talk of salience is needed for many purposes that have nothing to do with nudging and autonomy, including, of course, for virtue epistemology and for a general understanding of achievements. It would have been surprising if an understanding of how nudging upsets autonomy could be achieved without incorporating salience talk. More general problems about salience call for more general solutions. They do not pose a special problem for my use of salience talk here[55].

Let me mention four more points regarding the relevance of intention here. First, here too the scalarity of autonomy helps. For we do not have to say that either an item on a choice's causal history undermines its autonomy or it doesn't. We can say, for instance, that the intentional nudging in the cafeteria renders the choice (not to have bacon) less autonomous, and that in a specific context even the non-nudging placement of the bacon far from eye level does that – just to a lesser degree.

Second, the point made in this subsection plays a role also in responding to the oft-made point already mentioned earlier in the case – that there is no natural baseline from which nudging is a deviation. If I'm right that the intention to nudge is relevant as explained above, it shows how nudging is special among causal influences on choices, despite there being no natural pre-nudging default or baseline.

Third, what I am insisting on in this section is that the intention of the nudger often makes an autonomy-difference. I am not claiming, however, that such a nudging is always necessary for the relevant intervention to qualify as a nudge or even to offend against the value of autonomy. I would

[55] For a similar claim in the virtue-theoretical case – that reference to salience leaves much more work to be done, but is in no way vacuous – see Greco (2003, 132).

like to keep open the possibility of structural nudges – nudges that are a feature of a social structure, without a necessary connection to any specific agent and their intentions[56].

Lastly, the salience of the nudger's intention may be partly explained by the fact that autonomy is itself best understood as at least partly relational. This is so even if we don't go all the way endorsing a relational account of autonomy[57]. And the significance of the relevant relationship will re-appear in the final section, in discussing the question when nudging is wrong.

6.4 Always and necessarily

On this account, then, while the extent to which a specific case of nudging upsets autonomy varies, the fact that it does upset autonomy does not[58]. Any effective nudge in almost any context becomes a necessary part of an adequate full explanation of how it is that the relevant agent acted as they did, and (when this is the case) how it is that their action manifested non-alienation. Nudging – always and as a matter of necessity – upsets autonomy (to an extent).

You may be worried about this, because the literature mentions cases of nudging where, so some people seem to think, there's no autonomy deficit. Especially relevant here are cases of transparent nudges, self-nudges, and what may be called pre-emptive or counter-nudges. *Transparent nudges* are nudges whose nature is fully disclosed to the nudged[59]. Think of a cafeteria that has a sign at its entrance saying "Welcome to the nudging cafeteria. We've placed salads at eye

[56] I thank Tom Kohavi for a related suggestion. Hausman and Welch (2010) suggest that the problem with nudges is that the authorship of the relevant action now belongs with the nudger, not with the nudged agent. I reject this suggestion: for one thing, there may be more than one author of a relevant action. Relatedly, there may be cases of authorship-by-others that in no way undermines self-authorship. But also – the point in the text here is relevant, for in structural cases of nudges, if there are any, there is no other author. In general, the important question for me is not directly about the involvement of another agent, but about whether it's the agent's sovereignty that explains their achieving non-alienation.

[57] See Mackenzie and Stoljar (2000).

[58] With the sole exception of cases like those discussed in footnote 54 above.

[59] See Kiener (2021) for some discussion of transparency in the context of nudges. For the claim that transparent nudges can be effective, see Bruns et al. (2008) (I thank Eyal Zamir for the reference). Hansen and Jespersen (2013) put a lot of emphasis on transparency, though in a somewhat different (but related) sense to the one in the text. See also their critique of a Rawlsian publicity condition in our context.

level, and the plates are smaller than usual (in the US). If you want, though, you can find the fries just below, and feel free to use two plates." *Self-nudges* are nudges we deploy towards ourselves. I may, for instance, make sure when I bring chocolate to work to leave it at the department office, not in my own. *Pre-emptive (or counter-) nudges* are nudges meant to preempt or defeat other nudges, or other non-rational influences. For instance, the government may employ nudging in order to counteract the effects of (non-rational) advertising[60]. In all of these cases, the account in this paper applies: In all of these, a full explanation of the action and of the non-alienation manifested by it will have to invoke the nudging. So if you think that these cases – or even some of them – are in no way problematic in terms of the value of autonomy, you may think of this as a problem for my account.

But it is not. The main reason is that these cases too manifest autonomy shortcomings[61]. Recall throughout that autonomy comes in degrees – so that recognizing that in these cases too there's an autonomy deficit does not amount to declaring all of them non-autonomous. Let's revisit the cases, then (in reverse order): A governmental response to problematic advertising that fully respects the autonomy of its citizens will surely consist of exposing the manipulation mechanisms employed by the advertising, explaining to people how their effect is to be avoided, and so on. Preemptive nudging may be all-things-considered justified (a point I return to below), but it falls short of the full ideal of autonomy, intuitively understood, certainly when compared to the response just sketched. Similarly, while making sure there's no chocolate in my office may be justified (given what I know about my imperfections), a fuller autonomous achievement would have been to have it in my office, and eat moderately thereby responding directly to the balance of relevant reasons[62]. And while I agree that – and can explain why – typically a transparent nudge will give rise to a lesser

[60] Hausman and Welch (2010, 132) seem to claim that such nudges do not offend against autonomy.
[61] Waldron (2014) accuses Sunstein of being "remarkably tone-deaf to concerns about autonomy". I think he is right in this accusation, partly because Sunstein seems to think of the cases in the text as cases in which nothing by way of autonomy is at all missing.
[62] Bovens (2008, footnote 5) expresses the right kind of suspicion about self-nudges. But he seems to think that joining a self-professed paternalistic company is not problematic at all (in a case that resembles the transparently nudging cafeteria). I would say that such a case falls short of the full autonomy ideal, but that it may very well not be wrong (I get to this in the concluding section).

autonomy deficit compared to a non-transparent nudge (think of the role my decision to enter the transparently nudging cafeteria plays in explaining why I chose the salad), still in the transparently nudging cafeteria *something* by way of autonomy is lost; I am at least somewhat more autonomous in a scenario in which no nudging is going on, and I choose the salad for the right reasons. All of this holds even in cases in which the nudge is all-things-considered conducive to the nudged's autonomy (perhaps by nudging them away from self-destructive choices) – still, the nudging interaction falls short of the full autonomy ideal, in the way here described.

Even in the autonomy-best-case nudging scenarios, then, nudging upsets autonomy. And my account explains why this is so, and also sheds light on the extent to which this is so. This doesn't mean that nudging is always wrong: All it means is that nudging is never *perfect*. There's always something to be said against nudging – namely, that it falls short of the full autonomy ideal[63]. Whether it can nonetheless be justified depends on what else is at stake. I get back to the question when nudges are (even pro tanto) wrong in the last section.

6.5 But: The Good

There's an important difference between the virtue epistemological case and the autonomy case, one that you may think casts doubt on the analogy I have been making much of. As I have been using the analogy, non-alienation was the analogue of truth, sovereignty of justification (or of the exercise of epistemic virtues). But wouldn't a better practical analogue of truth – certainly one with better historical credentials – be The Good? Thus, in a specific case, I may be sovereign, in that my choice may determine things; I may show non-alienation, in that my choice may express my deep commitments; my non-alienation and my sovereignty may be related in the way emphasized

[63] Somewhat more precisely – nudging always and necessarily offends against the ideal of autonomy. Whether this is always and necessarily a reason counting against that instance of nudging depends on whether every manifestation of autonomy is of value and is reason-giving. I don't think I have a view on this. Perhaps, for instance, in some cases in which people don't care about their own autonomy (or in the cases of the kind discussed in Sunstein (2014 (c)) it ceases to be of value, or at least it ceases to give others reasons for action. I thank Saul Smilansky and Eyal Zamir for related points.

throughout this paper. But also, my choice may or may not be *the right choice*, the values I am most deeply committed to may or may not be *of genuine value*. Surely, this matters too. Where we have two factors in the epistemological case (truth, justification) we have three in the practical case (the good, non-alienation, sovereignty). And this, it may seem, renders the analogy less compelling.

In response, let me make the following two points. First, while I agree that there are these three factors doing work in the practical case, it's not obvious that all three are relevant *to the value of autonomy*. Whether they are depends on whether one can autonomously make bad choices. Myself, I think that the answer is (within some constraints) "yes" – I can see value, indeed, the values of sovereignty and of non-alienation, even in cases in which the relevant person's decisions and deepest commitments are substantively wrong. But I do not want to rely on this, so let me just note the relevance of this question here.

Second, there may be room for iterating the structure emphasized in this paper. That is, perhaps the ultimate practical achievement – regardless of whether or not we want to include it as a part of autonomy, or as something over and above autonomy – involves all three of the relevant features, suitably related: Perhaps, that is, the full achievement includes reaching The Good, in a way that aligns with one's deep commitments, in virtue of having sovereignty[64].

Analogies are helpful, when they are, up to a point. I am happy to concede that the relevance of The Good (or some such) makes the analogy I've been relying on between knowledge (a la virtue epistemologists) and autonomy more constrained. But it doesn't do enough to undermine the analogy's usefulness entirely. And indeed, if the suggestion in the previous paragraph can be made good on, the role of The Good may actually be explainable, to an extent, by employing the very structure inspired by that analogy.

[64] As before (see footnotes 22 and 24 above), this point says nothing about the desirable motivations of the agent – perhaps these should be concerned just with the good. I discuss this possibility further in my "Epistemic Autonomy May Not Be a Thing" (Manuscript).

6.6 A Few Problematic Cases

In this subsection I briefly mention some initially problematic cases for my analysis, and indicate how I think they should be dealt with. Because this paper is already long, the discussion will be very quick – it's meant to indicate possible problems and ways of coping with them, not to offer full discussions thereof.

Nudges can occur also when the nudged agent has no relevant deep commitments – think about a cafeteria owner nudging customer to pick one brand of chocolate rather than another, when the customer really doesn't care that much about the difference between the two. In such a case, it's not true that the problem with the nudge is that it severs the explanatory connection between the customer's sovereignty and noon-alienation, for the customer's non-alienation is just not relevant here at all[65]. The thing to say, I think, is that such nudging still offends against autonomy, and that non-alienation is relevant counterfactually – for such nudging would have severed the explanatory relation with non-alienation had it been relevant, had the customer cared about the difference between the two brands.

Relatedly, what should we say of trivial nudges – for instance, using footprints or arrows in order to nudge people into taking the stairs rather than the elevator (Hansen and Jespersen (2013, 21))? Many of these will also be nudges of the kind discussed in the previous paragraph, where the agent has no relevant deep commitments. Here, I want to insist, there is some offense against autonomy. It's just that in such trivial cases the offense is, well, trivial – autonomy is not of great value, perhaps sometimes even none at all, regarding such matters. And of course, there need be nothing wrong in such nudging (as I discuss in the next section).

Doesn't it follow from my account that a good way of making me more autonomous, or allowing me to manifest more by way of the achievement of autonomy, is to make the right or non-alienated decision *harder* for me? Perhaps the cafeteria owner should, then, take care to put the

[65] I thank Shir Nidam, Roy Kreitner and Courtney Cox for this objection.

bacon at eye-level, so that if I still resist temptation, this will be entirely creditable to me? And isn't this absurd[66]? Well, first, I think that autonomy – certainly one's own, but to an extent also that of others' – should not be our aim. It is of value, but values of things that are essentially by-products[67] should not be our aims. Autonomy, very often, should not be pursued, but sneaked-up on. I say much more about this elsewhere[68]. Second, perhaps the phenomenon here is a particular instance of one that occurs with all challenges, and so not a special problem for my analysis here. Third, and relatedly I can imagine contexts in which this result is not absurd at all – perhaps these are the contexts in which it makes sense to accept and give each other challenges. Lastly, even if there is an autonomy-reason for the cafeteria to place the bacon in the most tempting way, this reason may of course be outweighed by other reasons (including ones grounded in my autonomy).

Lastly, suppose you just inform me that the dish I am about to choose contains bacon, and I proceed not to take it (because I'm committed to keeping kosher). An adequate explanation of why it is that I did not choose the bacon-containing-dish is likely to refer to your informing me. Am I committed, then, to such informing offending against the value of autonomy[69]? Let me concede that it would be bad if I were – intuitively, there's no autonomy-problem here. I think I can avoid it, though. First, while the information is relevant to explaining my action, the specific way in which it was delivered is not (and indeed, if it is, then it may be a case of nudging, and an autonomy-problem may be present after all). Second, when you merely give me needed information, you're not bypassing my reasoning skills (as you arguably do in cases of nudging). Rather, you're feeding the information into them. Granted, this is a different condition from the one highlighted in this paper, but then again, the ambition here was limited to offering a diagnosis of what it is that goes wrong (in terms of autonomy) in cases of nudging. That there is more to say about other cases should not be a problem.

[66] For raising objections along these lines, I thank Korbinian Rüger and Jörg Löschke.
[67] In the sense developed by Elster (1983, Chapter 2).
[68] In my "Epistemic Autonomy May Not Be a Thing" (manuscript).
[69] I thank Till Grüne-Yanoff for this case.

7. But Is It Wrong? Some Final Thoughts

Nudging, I've been arguing, always upsets autonomy. It upsets autonomy not because it undermines sovereignty, and not because it undermines non-alienation, but because it undermines the "because" relation between them needed for the full manifestation of the value of autonomy. But even if all of that is correct, it still doesn't follow that nudging is even pro-tanto wrong. Whether it is, in a specific case, depends on whether the autonomy-related reason not to nudge matures into an (at least pro-tanto) duty. And so the important question here – for moral, and certainly for political purposes – is when is there a duty not to nudge? When do nudges wrong the nudged? When is nudging wrong?

The discussion in this paper does not on its own yield an answer to this question. But let me hint at a plausible answer and show how it nicely coheres with the thesis of this paper[70]. I don't think that we owe it to all others, as a general matter, to engage each other on perfectly autonomous terms. But very clearly, this is *sometimes* the case[71]. Sometimes, one agent does owe another precisely that. Recall an example from early on, about the possibility of you raising the question of your joint vacation location only after your partner has had a good meal. What exactly you owe them in this respect seems to be a rather intricate matter, that varies with the specifics of the relationship. In most relationships of this kind, you do owe them at least not to bypass their rational reasoning mechanisms entirely. You do not owe them to not even smile when you raise the issue (even if smiling will make them more favorably disposed). And so on.

A specific case of nudging is pro tanto wrong, I want to suggest, when the relevant relationship includes a duty not to upset the nudged's autonomy in the way (and to the extent) that

[70] A specific case of nudging may be wrong, of course, for other reasons (suppose it involves shifting the location of bacon in the cafeteria, and I, the owner, promised my wife never to do that). But this is not the kind of case we're interested in here. The condition for the wrongness of nudging in the text is limited to when nudging is wrong in virtue of its shaky relation with the value of autonomy.

[71] That nudging is sometimes – but not always – wrong is hardly news. See, for instance, Hausman and Welch (2010).

the specific nudging will. What this means, is that a general account of how nudging upsets autonomy will not, by itself, entail any answer to the question of the moral status of the nudge. For that, much more information is needed, typically about the nature of the relevant relationship. But the fact that nudging upsets autonomy does play a crucial role here – for what we ask about the relevant relationship is precisely whether as a part of it the parties owe each other (or one party owes the other) to engage, in the specific context, in terms that do not fall short of the autonomy ideal in the way that nudging (always and necessarily) does[72]. Notice that this is where, in my view, cases of self-nudges are special – not in their relation to autonomy (because in those cases too, something by way of autonomy is lost), but rather in the fact that they are not (ever, or at least typically) even pro tanto wrong. This also seems to me to be the case with regard to many cases of transparent, consented-to nudges.

Much of the nudging literature is conducted in the political context. Is it wrong, then, for the state to nudge its citizens in all sorts of ways? The question depends, I suggest, on whether – and more plausibly, when – the state owes its citizens to engage them in a way that doesn't offend against the value of autonomy in the way nudging does. And even in those cases – in politics or elsewhere – in which nudging is wrong precisely because of the way in which it upsets autonomy, it is still pro tanto wrongness that has been established. So it remains possible that other considerations outweigh this one, thereby rendering the relevant nudging all-things-considered morally justified[73].

Even in those cases in which this is so – indeed, even in those cases in which nudging is not even pro tanto wrong – we should not lose sight of the way in which it, as a matter of necessity, undermines full autonomy. Nudging may have many advantages, but in assessing its overall moral and political status, we should not ignore its normative shortcomings as well.

[72] I hope to discuss this in more detail in future work. There I hope to also show how the story just sketched in the text nicely generalizes to other cases of what may be called flawed consent – cases of manipulation more generally, coercion, and more.

[73] For an example of the complications that have to be discussed for a fuller assessments, see Teichman and Zamir (2021, 266, and the references there).

Luc Bovens (2009), “The Ethics of Nudge”, In Till Grüne-Yanoff & Sven Ove Hansson (eds.), *Preference Change: Approaches from Philosophy, Economics and Psychology* (Berlin: Springer, Theory and Decision Library A.) pp. 207-20.

Daniel Brudney and John Lantos (2011), “Agency and Authenticity: Which Value Grounds Patient Choice?”, *Theoretical Medicine and Bioethics* 32, 217-227.

Hendrik Bruns, Elena Kantorowicz-Reznichenko, Katharina Klement, Marijane Luistro Jonsson, Bilel Rahali (2018), “Can Nudges Be Transparent and yet Effective?”, *Journal of Economic Psychology* 65, 41-59.

Ryan Bubb and Richard H. Pildes (2014), “How Behavioral Economics Trims Its Sails and Why” *Harvard Law Review* 127, 1593-1678.

J. Adam Carter (2016), “Robust Virtue Epistemology as Anti-Luck Epistemology: A New Solution”, *Pacific Philosophical Quarterly* 97, 140-155.

(2022) *Autonomous knowledge: radical enhancement, autonomy, and the future of knowing* (Oxford: Oxford University Press).

John Christman (2009). *The Politics of Persons: Individual Autonomy and Socio-Historical Selves*. Cambridge. Cambridge University Press.

Stefano Della Vigna and Elizabeth Linos (2022), “RCTs to Scale: Comprehensive Evidence from Two Nudge Units”, *Econometrica* 90, 81-116.

Jon Elster (1983), *Sour Grapes* (Cambridge: Cambridge University Press).

David Enoch (2016), “What’s Wrong with Paternalism: Autonomy, Belief, and Action”, *Proceedings of the Aristotelian Society* 116, 21-48.

(2017), “Hypothetical Consent and the Value(s) of Autonomy”, *Ethics* 128, 6-36.

(2020) “False Consciousness for Liberals, Part I: Consent, Autonomy, and Adaptive Preferences”, *The Philosophical Review* 129, 159-210.

(2022), “Autonomy as Non-Alienation, Autonomy as Sovereignty, and Politics”, *The Journal of Political Philosophy* 30, 143-165.

(Manuscript) “Epistemic Autonomy May Not Be a Thing”

David Enoch and Levi Spectre (forthcoming), “There Is No Such Thing as Doxastic Wrongdoing”, forthcoming in *Philosophical Perspectives*.

J. Dmitri Gallow (2022), “The Metaphysics of Causation”, *The Stanford Encyclopedia of Philosophy, https://plato.stanford.edu/entries/causation-metaphysics/#Norm*

John Greco (2003), “Knowledge as Credit for True Belief”, in *Intellectual Virtue: Perspectives from Ethics and Epistemology* (Depaul and Zagzebski eds.) (Oxford: Oxford University Press), 111-134.

Daniel M. Hausman and Brynn Welch (2010), “Debate: To Nudge or not to Nudge”, *The Journal of Political Philosophy* 18, 123-136.

Pelle Guldborg Hansen and Andreas Maaløe Jespersen (2013), “Nudge and the Manipulation of Choice: A Framework for the Responsible Use of the Nudge Approach to Behaviour Change in Public Policy”, *European Journal of Risk Regulation* 4, 3-28.

Ralph Hertwig and Till Grüne-Yanoff (2017), “Nudging and Boosting: Steering or Empowering Good Decisions”, *Perspectives on Psychological Science* 12, 973-986.

Maximilian Kiener (2021), “When Do Nudges Undermine Voluntary Control?” *Philosophical Studies* 178, 4201-4226.

Suzy Killmister (2018), *Taking the Measure of Autonomy: A Four-Dimensional Theory of Self-Governance* (New York and London: Routledge).

Jennifer Lackey (2007), “Why We Don’t Deserve Credit for Everything We Know”, *Synthese* 158, 345-361.

Catriona Mackenzie, and Natalie Stoljar, Natalie (eds.) (2000), *Relational Autonomy*: *Feminist Perspectives on Autonomy, Agency, and he Social Self* (Oxford and New York. Oxford University Press).

M. Maier, F. Bartos, T. D. Stanley, and E. J. Wagenmakers (2022), "No Evidence for Nudging after Adjusting for Publication Bias", *Psychological and Cognitive Sciences* 119 (31).

Diana T. Meyers (1987), "Personal Autonomy and the Pradox of Feminie Socialization." *The Journal of Philosophy* 86: 619-628.

Robert Noggle (2022), "The Ethics of Manipulation", *Stanford Encyclopedia of Philosophy,* available here: https://plato.stanford.edu/entries/ethics-manipulation/

Marina Oshana (1998), "Personal Autonomy and Society", *Journal of Social Philosophy* 29: 81-102.

(2006), *Personal Autonomy in Society* (New York: Routledge).

Duncan Pritchard, John Turri, and J. Adam Carter (2022), "The Value of Knowledge", *The sTanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/knowledge-value/

Jonathan Quong (2011) *Liberalism Without Perfectionism* (Oxford: Oxford University Press).

Joseph Raz (1986) *The Morality of Freedom* (Oxford. Oxford University Press).

Andreas T. Schmidt and Bart Engelen (2020), "The Ethics of Nudging: An Overview", *Philosophy Compass* 15(4).

Ben Schwan (2022), "Why Decision-Making Capacity Matters", *The Journal of Moral Philosophy* 19(5), 447-473.

Ernest Sosa (2009) *Reflective Knowledge* (Oxford: Oxford University Press).

Natalie Stoljar (2014), "Autonomy and Adaptive Preference Formation." In Veltman. Andrea and Piper, Mark eds. *Autonomy, Oppression, and Gender*, 227-252. (Oxford. Oxford University Press)

Richard H. Thaler and Cass R. Sunstein (2008), *Nudge: Improving Decisions about Health, Wealth and Happiness* (New Haven, CT: Yale University Press).

Cass Sunstein (2014a), *Why Nudge: The Politics of Libertarian Paternalism* (Yale University Press).

(2014b), "Response", *New York Review of Books*, 23 October 2014.

(2014c), Cass R. Sunstein, Choosing Not to Choose, 64 *Duke Law Journal* 1-52.

Doron Teichman and Eyal Zamir (2021), “Symposium on Limitations of the Behavioral Turn in International Law: Normative Aspects of Nudging in the International Sphere” *AJIL Unbound* 115, 263-267.

John Turri, Mark Alfano, and John Greco (2021), “Virtue Epistemology”, *The Stanford Encyclopedia of Philosophy,* https://plato.stanford.edu/entries/epistemology-virtue/

Mikhail Valdman (2010), “Outsourcing Self-Government”, *Ethics* 120, 761-790.

Jeremy Waldron (2014), “It’s All for Your Own Good”, *New York Review of Books*, 9 October 2014.

Eyal Zamir and Doron Teichman (2018), *Behavioral Law and Economics* (Oxford: Oxford University Press).

Why Isn't (Purely) Epistemic Autonomy of Value?

David Enoch*

1. A Disanalogy

When we make a decision, it is often important for us to make the right (or *a* right, or a sufficiently good) decision. And we have different ways of making it likely that we will. Sometimes, such ways involve a role for others. This role may be that of someone we turn to for advice, or someone we run our arguments by to see how strong they are, but it may also be something like outsourcing: We may, at times, just let someone else make decisions for us, or blindly follow the advice of another. This is how I may go about making decisions about, say, retirement savings plans. And while sometimes people think that such outsourcing is problematic from the point of view of the value of autonomy, I don't think that this is so – far from being in tension with my autonomy, such outsourcing, in the right circumstances, may be a manifestation thereof (and may be rationally required, to boost). After all, what is important to me in making a decision about retirement plans is just that I make the right one (in terms of risk, how much to save, how to invest, and so on), and, as I fully recognize, outsourcing maximizes my chances of making the right decision[1].

Not all decisions are like this, though. With some decisions, getting them right is not the only thing that matters. With some decisions, it also intrinsically matters *who* makes them. In particular, in some of *my* decisions, it matters greatly that *I* make them. An oft-given example here is the choice of a romantic partner. With such a decision, it seems important that the person whose romantic partner is at stake make the decision. Undoubtedly, this is at least in part for instrumental reasons – making such decisions for yourself often increases the chances of the choice being a good one (according to whatever parameters make the choice of a romantic partner good). But just as clearly,

* For comments on earlier versions, I thank Dani Attas, Ittay Nissan-Rozen and Levi Spectre. This paper was presented as one of the Burman Lectures at Umeå, and at the annual conference of the Israeli Philosophical Association. I thank the participants for the helpful discussions that followed.

[1] If you sense a hint of Raz's service conception of authority here, you're right. But I am not here committed to all of its details, and I'm not really discussing authority here. For some discussion, see my "Authority and Reason Giving" (2014).

the instrumental considerations don't exhaust things here. Even in those cases – and how sure are we that they are farfetched? – in which the chances of a good choice are higher when the choice of a romantic partner is outsourced, we still think that there's importance in making the choice for ourselves[2]. This importance need not always outweigh all other considerations – if I'm just terrible at choosing romantic partners, and if outsourcing the choice to my mother can guarantee for me eternal romantic bliss, perhaps it would be rational for me to outsource[3]. The important thing for our purposes here, though, is that this is not always and necessarily the case – there are cases in which it makes perfect sense to refuse to outsource, even when outsourcing will bring about a better (expected) decision or choice, cases in which the identity of the decision maker makes a non-instrumental difference[4].

It is natural to think about such cases in terms of the value of autonomy. It seems of (not merely instrumental) value that we be part-authors of our life stories, that we shape our lives ourselves[5]. And while such autonomy may be consistent with outsourcing choices about retirement savings plans, it does not seem consistent with outsourcing choices of romantic partners. And because an autonomous life is – other things being equal – a better life than a non-autonomous life, it makes sense for one to be willing to pay a price in other values in order to secure more autonomy for oneself. Presumably, this is what one does when one refuses to outsource at least sometimes even when outsourcing will yield greater returns in other values.

Of course, I don't want to pretend that anything is simple here – in particular, it would be good to have an explicit, non-metaphorical story distinguishing cases where outsourcing is consistent with one's autonomy and cases where it isn't. And as things will turn out later in this

[2] See Raz's (2006, 1014) independence condition.
[3] In the text I put things as if the matters are dichotomous – either outsource, or not. But these are really matters of degree – I may ask my mother for advice, I may ask her to make a tentative decision but also share her reasoning, I may fully outsource.
[4] In the text I focus on the case in which it matters that the decision maker is the person whom the decision is about. There may also be cases in which the identity of the decision maker matters when it's not going to be that person either way – when a decision is outsourced, it may sometimes (non-instrumentally) matter whom it is outsourced *to*. But I won't be discussing such cases here.
[5] Raz (1986, 369).

paper, I'll have to qualify even some of the things already said. But for now, these claims about the practical case will do. Let's turn to epistemology.

When we wonder whether something is the case, or deliberate[6] what to believe, it's important for us to believe the truth on the relevant matter, and to avoid falsehoods. We have ways of making it likely that we will – different epistemic methods, like relying on perception, using inferences, and so on. Some of these ways involve a role for others: We may ask for their opinion on the matter or on related matters, we may ask how things look from their perspective, we may run our inferences by them to see how strong they are. At other times, we may rely on others in a stronger, outsourcing kind of way – we may just take their word for it, and form a belief entirely on the basis of their testimony. This is how we form – and should form – almost all of our beliefs about scientific matters we're far from experts on, about historical events reported in books, etc. Some seem to think that when we outsource belief-formation in this way, this is in tension with something worth calling "epistemic autonomy", but this just seems false to me right off the bat – and the analogy to the practical case helps to see why. Just like relying on others is one of the ways in which I may exercise my autonomy in improving the quality of my decisions, relying on others may be one of the ways in which I improve on the quality of my belief-formation. (I return to this in the next section.)

So far, then, the analogy between the practical and the epistemic holds rather unproblematically[7]. But let's now ask whether there are epistemic analogues of the choosing-a-romantic-partner case. Are there cases, that is, in which it is not-merely-instrumentally important that I form my own belief without outsourcing, that I "think for myself", and furthermore, where I should be willing to pay a price in epistemic value just for that? Suppose, then, that I'm trying to

[6] Terminology is difficult here. It's not clear that we actually *deliberate* about what to believe. Typically, we wonder *whether p*, not *whether we should believe p* (I discuss Transparency below). And we don't really ever *decide* to believe, so even if we do deliberate about what to believe, this deliberation does not conclude (as it presumably does with practical deliberation) with a decision. I think my points in the text do not depend on any tendentious choice of terminology here.

[7] In the epistemic case too matters are not as dichotomous as the test here seems to indicate. See footnote 3 above.

make up my mind whether some given mathematical formula is a theorem, or whether the currently suggested judicial overhaul in Israel is anti-democratic. Suppose I can either try to figure out these things for myself, or rely on others who, as I myself concede, are significantly more reliable than I am on such matters. I have a colleague who is (as I know) much better than I am at mathematics, say, and another who is (as I also know) much more reliable than I am on constitutional and political matters. Now suppose it seems to me – having surveyed the (first-order) evidence, and based on my own devices alone – that the formula is indeed a theorem, and that the judicial overhaul is not anti-democratic. But suppose that my colleagues report otherwise (on both matters). Remember that I recognize their superiority over me in terms of reliability on these matters[8]. So I also recognize that my chances of getting to a true belief – as well as my chances of avoiding a false one – go up if I rely on my colleagues than if I stick with the beliefs, or perhaps credence levels[9], called for by only procedures of my own device. Is it ever epistemically rational for me to nevertheless stick to my own devices here?

In the practical case, we saw that sometimes it makes sense to refuse to outsource even when one fully realizes that outsourcing increases the chances of getting a good decision. In the epistemic case, though, this does not seem possible. If you believe that your chances of reaching the truth and avoiding falsehood on the question of theoremhood of the relevant formula are highest if you just take your colleague's word for it, and yet you insist on not outsourcing and instead "thinking for yourself", you are necessarily being epistemically irrational. Depending on the details, you may be downright incoherent (if you simultaneously believe *This formula is a theorem* and *Given my colleague's input, this formula is more likely not to be a theorem than to be one.* I return to this incoherence below). In the relevant respects, the situation is not different from one where you insist

[8] It is often important to ask whether the relevant superiority should be understood in terms of their actually being more reliable, my believing that they are, or my justifiably so believing. (See section 5 of my "Not Just a Truthometer" (2011).) Here I am trying to avoid these complications by stipulating that all of these conditions are satisfied in the example in the text.

[9] If we're talking in terms of credence levels, we may need to add something like accuracy to the list of epistemic values. I don't think this makes a relevant difference here.

on forming a belief about the temperature of a liquid by dipping your finger in it and going by how it feels, rather than by relying on the reading of a thermometer that you yourself recognize is much more accurate and reliable. If this is how you form your belief, it cannot be epistemically justified, and even if you fluke your way onto the truth, your belief will not amount to knowledge[10]. And this seems true perfectly universally, regarding any proposition or subject matter. (Though the political case does seem harder. More below.) In the practical domain we noted that the presence of cases in which it makes sense to refuse to outsource even at a price seem to reflect the value of autonomy. If it's true, though, that in the epistemic domain there can be no such cases, this seems like an important disanalogy. Epistemic autonomy does not seem to be a thing[11].

Now, this disanalogy calls for explanation. Why does it make sense to sometimes refuse to outsource in the practical domain, but not in the epistemic domain? What explains why practical autonomy is of value, and epistemic autonomy is not? The main task of this paper is to step up to this explanatory challenge.

Perhaps, though, you are not entirely convinced. Perhaps, for instance, you think that while refusing to outsource in the mathematical case is never epistemically justified, refusing to outsource in some other cases (like perhaps the constitutional-political one) may be epistemically justified. And I should immediately concede that things are not as clear-cut here as the previous paragraph makes them seem (and some qualifications will emerge as we proceed). Now, as long as you agree with me that *some* disanalogy is present here – perhaps just that refusing-to-outsource much more often makes sense in the practical domain than in the epistemic one, or that it's less clear that it's ever epistemically justified than it is that it's often practically justified – you should already see that the disanalogy calls for explanation. Another, secondary task of this paper is to make more plausible the

[10] Everything I say in the text is, as far as I can see, consistent with any plausible theory on peer disagreement. For my own, see "Not Just a Truthometer" (2010).

[11] Zagzebski (2012) says that the ideal of epistemic autonomy is incoherent. But I don't accept much of her reasoning. It's not even clear we use the term "epistemic autonomy" in sufficiently similar a way. And notice that the related ideal Zagzebski does accept (and develops further in her 2013) – that of intellectual autonomy – is not directly related to my discussion here.

claim that epistemic autonomy is not a thing – or at least, that *there is* this initial epistemic-practical disanalogy[12].

In the next section (2), I clarify the issue and get some preliminaries out of the way. In the following two sections I discuss three possible explanations – one (3) in terms of incoherence, one (4) in terms of the distinction between the right and the wrong kind of reasons (for belief), and one (5) in terms of the different role value pluralism plays in the practical and epistemic domains All three, I conclude, have some merit, but neither is fully satisfactory as an explanation of the disanalogy. In section 6 I revisit the disanalogy, arguing that there may be less of it than meets the eye. We can see that, I argue, once we are aware of the possibility that autonomy is, for the person whose autonomy it is, essentially a by-product. A somewhat disappointing conclusion follows.

2. Distinctions, Distinctions, Distinctions

In this section I clarify further the kind of case I'll be focusing on, the epistemic case that manifests, so I claim, the disanalogy with the practical domain that calls for explanation. I do this mostly by utilizing a host of distinctions. Such a section is needed not just for general clarificatory reasons, but also because the term "epistemic autonomy" is used in the literature with somewhat different meanings, and it's important not to conflate them.

2.1 Nothing about Extreme Autarkies

The first point has already been made in the introduction, so we can afford to be quick here: if the autonomy ideal – either in the practical or in the epistemic domain – is to be at all appealing, it can't be about some extreme requirement that we all be minor autarkies, fully satisfying our own practical

[12] There's more than one way of developing a relation between the value of autonomy and the epistemic. One that I won't discuss here is the claim that something like autonomy is a necessary condition for knowledge. I'm not sure what I think about this claim, but I suspect it has some interesting relations to some of the points discussed later in this paper (for instance, autonomy being essentially a by-product). For this claim about the relation between knowledge and autonomy, see Carter (2022). For a precursor, see Sosa (2003, 174).

or epistemic needs by ourselves. In both domains, we rely on others all the time, as well we should[13].

Why is it, then, that philosophers often seem to say the opposite? This may be because of failures to appreciate some of the distinctions below, as I proceed to explain. Or perhaps they accord too much weight to the plausible thought that there's often something less impressive if one reaches the truth by relying on others than if one reaches the truth on one's own. Perhaps outsourcing sometimes renders the true belief less of an *achievement*, less something the believer should take pride in. This may be so – though perhaps sometimes we should take pride in realizing our own limitations and acting on such realization (by outsourcing), and this in itself may be an achievement, for imperfect creatures such as ourselves[14]. But anyway, even if relying on others may make the relevant belief (or decision) less of an achievement, it will still often be the right thing to believe or do[15]. (Compare: Even if it's more impressive if someone can accurately tell the temperature of a liquid simply by dipping one's fingers in it, still usually using a thermometer is the epistemically justified belief-forming method. And certainly, given a reading of the thermometer, one is not often justified in overruling it because of how warm the liquid feels to one, especially not because of the possible relevant achievement here.)

Be that as it may, when some epistemologists write about epistemic autonomy as if it involves utter epistemic self-reliance, it is important to remember not only that the epistemic "ideal" they are describing is no ideal at all, but also that neither is its practical analogue.

2.2 Global vs. Local Autonomy

The literature on (practical) autonomy sometimes distinguishes between autonomy understood globally, as a feature of a life, and autonomy understood much more locally, as a feature of specific

[13] See, for instance, Fricker (2006), Zagzebski (2012), Nguyen (2018), and Matheson and Lougheed (2022, 2-3) and the references there.
[14] For an argument utilizing this observation to defend moral deference, see my "A Defense of Moral Deference" (2014).
[15] The discussion of essential by-products below may be relevant here.

decisions or choices[16]. (And there can be intermediate notions of autonomy, applying to a segment of a life, and so on.) We can draw an analogous distinction in the epistemic case.

For instance, those who complain about relying on others not being autonomous, or who praise thinking for oneself, may insist, highly implausibly, that on every occasion in which you form a belief, you have to think for yourself, or else you are in violation of your epistemic autonomy. Or, they can merely say – much more plausibly – that a life devoid of any thinking-for-oneself is, for this very reason, less rich and valuable as a life, that such a life lacks something by way of epistemic autonomy (all the while agreeing that in many specific cases it makes perfect sense to rely on others).

Now, as already emphasized, the sense of epistemic autonomy I am working with is somewhat different from the sense that requires that you think for yourself – my question is not so much whether you should think for yourself, but whether it is ever epistemically justified to believe according to such self-thinking even when you know that your chances of having a true belief go up by outsourcing. But I want to note here the following point: Even though there's not much to the extreme autarky conception of autonomy, I do think that the more global intuition here expressed – that a life devoid of thinking for oneself is, other things being equal, less of value – is worth vindicating[17]. We'll have to see later on whether what we end up saying about epistemic autonomy can deliver on this promise. Still, the kind of case I am focusing here is local: you are trying to form a specific belief – about whether a certain formula is valid, or about the suggested judicial overhaul – and the question is whether it is ever epistemically justified to insist on doing it yourself rather than

[16] See, for instance, Oshana (2006, chapter 1).

[17] Here's Matheson (2022, 1) seemingly rejecting epistemic autonomy in the local context but expressing the autonomy intuition in the global context (without explicitly distinguishing between the two): "After all, when trying to find an answer to a question, we should take the best available route to the answer, and the most reliable route to the answer to most questions is to rely on the minds of others. At the same time, there is something defective about an intellectual life that outsources nearly all of its intellectual projects, …"

relying on others, when you acknowledge that the latter option is more likely to get you to the truth. This question is about local, not global autonomy[18].

2.3 Instrumental Considerations

It should be uncontroversial that there may be instrumental payoffs to thinking for oneself. Perhaps, for instance, sometimes refusing to rely on others can help improve one's own cognitive abilities. Or perhaps it is sometimes important that people think for themselves so as to achieve some social good, like perhaps the payoffs of a marketplace of ideas. Or perhaps something along these lines is true not of people in general, but of some subset thereof, like experts – that is, perhaps the reliability of experts as a class, or the epistemic health of a field, depends on each expert forming their opinion to an extent independently, and anyway, not by all relying on the best expert[19].

I want to distinguish clearly between the question I'm interested in – whether it is ever epistemically justified to refuse to rely on others one acknowledges are more reliable than oneself in forming a belief – and instrumental questions of this kind. Two points are relevant here.

First, we can invoke the distinction between epistemic and pragmatic reasons for belief. I discuss it below, but for now all we need is some of the intuitive data underlying such more sophisticated discussions: If I wonder whether the suggested judicial overhaul is anti-democratic, and you show me how useful it will be for me to believe that it is not (there is such a shortage of people with respectable academic credentials supporting the government on this topic, so if I support the overhaul I can become quite central), you have not supplied me with evidence that the overhaul's democratic credentials are impeccable, and if I form that belief on the basis of what you said my belief will not be justified in the standard, straightforward sense of justification, and it will certainly not amount to knowledge (even if true). Similarly, then, if you show me all the instrumental

[18] Let me flag a possible complication here. If we end up insisting that epistemic autonomy is not of value locally but is of value globally, we may be close to self-torturer-paradox (Quinn 1990) territory, where the iteration of seemingly rational local decisions amounts to a clearly irrational global strategy. I return to this paradox briefly below.

[19] Dellsen's (2022) central claim is that thinking for ourselves increases the reliability of experts.

payoffs of refusing to outsource, you will have perhaps given me reasons not to outsource, but you haven't made it the case that I'll be epistemically justified in ignoring the (second-order) evidence, and that I should go with the belief that I myself acknowledge is less likely to be true[20].

Second, recall the disanalogy again. In the practical domain, the strong intuition was that there's something *intrinsically* important about making one's own choice (of a romantic partner, for instance). This is what explains why it makes sense sometimes to be willing to pay a price in the goodness of the decision just in order to make it on one's own. True, it's possible to insist that it does make sense to pay a price in the goodness of the decision in order to make it on one's own, but that what explains this are *other* instrumental payoffs of making one's own decisions. But this will not be fully loyal to the underlying intuition. So if the best that can be done to vindicate something like epistemic autonomy is to offer such instrumental considerations, this will leave the disanalogy between the practical and the epistemic case intact, and thus far unexplained.

2.4 Politics

Discussion of epistemic autonomy sometimes seems to come with a political twist[21]. Thinking for oneself seems politically important, perhaps as an antidote to anti-liberal tendencies to blindly follow the leader.

The first thing to say about this is that really, this is a particular instance of the previous point – such justifications of refusing to outsource are instrumental, and so the discussion in the

---

[20] In the text I'm drawing together practical instrumental reasons for *belief* with such reasons for *following an epistemic procedure* (such as outsourcing). The two are not always on a par – for one thing, one may argue that while there's some category mistake in offering practical reasons for belief (see, e.g., Berker (2018), there's no such mistake in offering a practical reason for following an epistemic procedure, which is, after all, an action.

(Perhaps relevant here – the current lit on norms of inquiry?)

But I think I can afford this in our context. If you rely on, say, the reading of a measurement device because you have some evidence it's reliable, beliefs formed in this way may very well be epistemically justified and may amount to knowledge. If, however, you rely on the reading of the measurement device for practical, instrumental reasons (I gave it to you as a gift, and you know how happy I'll be if you rely on it in your experiments), a belief formed in this way will not be epistemically justified.

[21] …

previous subsection applies[22]. But this particular instance nevertheless merits discussion for the following reason.

The cases of interest here are cases where you consider deferring to someone *whom you take to be more reliable*, or whether you can ever be epistemically justified in ignoring them, and proceeding to form a belief in a way that you know is less reliable. This question is very different from the question whether you should defer to those in positions of political power[23]. And this is important here, for it may be the beginning of a debunking explanation of intuitions in favor of epistemic autonomy. Perhaps, that is, what many people are really – and justifiably – concerned about is not epistemic outsourcing in general, but rather epistemic outsourcing *to those in positions of power*. If so, we should recognize this concern (instrumental though it is), and in our context just put it to one side. For our context is one in which the controversial outsourcing is to those characterized by their epistemic, not political credentials. Indeed, the tendency to conflate the two may explain why outsourcing in the case of the belief about the anti-democratic nature of the suggested judicial overhaul seems more problematic than outsourcing in the case of a belief about some mathematical formula being a theorem. It is relatively rare that people in positions of political power show interest in mathematical theorems and in who believes them. In such a context, political concerns about epistemic outsourcing are less serious. When it comes to the current political-constitutional events in Israel, though, *all* politicians have an interest. Political concerns – including about outsourcing – take center stage. And in many ways, they should. But we can still assume them away, if need be explicitly, by stipulating that our question is about outsourcing to the reliable, not the powerful, even when the topic is political.

[22] Perhaps with the following complication: You may think that while deferring in the relevant cases makes mistakes less likely, it makes *the worst* mistakes more likely. (This needs to be supported, of course). If what characterizes the worst mistakes is that they are, say, politically dangerous, then all of this is entirely practical and instrumental. But if the criterion for the worst mistakes is epistemic, then the case becomes more borderline – it remains instrumental, but it's at least in the epistemic ballpark, it's epsitemish. See the discussion of purism and impurism below.

[23] See Nguyen (2018, 111) for a related point.

In fact, at this point we already have considerable resources for a debunking explanation of intuitions about epistemic autonomy: Those who seem to be for it may conflate global with local autonomy, may allow instrumental considerations a role where they shouldn't play one, and may let their legitimate concerns about deferring to political leaders to infiltrate our discussion of deferring to those whose credentials are epistemic. Even if the combined effect of these debunking explanations is quite significant, though, so that there's less pressure to attribute value to epistemic autonomy, we still need an explanation of the disanalogy between the practical and the epistemic case.

2.5 Paternalism

The discussion of epistemic autonomy in the literature is often and understandably bound up with a discussion of epistemic paternalism[24]. Without committing to any specific, precise definition, cases of epistemic paternalism are cases in which the paternalizer somehow intervenes in a believer's belief-forming in order to benefit the believer epistemically. For instance, suppose you know I'm way too impressed with TV personalities who speak in a deep voice, tending to give much too much weight to their opinions in forming my own. So you make sure that I am otherwise occupied when the most charismatic, deep-voiced, government-supporting TV person is on. You're doing this because you know that the judicial overhaul is anti-democratic, that I am likely to be misled by that person, and you care about my epistemic status, perhaps about my epistemic wellbeing. This, it seems to me, is a paradigmatic case of epistemic paternalism, and it is very natural to think that it involves an offense against my epistemic autonomy, in a way closely analogous to that in which practical cases of paternalism offend against the relevant agent's practical autonomy.

I agree, of course, that epistemic autonomy and such cases of epistemic paternalism may be interestingly related. Still, it's important to see that the topics, even if related, are nevertheless distinct. The question about paternalism is whether, or when, anything about *my* epistemic

[24] See, for instance, Jackson (2022).

autonomy gives *you* a reason not to intervene (even in order to improve my epistemic situation). The question I am interested here is whether *my* autonomy gives *me* a reason to form a belief in one way (by myself) rather than another (outsourcing) even when I realize that the latter is more reliable. When I insist that epistemic autonomy may not be a thing, I insist on a negative answer to the latter. This is consistent with a positive answer to the former. (One way of understanding the discussion below of autonomy as essentially a by-product is as supporting this combination of claims.) If so, the disanalogy between the practical and the epistemic domains stands when it comes to first-person judgments (whether my own autonomy gives me a reason for action or belief), but the third-person questions (whether you should intervene) gets an analogous answer in the two domains.

Let me note here another point about epistemic paternalism. I do think that the intervention in the example above offends against the believer's autonomy, and furthermore, that this is a reason not so to intervene (though it may, at times, be outweighed by other reasons, of course). What is much less clear, though, is that the relevant offense involved is against the believer's *epistemic* autonomy, at least not in my intended sense. So we need to be clearer about the intended, perhaps purer, sense of "epistemic" that I'm working with.

2.6 More Generally: The Purely Epistemic

Miranda Fricker (2007) defines epistemic injustice as injustice directed at someone *in their capacity as a knower*. Such injustice is, of course, a *moral* violation, as indeed all injustice is. But because it essentially involves the victim's capacity as a knower, and perhaps for some other reasons as well, it is natural to call it epistemic injustice. (And indeed, it is one of Fricker's contributions to show that the moral and the epistemic are more closely related than we may have thought[25]). Still, there is *a* narrower, perhaps purer sense of "epistemic", in which cases of epistemic injustice are not exactly

[25] Relevant here are also discussions of moral encroachment and of doxastic wrongdoing. See, for instance, Enoch and Spectre "There Is No Such Thing as Doxastic Wrongdoing" (forthcoming).

epistemic – they are not directly about what beliefs or credences are justified given a specific body of evidence. Of course, there's no point to fighting over terminology[26], so I'm not trying to convince you that cases of (for instance) epistemic injustice are not epistemic at all. I'm just trying to convince that we can make sense of the narrower category of the purely epistemic.

When I claim that epistemic autonomy is not of value I use "epistemic" in this narrower sense. I don't think, in other words, that anything about one's autonomy directly affects which beliefs one is justified in holding, what credences what is justified in having, etc. I remain open to the possibility that the value of autonomy makes a difference in all sorts of ways, perhaps even "in our capacity as knowers", or in a way relevant to the norms of inquiry[27]. Return, then, to the paternalism example. The intervention – hiding some evidence that is likely to mislead a believer, precisely in order to make it more likely that their beliefs are true and justified, and perhaps amount to knowledge – clearly offends against the believer's autonomy, something that is here of at least some value. And it does so in a way that is directly relevant to the believer's capacity as a knower. Perhaps this is enough, then, to conclude that what's being offended is the believer's epistemic autonomy, and that it is of value. Even if this is so, though, it doesn't show that epistemic autonomy in the narrower sense is of value[28]. It doesn't show that considerations about autonomy should play a role in forming beliefs, or that they are directly relevant to whether a belief is justified. When I claim that epistemic autonomy is not of value, it is this narrower sense of epistemic autonomy that I have in mind.

This applies to some of the previous subsections here. Perhaps autonomy in one's capacity as a knower is politically important, or perhaps more generally it's important for practical,

[26] I used to insist that only the narrower sense is really about the epistemic, and that the broader sense is best referred to as practical things in the vicinity of the epistemic, or that are just epistemically relevant (2016, 31). I have now clearly lost that minor and unimportant terminological battle.

[27] …

[28] In the text I don't distinguish between the relevant autonomy being epistemic, and autonomy having an epistemic value. Perhaps for some purposes such a distinction is important: We can perhaps ask about the general kind of autonomy whether it's (also) of epistemic value, or about epistemic autonomy whether it's of moral value. But I think for my purposes here – really, asking about epistemic autonomy whether it's of purely epistemic value – I don't need to worry further about this distinction.

instrumental reasons. If you want to call that "epistemic autonomy" and assign it value, I will happily grant you this way of speaking, but I will insist that this way in which autonomy is of value is not purely epistemic, it is not directly relevant to what beliefs one is justified in having.

Armed with this distinction – between the narrower, purely epistemic significance of autonomy and its wider possible significance – we can now more clearly see what's wrong with attempts to reject the explanandum of this paper, namely, the claim that autonomy has a kind of practical value that it lacks in the (purely) epistemic case. On the epistemic side: it is now clear that one can consistently accept the wider epistemic significance of autonomy while denying it the more purely epistemic value (insisting that one can't be epistemically justified in believing a proposition while rejecting, for reasons of autonomy, an epistemic procedure that one acknowledges is more reliable, and its opposite deliverance). And this observation should undermine any temptation you may have had to secure autonomy a more purely epistemic role. And on the practical side: The disanalogy survives. For in the practical domain, the value of autonomy seems to be relevant in (the analogue) of this narrower sense as well[29] – this, after all, was the point of the example of choosing a romantic partner.

Before concluding this section, let me address a worry. In this section, and to an extent elsewhere in this paper, I rely on a distinction between purely epistemic and other considerations. In the current epistemology literature, though, it seems that impurism is gaining grounds. Thoughts of pragmatic encroachment, of doxastic wrongdoing, and perhaps other thoughts as well, have been used to challenge the purity of the epistemic. Now, I have deliberately avoided assuming anything like epistemic impurism: I remain here neutral on the possibility of other considerations encroaching on the epistemic. I merely insist that even if they do, we can still isolate, at least sometimes and in the cases relevant here, the more purely epistemic from its environment. But perhaps impurists

[29] Things are a little complicated here by the fact that whereas in the epistemic case we can perhaps distinguish between the (purely) epistemic and the practical significance of autonomy, in the practical case the distinction between the narrow and the wider sense occurs *within* the practical. But I don't think this complication matters here.

should reject even this more minimal assumption? If so, and if impurism is true (or at least plausible), that could spell trouble here. Fortunately, then, I don't think that impurists should have any trouble with my minimal assumptions here. All that's needed to render such assumptions plausible are the intuitive examples I've been using: the personal and political payoffs of believing that the judicial overhaul is democratically kosher leave a narrower, purer sense of "evidence" or "reasons for belief" or "epistemically justified" unaffected; Even if we shouldn't hide evidence from people in order to promote their epistemic wellbeing, and even if this is so because of their autonomy, this falls short of showing that they should ignore, for reasons of their own epistemic autonomy, more reliable sources; and so on. If impurism requires rejecting such intuitive claims, so much the worse for impurism. More plausible versions of impurism will not reject them, and should therefore be entirely on board with my minimal assumptions about the purely epistemic.

2.7 Sovereignty and Non-Alienation

The value of autonomy is the value, perhaps roughly, of living one's life according to one's own values and deep commitments, of shaping one's life with one's decisions. But in many contexts it is important to distinguish between two different values here. One – which I call non-alienation – is the value of living one's life according to one's values and deep commitments. Another – the one I call sovereignty – is the value of having the last word on relevant matters, of being the one whose decision is, so to speak, law on those matters. Very often, these two values coincide, for in many cases if you get the last word on some issue, you're going to decide according to your deep values and commitments, and in many cases, allowing you to have the last word is an excellent way of making sure it's your commitments and values that shape things. But this is not always the case, and when the two values come apart – when, for instance, the best way of securing non-alienation

requires *not* letting you have the last word (perhaps because you are weak-willed) – the distinction becomes important[30].

When we're discussing epistemic autonomy, are we talking about sovereignty or non-alienation (or perhaps about both)? I think that both may be relevant, but in different cases in different ways.

In many cases, sovereignty is not really at stake. Even if you decide to outsource, and form a belief purely on someone else's word, it's still you who are making the final call here, as it were. Your sovereignty is not more threatened by outsourcing to another person than it is by "outsourcing" to a thermometer. Still, sovereignty may be relevant: First, perhaps this is a natural way of thinking about the paternalism case. Hiding some evidence from you may be the epistemic analogue of feeding you misinformation about the alternatives you have to choose from or restricting your options, cases which may be thought of as a violation of your practical autonomy in the sense of sovereignty. But these aren't the cases I mostly focus on. Those, recall, are cases in which you yourself wonder what you should believe, and whether anything about your autonomy is ever a relevant consideration. Second, recall again the choosing-one's-romantic-partner case. There, insisting that it's important that it be *my* decision does seem to be at least partly about sovereignty, even though the point above applies – even if I let my mother choose my romantic partner for me, it will still be me who lets her do that. This seems to indicate that at least some sovereignty concern remains in such cases. And this may apply to the epistemic cases I started with – where the question is whether it's important that I make up my own mind (rather than outsource) in the kind of way that can make a difference to which beliefs I'm justified in having.

Non-alienation too may be epistemically relevant. Suppose that Bas is deeply committed to inference to the best explanation not being an epistemically good rule of inference[31]. Now suppose

[30] I discuss the distinction in "Hypothetical Consent and the Value(s) of Autonomy" (2017). I discuss the relations between sovereignty and non-alienation in "Autonomy as Non-Alienation, Autonomy as Sovereignty, and Politics" (2022). The distinction – or one very close to it, sometimes put in other terms – is also used by Brudney and Lantos (2011).
[31] Van Fraassen (1989).

that Bas wonders whether the suggested judicial overhaul in Israel is anti-democratic. He can try to reason to a conclusion all by himself, of course. But he doesn't think he's very reliable on such questions. In fact, he recognizes that Gil is more reliable on such matters. But he also knows that Gil routinely uses inference to the best explanation, and is likely to do so here as well[32]. Can Bas be justified in ignoring Gil's testimony (say, that the overhaul is anti-democratic) and believe that the overhaul is not democratically problematic, for the reason that Gil's reasoning is not in line with Bas's deepest epistemic commitments? Can Bas justifiably think something like "Yeah, relying on him will make it more likely that I believe the truth on this matter, but it's also important that I form beliefs according to my own epistemic commitments, so I'm going to ignore Gil's testimony?". To insist on epistemic autonomy in the sense of non-alienation being of value is to answer in the positive. I, of course, don't[33]. In many practical cases, though, analogous concerns do make sense – think of "this wouldn't be my style" as a reason for action. (But see the discussion of pluralism below). So the disanalogy between the epistemic and the practical domains is still in need of explanation.

2.8 So:

We are in a position, then, to conclude this long, perhaps somewhat tedious section, and state precisely what it is that I deny when I deny that epistemic autonomy is of value.

The question whether epistemic autonomy is of value – in the sense I intend it to have here – is the question whether a believer is ever justified in believing or forming a belief for an autonomy-related reason, if need be at the expense of likelihood of truth. The questions is a local one – about a specific belief – not about a general global belief-forming strategy; it's entirely about epistemic

[32] Harman (1965).

[33] Anyway, not for first-person cases. Perhaps third-person cases are different – perhaps, for instance, an intervention may be justified partly because it helps someone form beliefs more in line with their deep epistemic commitments. But first, even in the third person, I'm not sure this is ever the case – at least not at the expense of truth (or some such epistemic aim). And second, the discussion below of epistemic autonomy as essentially a by-product is relevant to this use pf the distinction between first- and third-person perspectives here.

justification – not about anything instrumental; political considerations having to do with deferring to those in power are strictly speaking irrelevant one way or another here; more generally, the question concerns the purely epistemic; it is not in the first instance about a possible decision to think for oneself, but it's related to it, of course; and it may be either about sovereignty (about forming my belief all by myself) or about non-alienation (about reasoning styles, perhaps, or anyway about one's own deep epistemic commitments).

Much of the (not too big[34]) literature on epistemic autonomy focuses on other phenomena. But there are in the literature also discussions that come very close. One example is Dellsen's (2022) "puzzle of epistemic autonomy" (though it is primarily not about which beliefs are justified in the presence of autonomy considerations, but about the status of a general strategy of thinking for oneself). Another is Huemer's (2005, 526) emphasis on the agent-centeredness that has to characterize epistemic autonomy. For if we do accord weight to epistemic autonomy, it follows that in the same circumstances, privy to the same evidence, different people may be justified in having different credences and beliefs: If Bas and Gil are both justified in giving greater weight to their respective "styles" of reasoning when it comes to inference to the best explanation, they may end up with different justified beliefs on many occasions[35]. And even if it's ok for you, say, to resist the testimony of those you concede are much more reliable than you are in order to think things out for yourself, and even if the belief or credence you end up having is justified, still this is clearly not a reason *for me* to endorse that credence or belief. Putting things in this way makes it even clearer how implausible it is to attribute value to epistemic autonomy[36].

---

[34] See Matheson and Lougheed (2022, 1).

[35] I briefly discuss epistemic permissiveness in the next section. But the point here is stronger: Assuming epistemic autonomy is of value, we can presumably describe cases in which it's not merely permissible for two believers to differ in their beliefs or credences, but rather it's *required* of each to have a different belief or credence (as is probably the case in the practical domain, where one may be required to give some weight to one's own autonomy).
And there's another complication here. It may be thought that even if IBE is a good rule of inference, Bas's belief that it isn't defeats his initial justification (which he as just like the rest of us) for using IBE. If this is so, then this will ground another way – different from the one in the text – in which people with the same evidence may be justified in different beliefs or credences.

[36] It is, of course, unobvious how agent-relativity is to be understood. For my own attempt, see my "Backgrounding Agent-Relativity" (manuscript).

So we still need an explanation of the practical-epistemic disanalogy.

3. Possible Explanation: Wrong Kind of Reasons

It is common to distinguish – across a wide range of attitudes, of which belief is central – between reasons of the right kind and reasons of the wrong kind for having them. The ontological argument for the existence of God gives, if successful, a reason of the right kind for belief in God's existence: If successful, it gives evidence supporting the existence of God, it makes such a belief epistemically justified, it could render it knowledge, it's a reason *for* which one may directly believe that God exists, it has the right kind of connection to truth. Pascal's wager, though, even if successful, doesn't give a reason of the right kind for belief in God. Perhaps it manages to give *a* reason for such a belief (though even this is controversial[37]), but even if it does, it doesn't give evidence for the existence of God, it may make such a belief justified but not epistemically justified, it could not render that belief knowledge, arguably it's not a reason *for* which one can directly believe that God exists, it arguably does not have the right kind of relation to truth. A lot here is controversial, including how to best give a theoretical account of the distinction between reasons of the right and the wrong kind, when it comes to beliefs, and more generally[38]. But it is not controversial, I think, nor should it be, that some distinction along these lines is important in many contexts.

You wonder whether that formula is a theorem. Reasons are offered to think that it is: a (purported) proof – but you are, of course, fallible about recognizing validity; the testimony of an expert; perhaps the fact that it's structurally very similar to formulae already known to be theorems; how good it will make you feel to believe it's a theorem; that you are more likely to get a good grade in that test if it is (because that's what you wrote there). It seems clear – intuitively, pre-theoretically – that the first three (proof, testimony, some inductive claim about structure) are, if successful,

[37] Again see Berker (2018).
[38] See, for instance, Rabinowicz and Rønnow-Rasmussen (2004), Hieroymi (2005), Schroeder (2012).

reasons of the right kind to believe that the formula is a theorem. The others (having to do with some pragmatic payoff of believing it's a formula, and with wishful thinking) are, even if successful, reasons of the wrong kind for belief. Notice that they remain reasons of the wrong kind for belief even if offered by the believer herself. ("Why do you believe it's a theorem? Well, believing that it is just makes me so happy." We can sometimes make sense of such an answer, but it's very hard to think of it as a serious epistemic justification, or even an attempt at one). Suppose, then, we ask you why you believe that the formula is a theorem, especially given that those you acknowledge are much more reliable than you say otherwise, and you respond with "Sure, relying on them makes it more likely that I'll reach the truth here, and I understand that they say it's not a theorem. But it's important for me to think for myself. And to me, this proof (the one the experts are saying is fallacious) seems valid." Or perhaps you answer with "Yes, I understand they're more reliable, but their style of reasoning is not mine, and it's important that I stick to my style.", or something of the sort. Do these reasons sound more like the paradigmatic reasons of the right kind for belief (proof, testimony, etc.) or more like those of the wrong kind (wishful thinking, etc.)?

In terms of the (soft, non-committal) criteria above: Do these autonomy-based reasons give evidence for the claim that the formula is a theorem? Could they render that belief epistemically justified? Could they render it knowledge? Can you believe that it's a theorem directly *for* such reasons? Do they have the right kind of connection to the truth of that belief? Myself, I answer all of these questions in the negative. (I return to the fact that not everyone will in a minute.) If so, autonomy-reasons are reasons of the wrong kind for beliefs.

This may already justify some suspicion towards epistemic autonomy. But it may do more than that – it may explain the disanalogy between the practical and the epistemic significance of autonomy. For no similar wrong-kind-of-reasons intuitions arise when autonomy is invoked in the practical domain. If I say something like "I understand that my mother's choice of a romantic partner for me may be better than my own, but still, it's important that the choice be mine.", and I offer this as a reason for choosing one partner rather than the other, this in no way feels like the wrong kind

of reason for that choice[39]. If so, we have this clear difference between the practical case, where the value of autonomy grounds reasons of the right kind, and the epistemic case, where the reasons (if any) grounded in the value of autonomy are of the wrong kind. And this, it may seem, is a good explanation of why it is that the value of autonomy plays out differently in the practical and the epistemic cases.

But this, as an explanation, won't do. One reason for this is that intuitions about right and wrong kind of reasons sometimes vary, so that the starting point of this suggested explanation may be challenged – someone who is more into epistemic autonomy than I am is likely to insist that the autonomy-based reasons above may after all be epistemic reasons of the right kind. And given the controversy about how the distinction is best captured and what tests better indicate reasons of the wrong (or right) kind, it's going to be hard to make progress on this. Relatedly, and more problematically still: This explanation too immediately raises another explanatory challenge – for why is it that autonomy-reasons are reasons of the right kind in the practical domain and reasons of the wrong kind in the epistemic domain? And here, it seems, this question is so very close to the question we started with, that it doesn't seem like any explanatory progress has been made.

A possible way to defend this explanation at least to an extent would be to offer some substantive answer to the question it raises – why autonomy generates right-kind-reasons in the practical case but wrong-kind-reasons in the epistemic case – that is sufficiently "far" from the current concerns, so as to generate independent plausibility. Perhaps one plausible line of thought here utilizes Transparency, the claim, roughly, that deliberation whether to believe p typically reduces to deliberation whether p[40]. Because for most propositions p, autonomy considerations are

[39] This may be because there are no wrong-kind reasons for actions, but I'm not sure about this. For attempts to draw similar distinctions within the pratical, see Schroeder … Raz …
Assuming something like the distinction applies within the practical, the tests for what qualifies as a right kind of reasons will for the most part have to be different in the epistemic and practical cases – obviously, nothing about the relation to evidence, truth, or knowledge is directly relevant in the practical case. But this doesn't show that no analogous distinction between right and wrong kind of reasons applies. And some of the tests from the epistemic case may apply in the practical one as well – for instance, we can still ask whether a reason is one the agent can directly respond to with the relevant action.

[40] See, for instance, Shah (2003). I thank Levi Spectre for relevant discussion.

irrelevant to whether p, granting them weight in the deliberation whether to believe p will amount to a violation of Transparency. Perhaps this is what makes autonomy reasons wrong-kind-reasons when it comes to beliefs. But nothing analogous to Transparency applies in the practical domain, so autonomy-reasons may very well be right-kind-reasons in that domain. If this explanation works – and if the fact that Transparency applies in the epistemic but not the practical domain is not deeply mysterious, or is perhaps a good explanation stopping point – then perhaps the explanation in terms of the distinction between the right and wrong kind of reasons can carry some of the explanatory weight here after all.

4. Possible Explanation: Incoherence?[41]

You're wondering, then, whether the suggested judicial reforms are anti-democratic. You review the (first-order) evidence, and come to tentatively conclude that they are not. You also realize that this colleague of yours is much more reliable on such matters, and that she is strongly convinced that these reforms are anti-democratic. You understand that relying on her will get you a much better chance of reaching the truth. She is likely right, as on such matters she often is (much more often than you are, as you realize). But you choose not to outsource, for autonomy-related reasons. What, overall, do you believe about the judicial reform? On the one hand, you believe that it is not anti-democratic. On the other hand, you have such beliefs as that it is more likely that it's anti-democratic than that it isn't; that it's probably anti-democratic; that your colleague's belief that it's anti-democratic is likely true; that your belief that it isn't is likely false. Let's focus, then, just on these two conjunctions: *The reforms are not anti-democratic, but I'm probably wrong about that;* or *The reforms are not anti-democratic, but it's more likely that they are anti-democratic than that they are not.* These are not contradictions, exactly, but they are *very* close: You both believe a proposition and that it is less likely to be true than its negation. This seems incoherent (as I quickly noted in the

[41] For discussions relevant to this section, I thank Ittay Nissan-Rozen and Levi Spectre.

introduction). The autonomy consideration – even if (as we're assuming for reductio) gives you a reason to refuse to outsource, does not render your overall beliefs here coherent. And it's very hard to see how your beliefs can be described here without such incoherence[42].

In the practical domain, though, no similar incoherence is involved. You have tentatively decided on a romantic partner (Bachelor #1) you want to pursue. You also know that your mother voted otherwise – for Bachelor #2 – and that she is much more likely to be right on the matter than you are. So you believe that Bachelor #2 is likely the better romantic partner for you, but you intend to choose Bachelor #1. This belief-intention combination is perhaps not the model of coherence (more on this shortly), but it is surely not as badly incoherent as the combination of beliefs in the previous paragraph.

Perhaps this, then, is the explanation we're after for the disanalogy between the role autonomy plays in the practical and the epistemic domains. In the epistemic domain, according weight to autonomy will result in incoherence. Not so in the practical domain. This is why autonomy cannot play in the epistemic domain a role analogous to that it plays in the practical domain. If this is so, what initially seemed like a substantive, evaluative disanalogy is reduced to a formal one – which would in itself be an interesting result.

It's not entirely clear that there's no incoherence in the combination of the belief that Bachelor #2 is the better choice and the intention to choose Bachelor #1. True, it's not *the same* incoherence there is in the combination of beliefs above, but perhaps this is just a function of the fact that the epistemic case is, well, epistemic, so that it's all about beliefs, whereas the practical case is about actions (and beliefs). We shouldn't expect, then, the incoherence to be precisely similar. And it does seem that if there's no further explanation, combining the intention to choose Bachelor #1 with the belief that Bachelor #2 is the better choice (and perhaps even the belief that

[42] I'm sure that this is closely related to some of the topics discussed above, like for instance Transparency, and the distinction between reasons of the right and the wrong kind.

one should choose Bachelor #2, or that Bachelor #2 is likely to be the better partner) at the very least gives rise to a tension.

The natural thing to say, though, is that there is no incoherence after all in the practical case, but for another reason. When you decide to go for Bachelor #1, the considerations you take into account – your utility function, if you will – include not just all the merits of Bachelor #1 compared to those of Bachelor #2. Rather, you also assign value to your independence, or autonomy. The alternatives you have to choose between are not *Bachler #1* vs *Bachelor #2,* but rather, roughly, *Bachelor #1 + autonomy* vs *Bachelor #2 (without autonomy)*. And as between *these* options, you prefer the former[43]. Your mother, however, merely prefers (and is likely right in so preferring) *Bachelor #1* over *Bachelor #2*. In this way, no incoherence remains: It's possible that Bachelor #2 is better than Bachelor #1, but that Bachelor #1 + autonomy is better than Bachelor #2 without it.

But this means that, first, the disanalogy is not after all formal – it's really about whether autonomy is of value in the relevant domain, whether it's included in the agent's utility function, or some such. And second, of course: This means we're really back at square one. Assigning autonomy value in one's practical deliberation does not result in incoherence, whereas in the epistemic domain it does. This is, pretty much, what we set out to explain. It's hard to see that progress has been made.

5. Possible Explanation: Pluralism

Talk about the optimal choice of a romantic partner (for one) is always suspicious, regardless of who it is who is actually making the choice. And the reason is very clear – there is a huge number of factors that go into making a romantic partner a good romantic partner. And there are different ways in which a romantic partner can be good. It is highly implausible to think that there's a single,

[43] On this picture, the relevance of the value of autonomy is optional, it depends on whether the relevant agent cares about autonomy. Otherwise, it has not value in the relevant case. I'm not sure whether this is a problem.

unique way of factoring in all of them so as to reach a precise ordering of possible romantic partners, from best to worst. It is much more plausible to think that when it comes to a choice of romantic partners, there is a pluralism of relevant values, and a wide range of incommensurability: Even just for you, there are many good romantic partners (of course, there are also many bad ones), who may be very different from each other, good in different ways to different extents, and yet such that none is better than the other (nor need they be equally good). Even if two potential partners are not incommensurably good – even if one of them is better than the other – it's still probably true that the less good one has many good-making features. And it's natural to think that this combination of pluralism and incommensurability opens up room for the value of autonomy to play its role: For when it comes to a choice between incommensurably good potential partners, autonomy can reign unchallenged. And even in cases of one potential partner being better than another, the good-making features of the less good candidate still make choosing him intelligible, perhaps a permissible option. It makes no sense, the thought seems to be, to give weight to the value of autonomy even when this will lead to a choice of a clearly bad option over a clearly good one. It's just that when it comes to choosing a romantic partner, this is hardly ever the case. On this picture, then, and perhaps roughly, autonomy is of value only *in pursuit of the good* (even if not necessarily the optimal). As a result, only in the presence of value pluralism and incommensurability (very common in the practical domain) does it make sense to assign one's autonomy value in one's deliberations[44].

If it can plausibly be argued, then, that there are no analogues of pluralism and incommensurability in the epistemic domain, this will be the explanation we need for the disanalogy between the role of autonomy in the practical and the epistemic domains. And pluralism does seem far less plausible in the epistemic domain. Just think of the multiplicity and variety of good-making features of, say, romantic partners. There is nothing resembling this kind of richness when it comes

[44] This picture is, of course, very Razian. I don't want to commit to any precise details here about how best to understand Raz, and I intend to address in detail Raz's view in future work. For now, let me note that the claim that autonomy is only of value in pursuit of the good, and the close connection between autonomy, pluralism, and incommensurability, all come from Raz (1986, chapters 14 and 15).

to beliefs, at least when those are evaluated in the purely epistemic way highlighted above. Truth is one such value, of course. So is avoidance of falsehood (more on this below). And what else? Perhaps accuracy is of such purely epistemic value, and perhaps it's not entirely reducible to the value of truth. Perhaps understanding is[45]. But this is pretty much it. Furthermore, in the purely epistemic domain, even these values play only a non-instrumental role – even if it can be shown, for instance, that if you believe that the relevant formula is a theorem this will maximize your true beliefs or understanding over time, this is not any reason (of the right, purely epistemic kind) so to believe[46]. Even if there is some room for pluralism (and with it, for incommensurability) in the epistemic domain, there's much less of it compared to the practical domain. So is this all the explanation we need for the disanalogy between these domains when it comes to the value of autonomy?

This explanation predicts that in epistemic cases in which pluralism *is* in play, autonomy may after all have a role to play. So think of cases in which it matters greatly how much weight the believer assigns to reaching truths and how much weight they assign to avoiding falsehoods. These comparative weights will affect how risky they are willing to be in forming beliefs: Someone who is most concerned about avoiding falsehoods will approximate the Cartesian policy of getting rid of all uncertain beliefs, whereas someone who places more weight on having true beliefs will be willing to take greater risks in their belief-forming. It's not implausible to think that there's more than one epistemically permissible way to go here[47]. Perhaps some different ways of assigning relative weight to these two epistemic values – reaching truths and avoiding falsehoods – are incommensurably good. And so perhaps here it makes sense for a believer to assign some weight in her epistemic

[45] In the context of a discussion of moral deference, for instance, many insist on the epistemic significance of understanding. See Nickel (2001), Hopkins (2007), and Hills (2009).
In terms I get to later in the text, I think that the most natural way to think of understanding, even if we acknowledge that it is of epistemic value, is as a state that is essentially a by-product.

[46] I have to say this seems intuitively obvious to me, but I've learned that not to all. For some relevant discussion, see Berker (2013).

[47] For some relevant discussion, see Weintraub (2013).

deliberation to what she can think of as her own style, or her own values and commitments, or to the ultimate belief being determined by herself, without outsourcing.

I'm not sure what exactly to say about such cases. I agree that attributing weight to epistemic autonomy is much more plausible in such cases than in others, where pluralism and incommensurability seem less plausible. Perhaps this already can serve as some confirmation of the explanation of the practical-epistemic disanalogy offered in this section[48]. Whether we should infer according to IBE, it seems plausible to assert, is not a matter for personal style. But how risky we should be in forming beliefs? Perhaps, within some bounds, that may be a matter of style, and if so, this allows for permissible variation across individuals.

In the practical domain, we insisted, it makes sense to sometimes be willing to pay a significant (if not unlimited) price in the quality of a decision just in order to make it oneself, or to have it reflect one's deep commitments. If this is so, it means that autonomy comes into play in the practical domain even in cases where the relevant options are not incommensurable, indeed, when one of them is clearly better than the other. But perhaps this too can be fed into the suggested explanation of the practical-epistemic disanalogy: In the practical domain, it seems plausible to say that even the clearly inferior choice of a romantic partner has *something* to be said for them, they too will almost always have good-making features. Perhaps this is why autonomously choosing them is of value, even though it's far from the optimal choice. In the epistemic domain, though, because there's a relative paucity of relevant values, often inferior options will have nothing or very little to be said for them. Perhaps this is why autonomy in that context loses its value[49].

[48] I'm not entirely sure, though. Recall that the test for the value of epistemic autonomy was that it makes sense to be willing to pay a price in reliability (or some such) just in order to manifest sovereignty or non-alienation in one's belief formation. In the case in the text, though, this is not precisely what's going on – because the underlying commitments or values or choices are precisely *about* the relative values of truth and avoiding falsehoods (reliability, presumably, comes along with these). That is, the purported reasons of personal "style" are not weighed against reasons of truth and reliability, but are best seen as offering competing interpretations thereof.

[49] Relevant here is also some unclarity in how to understand epistemic permissiveness exactly: Permissiveness is sometimes understood as the claim that more than one epistemic response (a belief, a credence, a suspension of judgment) is maximally rational (White 2005). At other times, though, it is more naturally understood as the claim that even less than maximally rational responses are nevertheless permissible (this, it seems to me, is the reading of epistemic permissiveness that is more closely analogous to moral

I am sure this story goes at least some of the way towards explaining the disanalogy between the practical and epistemic significance of autonomy. But I am not sure it's all we need. One reason is that I have some doubts about the relations between autonomy, pluralism and incommensurability this story utilizes. I don't deny such relations – I am just not convinced that the relations are *that* close. In particular, I think that autonomy is sometimes of value even in pursuit of the bad. But this is a matter for another occasion[50].

My second reason for thinking that this may not be the entire story here is more speculative. On the suggested story, the fact that value pluralism characterizes the practical domain more clearly than it does the epistemic explains why autonomy is more clearly of value in the former than the latter. But this raises the question – why *is it* the case that value pluralism is such a clear feature of the practical, but not of the epistemic? What has been achieved is that one practical-epistemic disanalogy has been explained by reference to another, but now this other disanalogy stands in need of explanation. It is tempting to think that no progress has been made.

This would be too quick, though. First, local explanatory progress is still explanatory progress. That more progress remains to be made does not show that no progress has already been made. Second, explanatory chains have to come to an end somewhere, and not all points are equally good explanation-stoppers. It may be argued that the value-paucity of the purely epistemic – having to do with truth and falsehood, primarily, and perhaps some other closely related values – is as good a place as any to bring one's explanatory chain to a halt. In other words, it may be argued that while it is not plausible to accept the disanalolgy regarding the value of autonomy as brute, accepting the disanalogy regarding value pluralism as brute comes relatively painless.

These are legitimate dialectical moves, it seems to me. And so, the explanation in terms of value pluralism and its different role in the practical and epistemic domains clearly has some merit. But these dialectical moves do not render the objection entirely weightless. So this objection –

---

permissibility). The point in the text here is that the gap between these two readings is rather small, given the paucity of epistemic values.

[50] "Revisiting Raz on Autonomy" (work in progress).

certainly together with the doubts about the relation between autonomy and pluralism – show at least that more may be hoped for.

6. Less of A difference than Meets the Eye?

In this section I don't offer an attempt at an explanation of the practical-epistemic disanalogy. Rather, I argue that there may be less of it than meets the eye.

So far, I've been assuming a rather strong relation between the *value* of autonomy and it being *reason-giving*. I insisted that it doesn't make epistemic sense for a believer to resist the truth or a more reliable epistemic method just for autonomy-related reasons, and from this claim – about epistemic autonomy not being reason-giving – I rushed to conclude that epistemic autonomy is not of value. And in many contexts the relation between values and reasons is indeed rather close – we often have reason to pursue what is of value. But not always.

An important – and quite large – family of cases where we don't are cases Elster (1983) famously calls states that are essentially by-products (I will sometimes refer to them also as essential by-products). Such states "can only come about as the by-product of actions undertaken for other ends" (Elster 1983, 43). Spontaneity is one of Elster's examples, as is sleep, political participation, forgetfulness, some aesthetic value, and much more[51]. Essential by-products may be of value, but even when they are, this doesn't give us reason to pursue them, because pursuing them guarantees not getting them. The way to get them is to sneak up on them, intentionally pursue other things, thereby achieving the essential by-products as well. (And sometimes there are indirect means of self-management that may work – think, for instance, about intentionally having a drink in order to become more spontaneous).

[51] Here is Elster in a hyperbolic moment: "… all good things in life are essentially by-products." (Elster 1983, 108).

Might autonomy be an essential by-product[52]? In particular, is epistemic autonomy? If it is, this will sever the tie between questions about reasons to pursue autonomy and autonomy being of value. Perhaps, in other words, it *is* of value – even if purely epistemic value – to shape one's beliefs in accordance with one's deep epistemic commitments, and by making one's own mind without outsourcing, but perhaps this is the kind of value that can only be sneaked up on, or achieved indirectly? Perhaps it is only by pursuing truth (or avoidance of falsehoods, or accuracy, or some such) that one may come to instantiate the essential by-product value of epistemic autonomy? If so, this will explain why it is that it doesn't make sense to refuse (e.g.) to outsource just in order to achieve epistemic autonomy, consistently with still attributing value to epistemic autonomy.

How plausible is it that epistemic autonomy is an essential by-product? To me, quite plausible indeed, perhaps partly because at least on many occasions this seems to me to be true about autonomy in general. Think, for comparison, of excellence. One can value excellence, I guess. But one doesn't pursue excellence directly[53]. Excellence is something that may characterize one's pursuit of other things, and may be achieved in pursuing such other things. Excellence, even if of value, is not itself the *end* one pursues. Rather, excellence's significance seems to be primarily *adverbial*, one pursues other things excellent*ly*. Of course, excellence itself can be pursued indirectly, one can utilize self-management tools in this way, and so on. But still, excellence is not, at least not primarily, itself an end, or something we have reason to pursue. I would say precisely the

[52] Elster (1983, 52) does briefly mention something about autonomy in this context. But he doesn't argue that it's an essential by-product.

[53] We may want to distinguish between different ways in which essential by-products cannot be pursued intentionally and directly: Sometimes the impossibility is conceptual or logical, as is arguably the case, for instance, with spontaneity and forgetfulness. At other times the impossibility is empirical – Elster's convincing example here is that of sleep, where there's no logical impossibility to the thought of trying hard to fall asleep, but we know, on empirical grounds, that this is highly unlikely to succeed. (Elster (1983, 57) is explicit about the distinction between conceptual and empirical impossibility here.) Yet another kind of impossibility – Elster is not explicit about it, but some of his examples fit this description – are cases where there's neither a conceptual nor an empirical impossibility to intentionally and directly pursue the relevant essential by-product, but doing so will defeat the purpose of doing so, will render the relevant thing achieved devoid of the value it sometimes has. It may be thought of, then, as a kind of evaluative impossibility or self-defeat. Think, for instance, of pursuing heartache because one believes that a life devoid of heartache is less rich, or Elster's (1983, 59) examples of despair, some aesthetic values (1983, 79), and political participation (1983, 100). Pursuing excellence directly and intentionally is often empirically impossible, and sometimes it is self-defeating in the evaluative kind of way.

same about autonomy, at least as understood in the first-person. Autonomy is not, as it were, its own end or project. Rather, one engages in other projects autonomous*ly*. One achieves autonomy – the value of being a part-author of one's own life – by shaping and telling a life story full of other value. Not – or not primarily, anyway – by pursuing autonomy directly. Indeed, there seems to be something objectionably fetishistic about someone guiding their actions and decision not by substantive values and projects, but merely by the aim of becoming autonomous or manifesting autonomy[54].

This does not mean, of course, that autonomy may not be reason-giving. For one thing, one person's autonomy may give *another person* all sorts of reasons[55]. And even in one's own case, the value of autonomy may give reasons for other actions – indirectly pursuing autonomy by sneaking up on it (say, engaging projects that are more likely to lead to an autonomous life). But it is, I tentatively conclude, quite plausible to think that autonomy is essentially a by-product.

And this seems especially true of epistemic autonomy, on the assumption that it *is* of value. If one conducts one's epistemic affairs – pursuing truth or understanding or some other purely epistemic aim, if there are any – according to one's deep epistemic commitments, and utilizing one's own devices, one will thereby be sneaking up on the value of epistemic autonomy. But this doesn't mean that one should treat achieving epistemic autonomy as something one has (purely) epistemic reasons to do, and certainly not as something that can justify ignoring evidence, or in some other way paying a price in the hard currency of likelihood of truth. Doing that – pursuing epistemic autonomy directly in this kind of way – will amount to an instance of "the fallacy of striving, seeking and searching for the things that recede before the hand that reaches out for them." (Elster 1983, 108).

[54] Here too I think that what we have is not the essential by-product that it's logically or conceptually impossible to pursue directly. Rather, it's a combination of the empirical impossibility of pursuing it directly, at least often, and the evaluative one – as is exemplified by the fetishism point in the text.

[55] See here Raz's (…) claim that one's wellbeing – for him, closely tied to the value of autonomy – gives reasons for actions not, or not primarily, for one but for others.

I hope that you will agree that this way of thinking of autonomy in general and of epistemic autonomy in particular is interesting and promising. But how is it related to the task at hand, namely explaining the disanalogy between the role autonomy plays in the practical and the epistemic domains? As I said in the beginning of this section, these observations show how there may be less of a disanalogy here than meets the eye. For now we know that attributing value to epistemic autonomy is consistent with epistemic autonomy not being directly reason-giving, and so perhaps we should be more willing to attribute value to epistemic autonomy, at least sometimes.

Some disanalogy nevertheless survives. To see this, we can focus first on the reason side, then on the value side. First, then, reasons: In the practical case too, I think, autonomy is often an essential by-product. But not always: In the practical domain, it seems plausible to assume, at least sometimes the value of autonomy is directly reason-giving in the way that epistemic autonomy is not. This, after all, was what we said about the case of choosing one's partner, where it seemed at least an intelligible possibility to refuse the (otherwise) optimal choice precisely because of the value one places on autonomy. This alone suffices to show that at least some disanalogy between the practical and the epistemic domain remains. And on the value side too, I am not sure that the two domains are exactly analogous. I find the thought that there is (not merely instrumental) value in shaping one's life according to one's commitments and values, and with one's decisions, extremely plausible. I don't find the epistemic analogue of this claim remotely as plausible, and this even when we fully appreciate that accepting such a value need not immediately commit us to it generating epistemic reasons (because epistemic autonomy may be essentially a side-constraint).

For both these reasons, then, I believe that some disanalogy between the evaluative and normative status of autonomy remains. But there's less of it than we may have thought before appreciating the plausible possibility that epistemic autonomy is an essential by-product.

I find the idea that autonomy – epistemic and otherwise – is often essentially a by-product highly promising, and there's much more to be said about it. I mention two further points in a long

footnote[56]. For our purposes here, though, the point above – about there being less of a disanalogy than we may have thought – will suffice.

7. Conclusion: Not All Explanations are All That Beautiful

We set out to explain the apparent disanalogy between autonomy's role in the practical and the epistemic domain. An ideal solution would be one, clean, beautiful explanation that reveals something deep and clear about, perhaps, the nature of actions and the nature of beliefs, in a way that makes the disanalogy precisely what you would expect. This is the kind of explanation I wanted to come up with when I started thinking about this disanalogy.

But we don't always get what we want. Instead, what we've arrived at is the following messier picture. First, there's less of a disanalogy than meets the eye. Second, there's a host of explanations, all perhaps shouldering some of the explanatory burden but none of them the ideal, one-swoop beautiful explanation that may have been hoped for. Considerations of incoherence and of the distinction between reasons of the right and of the wrong kind seem to render the disanalogy more intelligible, perhaps to place it in a wider context that makes some sense of it, which is already

[56] Recall the distinction between local and global autonomy – both practical and epistemic. Perhaps it may be thought that both are of value, that local autonomy is essentially a by-product, and that global autonomy is sometimes directly reason-giving. That is, in the epistemic case, that on no specific case of shaping a belief does it make sense to treat one's autonomy as giving a reason (not to outsource, say), but that it does make sense to directly take measures to shape one's epistemic life so as to make it more autonomous. If so, there will be a tension between the local and the global – a strategy that is rational at each relevant local point (no weight to autonomy) will lead to a globally irrational strategy. If this is correct, then the problem here is similar in structure to that discussed in the context of the self-torturer paradox (Quinn 1990). Perhaps recognizing that sometimes a value is present in both the local and the global context, but is an essential by-product in one of them, can go some way towards explaining the problem in such cases – the general problem highlighted by the self-torturer paradox, not just in the context of epistemic autonomy – and perhaps even towards solving it. Recall also the distinction between right-kind- and wrong-kind-reasons. It is natural to think that there's some relation between that distinction and the category of essential by-products. Perhaps reasons to directly pursue essential by-products are always of the wrong kind, or perhaps, more plausibly, some subset of them is (perhaps, for instance, those such that directly and intentionally pursuing them defeats their value; see footnote 50 above). Elster (1983, 51) himself seems to suggest something along these lines when he ties the fact that beliefs don't respond to instrumental reasons with the fact that useful beliefs are essentially by-products. If there is such a relation, it is natural to expect some close relation to arise between the relevance in our context of the distinction between reasons of the right and of the wrong kind and the relevance of the category of essential by-products.

some explanatory progress, but it's not clear how deep at the end of the day that progress is. The fact that the practical is so rich in values compared to the relative paucity of epistemic values seems to explain more here, though again, perhaps this explanation too does not deliver all the explanatory payoff that could be desired. And perhaps that's all that can be said by way of explaining the disanalogy. At least, it's all I now have to say about it.

Such messy explanations sometimes have advantages. Here's one: Had we found the kind of large and deep single explanation, in terms of a profound difference between the nature of beliefs and actions, we would have been forced to accept the disanalogy as applying in a clean-cut way as well – autonomy would have had to play one role in the case of *all* beliefs, and another in the case of *all* actions. If I'm right, though, that the explanation is messy, it opens the door to interesting hypotheses about the messiness of the explanandum: Perhaps, for instance, autonomy may be just as normatively irrelevant in some action cases as it arguably is in the belief case. In some cases, I've argued, it makes sense to be willing to pay a price in the quality of the decision just in order to make it oneself, but perhaps this is so for some practical cases – perhaps personal ones like that of choosing a romantic partner – but not in others. It does seem problematic, I think, to be willing to pay a price in the quality of a *moral* decision just in order to make it oneself[57]. And perhaps this is partly so because in moral decisions there's less relevant pluralism and incommensurability than in many personal cases (or perhaps for some other reason)[58]. And perhaps – though I can't think of an example – there are even epistemic cases that behave, in terms of the relevance of autonomy, more like the practical ones.

[57] See my "A Defense of Moral Deference" (2014).
Notice that this points pulls, to an extent, in the opposite direction from the observation above (in section 2.4) that in political contexts we seem to find outsourcing more problematic.

[58] Relatedly, perhaps what does the difference here, at least partly, is that in many of the practical cases we've been discussing the relevant reasons have to do with the agent's own wellbeing or are in some other way personal. Epistemic reasons seem – at least paradigmatically, but perhaps always – to be impersonal. And moral reasons, or at least many of them, seem to be in this way more like the latter than like the former. I thank Ittay Nissan-Rozen for relevant discussion.

Selim Berker (2013), "The Rejection of Epistemic Consequentialism", *Philosophical Issues* 23, 363-387.

(2018), "A Combinatorial Argument against Practical Reasons for Belief", *Analytic Philosophy* 59, 427-470.

Daniel Brudney and John Lantos (2011), "Agency and Authenticity: Which Value Grounds Patient Choice?", *Theoretical Medicine and Bioethics* 32, 217-227.

J. Adam Carter (2022), "Epistemic Autonomy and Externalism", in Matheson and Lougheed (2022), 21-38.

Finnur Dellsen (2022), "We Owe it to Others to Think for Ourselves", in Matheson and Lougheed (2022), 306-322.

Jon Elster (1983), *Sour Grapes* (Cambridge: Cambridge University Press).

David Enoch, (2011) "Not Just a Truthometer: Taking Oneself Seriously (But Not Too Seriously) in Cases of Peer Disagreement", *Mind*, 119, 953-997.

(2014a) "Authority and Reason-Giving", *Philosophy and Phenomenological Research* 89, 296-332.

(2014b) "A Defense of Moral Deference", *The Journal of Philosophy* 111, 1-30.

(2016) "What's Wrong with Paternalism? Autonomy, Belief, and Action", *The Proceedings of the Aristotelian Society* 116, 21-48.

(2017) "Hypothetical Consent and the Value(s) of Autonomy", *Ethics* 128 (2017), 6-36.

(2022) "Autonomy as Non-Alienation, Autonomy as Sovereignty, and Politics", *The Journal of Political Philosophy* 30 (2022), 143-165

(Manuscript), "Backgrounding Agent-Relativity".

(Work in progress) "Revisiting Raz on Autonomy".

David Enoch and Levi Spectre (forthcoming), “There Is No Such Thing as Doxastic Wrongdoing”, forthcoming in *Philosophical Perspectives*.

Elizabeth Fricker (2006), “Testimony and Epistemic Autonomy”, in Jennifer Lackey and Ernest Sosa (eds.) *The Epistemology of Testimony* (Oxford: Oxford University Press), 225-245.

Miranda Fricker (2007), *Epistemic Injustice: Power and the Ethics of Knowing* (Oxford: Oxford University Press).

Gilbert H. Harman (1965), “Inference to the Best Explanation”, *Philosophical Review* 74, 88-95.

Alison Hills (2009), “Moral Testimony and Moral Epistemology,” *Ethics* 120, 94–127.

Robert Hopkins (2007) “What Is Wrong with Moral Testimony?” *Philosophy and Phenomenological Research*, 74, 611–34

Michael Huemer (2005), “Is Critical Thinking Epistemically Responsible?”, *Metaphilosophy* 36, 522-531.

Pamela Hieronymi (2005), “The Wrong Kind of Reason”, *Journal of Philosophy* 102, 437-457.

Elizabeth Jackson (2022), “What’s Epistemic about Epistemic Paternalism”, in Matheson and Lougheed (2022), 132-148.

Jonathan Matheson (2022), “Why Think for Yourself?”, *Episteme …*

Jonathan Matheson and Kirk Lougheed (eds.) (2022) *Epistemic Autonomy* (New York and London: Routledge).

C. Thi Nguyen (2018), “Expertise and the fragmentation of intellectual autonomy”, *Philosophical Inquiries* 6 (2):107-124.

Philip Nickel (2001), “Moral Testimony and Its Authority,” *Ethical Theory and Moral Practice* 4, 253–66.

Marina A. Oshana (2006), *Personal Autonomy in Society*. New York: Routledge.

Warren S. Quinn (1990) “The Puzzle of the Self-Torturer”, *Philosophical Studies* 59, 79-90.

Wlodek Rabinowicz and Toni Rønnow-Rasmussen (2004), “The Strike of the Demon: On Fitting Pro-Attitudes and Value”, *Ethics* 391-423.

Joseph Raz (1986) *The Morality of Freedom* (Oxford: Oxford University Press).

(2006) "The Problem if Authority: Revisiting the Service Conception", 90 *Minnesota Law Review*, 1003-1044.

Mark Schroeder (2012), "The Ubiquity of State-Given Reasons." *Ethics* 122, 457–88.

NIshiten Shah (2003), "How Truth Governs Belief", *The Philosophical Review* 112, 447-482.

Ernest Sosa (2003) "The Place of Truth in Epistemology", in M. De-Paul and L. Zagzebski (eds.), *Intellectual Virtues: perspectives from ethics and epistemology* (pp. 155–179). Oxford: Oxford University Press.

Bas C. van Fraassen (1989), *Laws and Symmetry* (Oxford: Oxford University Press).

Ruth Weintraub (2013) "Can Steadfast Peer Disagreement Be Rational?", *The Philosophical Quarterly* 63, 740–59.

Roger White (2005) "Epistemic Permissiveness", *Philosophical Perspectives* 19, 445–59.

Linda Zagzebski (2012), "Ethical and Epistemic Egoism and the Ideal of Autonomy", *Episteme* 4, 252-263.

(2013) "Intellectual Autonomy", *Philosophical Issues* 23, 244-261.