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 this diagnosis. And let me remind you of the two points with which I concluded the previous section: First, I noted there

[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).

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].

6. Details and Implications

[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.)

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

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

[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.

[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.

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

[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.

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 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

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

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 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

[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.

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.

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

[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.

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 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].

[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).

[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.

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 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

[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.

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 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

[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).

(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 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.

[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.

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.

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

[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).

– 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 all65. 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 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]

[65] I thank Shir Nidam, Roy Kreitner and Courtney Cox for this objection.
[66] For raising objections along these lines, I thank Korbinian Rüger and Jörg Löschke.
[67] In the sense developed by Elster (1983, Chapter 2).

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.

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

[68] In my "Epistemic Autonomy May Not Be a Thing" (manuscript).
[69] I thank Till Grüne-Yanoff for this case.

"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 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

[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).

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).