10

The ethics of using boosts to change behaviour

Till Grüne-Yanoff

So much of what policymakers do is aimed at behaviour change. Striving to get individuals to reduce their carbon footprint, litter less, cause fewer accidents, or place fewer burdens on the healthcare system, policymakers draw on various tactics. Traditionally, they have deployed the 'carrot' of incentives, the 'stick' of regulatory penalties or the 'sermon' of public messaging. However, faced with low success rates and high political and financial costs of these approaches, policymakers are keen to employ all the help they can get. They have been increasingly turning to subtle and more sophisticated applications of insights from behavioural science (see the chapter by in this volume Abrahamse and Contzen (2025)) to develop policies that achieve the desired changes in behaviour. And the potential impact of these approaches is further expanded by various behaviour change technologies becoming more widely available and easily integrated into public policy initiatives. With technologies at policymakers' disposal for monitoring, notifying, personalizing and even intervening, pressing questions emerge about what principles ought to guide the use of these technologies and their implementation into public policies (see the overview in this volume by Oinas-Kukkonen and Agyei (2025)).

In this chapter, I discuss the ethical acceptability of 'boosting', contrasting it with other behavioural interventions, especially 'nudging' (e.g. Thaler and Sunstein 2009). I will argue that boosting, with its focus on enhancing individual competences rather than manipulating choice architectures, offers a promising framework for navigating the ethical complexities of policymaking, including those that make use of BCTs. By emphasizing the development of skills and competences, boosting potentially aligns more closely with values of autonomy and self-determination than does 'nudging', with its tendency to bypass the agency of those whose behaviour they seek to change.

Boosts can be defined as *policymakers' effortful interventions on groups of people with the aim of improving people's competences and through that improve their decisions according to their own standards* (Grüne-Yanoff and Hertwig 2016; Hertwig and Grüne-Yanoff 2017). This might lead critics to one of the following ethical complaints. First, because they intervene in people's behaviour, boosts might be criticized for limiting people's *freedom*. Second, because they aim to affect people's beliefs and attitudes, boosts might be criticized for undermining people's *autonomy*. Third, because boosts are often justified as improving people's decisions by their own standards, they might be

criticized as *paternalistic*. Finally, because they typically intervene in groups of people, for whose individual members they might exhibit different degrees of effectiveness, boosts might be criticized for increasing *inequalities* in such groups. The following sections engage with each of these criticisms in turn. With reference to a series of cases, I articulate what each of these ethical principles demand of behavioural interventions, and whether the concept of boosts – as well as the examples I discuss – satisfies them to a higher degree than other kinds of behavioural interventions.[1]

How boosting preserves freedom

Boosts interfere in people's choices. Any choice interference is suspect of being freedom-reducing. Therefore, do boosts reduce freedom? Many authors have defined freedom as the *absence* of obstacles, constraints or interference from others. More specifically, an agent's freedom is reduced if and only if (i) the set of choice options the agent has is reduced, and (ii) this reduction is attributable to another agent. Because it is characterized by an absence, the concept is often called *negative liberty*.[2]

The second condition states that whatever boosts do, their effect is at least partially attributable to another agent external to the affected individual. That seems correct: boosts are policymakers' interventions aiming to improve people's decisions. Some risk literacy boosts, for example, transform opaque representations (e.g. single-event probabilities) into transparent ones (e.g. frequency-based representations). Increased risk literacy, so the argument goes, improves people's understanding of the uncertainty they face and allows them to make better-informed choices (Gigerenzer et al. 2007; Spiegelhalter, Person and Short 2011). Some motivational boosts, to take a different illustration, train people to make strategic use of automatic processes. By encouraging people to describe how they intend to implement a plan, for example, the implementation intention boost makes it more likely that people realize their long-term desires at the expense of short-term whims (Gollwitzer 1999). Thus there is another agent (the policymaker) who intends to cause some change in people's choices, and only those boosts that actually have the desired effect on people's choices will be considered successful.

What about the first condition, then: do boosts reduce people's set of choice options? This is a trickier issue. Negative liberty is sometimes said to be violated only through someone else's physical coercion. When I am grabbed, held down, locked up, pushed back, then my liberty is obviously limited. Boosts do not engage in such tactics, nor do virtually any of the policies or technologies discussed by other authors in this book. In this very narrow sense, behavioural policies generally do not constrain liberty. But such a concept is obviously *too narrow* to capture what most people mean when they talk about violation of liberty. Consider the case of Jill, who would park her car in front of her house, but parking is restricted there and incurs a hefty fine, so she doesn't. She has the option to park her car in front of her house, but the legislator has shaped these options in such ways that she doesn't *want* to choose it. She is not physically coerced. What prevents her from choosing these options is her own motivation. Yet most agree

that the legislator constrains Jill's liberty. How can the concept of negative liberty capture this intuition? The trick is to understand that the options in question are more complex than simply 'parking on front of one's house'; they come attached with various costs and benefits. People want to choose these options only if the attached costs are not too high. The policymaker, by increasing the costs attached to these options – without directly removing them – makes them sufficiently unattractive in the eyes of the subjects that they would not want to choose them anymore. Jill would have parked in front of the house had the legislator not shaped the options in such ways that they no longer wanted to choose them. More formally, I might say that the parking regulation removed the option <park here, no fine> from her option set, leaving her with the option <park here, pay fine>, which she does not want. Thus, it is the policymaker's external influencing the options, in conjunction with the subjects' preferences, which brings about the changed behaviour. And because of this external influencing of options the policies make Jill less free.[3]

Negative liberty understood in this way offers insights into how behavioural policies, including boosts, might affect liberty. Are behavioural policymakers, like the legislator, nonetheless permitted to constrain freedom? That depends, as I now argue, on *how* the interventions affect people's choices. Consider the following example.

> Karen drives a big SUV. The policymaker mandates that all cars prominently display their fuel economy in gallons per mile, thus allowing everybody to judge her choice of car model. Karen does not want to be identified as driving a gas guzzler, so he changes to a more economical car.

There are no threats or coercion in this case, and yet I think that it makes Karen less free. The intervention, which actually has been proposed as a nudge (Thaler and Sunstein 2009, 192), raises the immaterial costs of SUV ownership, so that Karen no longer desires it – not much different than if a fine was imposed on driving an SUV. I see no reason why the liberty-reducing effects of material and immaterial costs should be assessed differently, and thus conclude that if Jill was made less free, so was Karen.

Now consider the following example:

> Larry discounts the future in strongly hyperbolic ways: when asked now, he prefers increasing in a year from now his retirement savings contribution (and thus consuming less next year) to keeping it unchanged. But he also prefers keeping things unchanged now, allowing him to consume more in the current year. The policymaker, who wants Larry to save more, therefore asks Larry to sign now a new retirement savings plan for next year.

This is again an intervention that behavioural scientists have proposed (Thaler and Benartzi 2004). Here, again, there is no direct coercion: the influence goes through Larry's own desires. And as in Karen's case, what Larry desires depends on the policymaker's influence. Yet in contrast to the former, the intervention does not affect his options at all; it only changes when Larry is choosing between them. Because Larry

has time-dependent preferences, this intervention changes his choice, even though the options and Larry's preferences remain unaffected. Consequently, this policy does not violate his negative liberty, either. Those who feel uneasy with this conclusion – including us – need to look for another concept of liberty in order to explain their intuition why these interventions make Larry less free. But before I get there, there is one other scenario to consider.

> Maggie enjoys eating mature cheeses. She also suffers from unpleasant digestion issues. Her doctor, suspecting a connection, helps her to determine whether she is lactose intolerant, which it turns out she is. The GP recommends a diet without cheeses to avoid the symptoms.

Although the case does not involve coercion, the doctor does influence Maggie's attitudes in such a way that she chooses against what she once preferred. However, this influence does not operate through changing the options under consideration. The intervention tells her that cheese has adverse effects on her well-being, and she therefore no longer desires to eat it. Had she known this before, she would not have had an all-things-considered preference for cheese. Thus, the intervention does not change the options (as it did in the case of Jill and Karen); nor does it harness some cognitive bias in order to make Maggie choose not to eat cheese (like in Larry's cases). Rather, it makes clearer what the (unchanged) option actually *is*. Such an intervention is not freedom-reducing: it does not prevent Maggie from eating cheese, nor does it deteriorate the option (e.g. through a fine or an additional psychological cost); instead, it reveals more of the true nature of the option, thus making Maggie realize that she never wanted it in the first place.

Maggie's case concerns a medical intervention on a single individual, while boosts are interventions on populations. Yet it shares all ethically relevant features with boosts. Boosts, according to our definition, aim to improve people's competences and through that improve their decisions. 'Competences' is a concept with many meanings, but most agree that it is a 'roughly specialised system of abilities, proficiencies, or skills that are necessary or sufficient to reach a specific goal' (Weinert 2001, 45). In this sense, Maggie's doctor's intervention is competence-enhancing. In contrast, interventions like those on Jill, Karen and Larry do *not* improve competences. They affect behaviour through other pathways and thereby sometimes reduce the subjects' freedom. Increasing competences, however, can *never* decrease negative freedom.[4] As illustrated in the cases of Maggie, increasing competences helps reveal the true nature of the available options without altering them or supports the creation of new options, thus not only preserving freedom but even increasing it. Boosts, therefore, are not freedom-reducing, in contrast to some other behavioural policies.

Of course, this argument presupposes a notion of negative liberty, which focuses on options available to an agent, and the actual reduction of options by another agent's doing. As I noted earlier, this notion sometimes gives unintuitive results. For example, a benevolent dictator or a powerful but cautious hypnotist might not constrain subjects' choices, even though she has the power to do so. From a negative freedom perspective,

because they enjoy the absence of actual interference, the subjects are considered not constrained in their freedom. But that seems counterintuitive: the dictator and the hypnotist hold the means to control the subjects' choices, and can do so at their arbitrary will. The alternative concept of *republican freedom* characterizes liberty by the absence not of actual interference, but by the absence of any structural dependence on arbitrary power or domination – an institutional arrangement of laws and checks that guarantees that such acts of arbitrary control are ruled out (Skinner 2002; Pettit 1997).

The republican freedom concept helps us to clarify the distinction between the intervention on Larry vs. the intervention on Maggie, which the negative liberty concept could not properly differentiate. Larry does not experience a reduction of choice options. However, his choices are subject to a systematic policy influence, and this constitutes an arbitrary power that the policymaker has over him. Thus, from a republican perspective, Larry suffers from a reduction of their freedom. Contrast this with Maggie's case: the GP and the therapists can also effect a change of Maggie's choices, making her forego things that she highly valued before. But this power is not arbitrary. The doctor cannot at will make Maggie switch back and forth between choosing different options, as the policymaker seems to be able to do with Larry. Instead, doctors provide knowledge and competences that their patients will adopt only if they deem them true or useful. As a result of the intervention, Maggie learns more about the world she inhabits and acquires strategies to achieve her goals in a better way. Once she has acquired these competences, there is no way of going back.[5] Thus boosts, in contrast to interventions exemplified in Larry's case, are not liberty-reducing even under a republican interpretation.[6]

How boosting preserves autonomy

Behavioural policies might be morally or politically unacceptable even if they do not violate negative or republican freedom. For example, even if an intervention does not reduce an agent's set of options, it might reduce their autonomy – their ability to govern or realize themselves, acting out of their *own* power to decide what to do. Consider this scenario:

> Nigel suffers from a disease that causes strong pain, which keep him confined to his bedroom. A new drug, prescribed by his doctor, allows him to again socialize and work. Nigel also becomes addicted to the drug: he strongly craves consuming it, and spend many hours of his newly-won freedom to secure its supply and to indulge its effects. Nigel wishes he were not addicted; a vague sense of loss looms over him, he would prefer spending all these hours on some more rewarding activity instead, and he cannot justify to himself why he continues consuming the drug.

Nigel is an exemplary case of a heteronomous, non-self-governing agent, according to the following four criteria. First, his urge to consume the drug conflicts with his other

desires for a more rewarding activity, with his sense of loss, and his explicit preference not to be addicted. Those desires arguably represent *his own point of view*. As he follows an urge that conflicts with these, his resultant actions result not from his own power to decide but from something else. This *incoherence* makes his actions heteronomous. Second, Nigel has no reason for indulging his drug habit, but he has at least one reason against it: it prevents him from more rewarding activities. Nevertheless, Nigel does what he has no reason to do. He thus fails to appreciate what he has reason to do and is not moved by these reasons. This lack of *reasons-responsiveness* makes his actions heteronomous. Third, Nigel fails to draw inferences from what he wants and believes (in particular relating to those other more rewarding activities), and fails to rationally reflect upon his addiction from that perspective. A rational agent, in contrast to Nigel, would recognize that it is the consequences of the addiction that prevent her from engaging in those more rewarding activities, and if concluding that those are more desired than continued drug consumption, would take whatever means available to suppress the urge. Nigel does not do this, thus exhibiting a *non-responsiveness to practical reasoning* that makes his actions heteronomous. Fourth, Nigel's actions can be fully explained as the effects of his medical history and his subsequent addiction. Because his actions are explainable by references to causal powers that are independent of him, his actions are heteronomous.[7]

These four criteria do not always coincide; a lot of the literature focuses on such more complex cases, and a lot of ink has been spilled in favour of identifying autonomy violation with one criterion or another. I will not commit to one single view here, but instead follow a cluster approach: if all or most of these criteria are satisfied, as in Nigel's case, then I consider it a violation of autonomy. After all, autonomy is intuitively seen as concerning the independence and authenticity of the desires (values, emotions, etc.) that move one to act in the first place, and each of the criteria tries to make precise one aspect of this vague intuition. Trying to meet many of these criteria might be the best one can do; in any case, it is the best I can do in the confines of this chapter.

Note that autonomy characterized in this way differs substantially from negative freedom. Arguably, taking the drug *increased* Nigel's freedom: he is not confined to his bedroom anymore, but can venture outside to socialize or work. The subsequent addiction, however, makes him spend this new-gained freedom in ways that do not realize his own desires and values; instead of his genuine, authentic desires, it is this urge, imposed by the addiction and – at least sometimes – disavowed by Nigel. The doctor's intervention – prescribing the drug – thus increased his freedom but decreased his autonomy. Conversely, there might be cases where an intervention reduces freedom but increases autonomy – for example, when a reduction of options leads agents to reflect more on their desires, makes them more perceptive of their genuine reasons, reduces the causal influence of external powers, and increases the agent's internal coherence.[8] To assess a policy for its autonomy-reducing effects is thus independent from investigating its freedom-reducing effects.

Furthermore, autonomy can be compromised *without* any clearly identifiable external influence. A person who lacks the necessary knowledge and competences might have difficulties reflecting on their desires, might not be perceptive of their genuine

reasons, or might not be capable of achieving internal coherence – even without, or perhaps even because of, the absence of any external interference. Furthermore, even if another person does cause the autonomy reduction, it might often not be intended, as in Nigel's case. Thus, a person's desire might be heteronomous, according to three of the four criteria discussed above, without being intentionally caused by another agent. This again stands in contrast to negative freedom, which cannot be reduced without another person's external influence.

Autonomy of desires is thus determined by the *ways* this desire is formed – whether as a result of rational reflection on all the options available, or as a result of pressure, manipulation, impulses or ignorance.[9] This might seem like a very demanding a concept, but in fact it isn't. Take for example urges and impulses. Acts driven purely by impulses cannot be autonomous. However, this does not mean that the autonomous agent cannot choose according to their impulses or urges. All that is required is that the agent, upon rational reflection, embraces the urge (in the sense of ensuring it is coherent with their other desires, or by citing reasons for acting in such 'impulsive' ways). Imagine Nigella, who experiences the same conditions as Nigel, but, unlike him, is genuinely content with her addiction; does not think her time could be spent in a more rewarding fashion elsewhere; and justifies her continued drug use by the comfort it provides. Nigella is an autonomous addict. The important difference between Nigella and Nigel is that she regards her own judgement about continuing her drug habit as authoritative, while he does not. Thus, it is the attitude (expressed as coherence, reflection or genuine reasons) towards the drivers of one's actions – the desires, emotions, urges or whatever – that makes actions autonomous or heteronomous. Agents, therefore, can follow their impulses, or follow the command or advice of someone else, and nevertheless remain autonomous, as long as they regard their own commitments to these acts, their own judgements and decisions that they should act in this way, as authoritative.

Let's apply the autonomy concept to the behavioural policy scenarios I discussed above.[10] Karen is fine with driving a gas guzzler but doesn't want others to shame her for it. As soon as she is required to prominently display her car's economy, she adjusts her actions in accordance with her own desires and switches to a different model. There is no hint in the scenario that these are not her own reflected and reasoned desires, so her switch to a different model is an autonomous action. The fuel economy display mandate thus does *not* reduce Karen's autonomy, although it reduces her freedom.

Larry's preferences between saving and consuming depend on his position in time relative to the time of potential consumption. The closer he comes to this point, the more he prefers consumption over saving. In the context of managing one's personal assets, such preferences cannot be considered coherent: every plan that Larry forms in this regard will be defeated by his later preference change. Reflecting on this should lead an autonomous person either to give up on such plans entirely or to take measures that such stabilize plans (i.e. various self-committing strategies). Given that Larry has not done so, his sensitivity to intervention might indeed be considered evidence of his lack of autonomy in this regard. Indeed, defenders of nudges often argue that the subjects being swayed so easily shows that their relevant preferences

before the intervention took place were not autonomous and that one cannot reduce the autonomy of an already heteronomous agent.[11] I do agree with this argument for cases in which the policymaker knows that the act in question is fully heteronomous. However, a person can authoritatively judge that the choice options are a toss-up; such incomparability judgements can be perfectly coherent with a person's other attitudes, reflecting authentic desires and genuine reasons. Being undecided prior to an intervention is *not* evidence that she lacks autonomy in this regard.[12] Consequently, if there is reason to believe that Larry lacked autonomy to start with, the intervention should not be considered autonomy-reducing. If, however, the intervention swayed Larry to one of two options which he rationally held to be incommensurable, then it should be considered autonomy-reducing.

Thus, interventions are generally autonomy-reducing if subjects (who were not fully heteronomous at the start) are not aware of them, if they foster impulsive reactions or prevent reflection. In contrast, interventions that are transparent and inhibit neither reflection nor reasoning are generally not autonomy-reducing. In particular, this holds for competence-fostering interventions like boosts: they either provide information that subjects can employ in their deliberation, or they provide subjects with new choice alternatives that subjects can make use of if they so desire. Maggie, for example, when learning about her lactose intolerance, might choose not to consume any more cheese. But if she does, then that choice must be motivated by her own desires or reasons – the information alone could not motivate her abstinence. More generally, to be boosted successfully, subjects must be motivated at various stages: (i) to expose themselves to the intervention, (ii) to engage with its content to acquire the competence, and (iii) to apply the thus-acquired competence for particular purposes.

For boosts to work, therefore, what is required is that they engage people's own motivation. Because they only work by engaging people's own desires, boosts *cannot* be autonomy-reducing: in every case, the motivational drivers of their boosted action are their own. This even holds when the boost affects values or desires itself. Indeed, not only do boosts not reduce the autonomy of the boosted, in many cases, boosts actively increase the autonomy of the boosted. The ability to form coherent motivations, to rationally reflect on one's motivations or to identify genuine reasons for them all are forms of cognitive competences. Boosts that improve people's abilities to, for example, search and screen information, organize their deliberation and come to clear conclusions all contribute to such competences, and therefore directly increase a person's autonomy.

The modest (and unproblematic) paternalism of boosting

Paternalism involves interference with a person's actions or knowledge for the purpose of advancing that person's good. Boosts are paternalistic in this sense.[13] Paternalistic interventions are often criticized for being freedom- and autonomy-reducing, along the lines discussed in the previous two sections.[14] But paternalistic behavioural policies have also have been criticized on epistemic grounds – namely that the policymaker

rarely has the knowledge needed in order to be reasonably certain that the interference indeed is beneficial for them. It is those arguments that I engage with in this section.

Such *knowledge-deficit* arguments against paternalism consist of the following steps. First, they assume that the paternalism in question respects subjective evaluations as its normative basis. That is, it considers the satisfaction of a person's genuine preferences and values to constitute what is good for that person. Such 'libertarian paternalists' seek to improve people's choices so that they are 'better by their own lights' (Sunstein and Thaler 2003, 1163). Other paternalists, in contrast, rely on objective principles instead – for example, a medical doctor who performs a stenting procedure on a heart attack patient, even though the patient expresses a credible desire not to be treated. It is much more difficult to know what an individual really, genuinely wants than to determine that she is better off according to some objective standard (e.g. her chances of survival). The knowledge-deficit critics claim that libertarian paternalists must know the details of people's evaluations in order to justify their preferences, and that such knowledge in the required detail is unattainable because it 'consists of subjective attitudes, perceptions, beliefs, and tastes. It includes personal strategies whose application depends on the idiosyncratic environments, routines, and social contexts of countless individuals' (Rizzo and Whitman 2020, 279). Thus, these critics conclude, the paternalists do not even know what the good is that they claim to help people attain.

Second, even if the goal was known, the paternalist would have to justify the intervention as promoting this goal. This requires knowledge not only about the intervention itself but also about the subject being intervened on. The doctor treating the heart attack patient of course needs to know how to perform the stenting procedure but also needs to know that the patient does not suffer from adverse conditions which, in combination with this treatment, would worsen the patient's condition further. Doctors therefore get extensive training in correct diagnosis and treatment procedures to ensure that they have the required knowledge.

Critics of paternalistic behavioural interventions have argued that behavioural policymakers, in contrast, often do not and cannot have the required knowledge. The factors that allegedly prevent people from making decisions in their own best interest are highly context-dependent, varying from person to person, place to place, and time to time; they also vary with the experimental method of how such factors are elicited. The same holds for the way people react to the various proposed interventions. Furthermore, such policies typically intervene on a population, and policymakers typically do not know the distribution of the relevant properties in the population. Therefore, behavioural policymakers don't really know the ailment of their 'patients', and thus cannot really know what they are treating and what result to expect. 'If behavioural paternalism is to clear the high bar its own creators have set for it – to provide an evidence-based policy programme that reliably improves personal welfare from the perspective of the individual – then its practitioners need to have the vast body of knowledge I have discussed. For most proposed interventions, they do not have it.' (Rizzo and Whitman 2020, 280). As this knowledge-deficit is pervasive, paternalistic behavioural policies generally are not justified.

To some extent, I agree with this knowledge-deficit criticism. In some of our earlier scenarios, the policymaker seems to lack the knowledge required to justify interventions. Think, for example, of Larry, who has time-dependent preferences for or against saving more. Which preferences should the policymaker heed? Larry might well have some higher-order evaluations – he might prefer to prefer saving more. But the intervention is supposed to be performed on a large group of individuals, so the policymaker is not likely to know such subtle attitudes of members in that group. Furthermore, they don't know whether Larry has devised his own ways to overcome preference instability, and whether their intervention might undermine such self-controlling strategies, actually making things worse for Larry. The point here, again, is that the policymaker doesn't know. And unlike a doctor who can check with the specific patient for allergies or other adverse indications before administering a treatment, the behavioural policymaker typically does not and cannot acquire this individual-specific information. Yet it would be needed to justify the intervention.

I part ways with the critics, however, in their claim that knowledge deficits are pervasive and generally leave paternalistic behavioural policies unjustified.[15] Instead, while knowledge deficits are an important problem to reckon with, there are many ways to design paternalistic interventions that avoid it. A first strategy is to design interventions in such a way that they are *innocuous* for those not targeted. To use another medical analogy, people have heterogeneous vitamin C needs – depending on their sex, small intestine capacity, diet and the air they breathe. But as excess amounts (up to a limit) are not harmful and are simply excreted, experts give general recommended doses and let people decide for themselves whether they want to ingest more – with the side effect being no more than the production of expensive pee. For vitamin C dose recommendations, specific knowledge seems unnecessary, and the general patterns easily discovered by science suffice (but don't do that with insulin!). Behavioural policy proponents have suggested that their interventions are similarly innocuous, for example in that they are 'asymmetric' – they only affect those who behave irrationally (Camerer et at. 2003, 102) – or in that they are 'cheap to avoid' – people who don't want to be affected can opt out without any substantial costs (Thaler and Sunstein 2009, 6). Unfortunately, these remain largely unsupported hypotheses, and supporting them might itself require particular and contextual knowledge.

This is where boosting comes in as an alternative strategy. To be effective, boosts *must* engage subjects' motivation. Consequently, only those who consider themselves in need of such competence boosts will choose to listen, learn and apply. For example, in Maggie's case, the doctor need not determine whether the discomfort from lactose intolerance outweighs the pleasures of eating cheese – that is for Maggie to decide; the doctor or policymaker only provides the competence to make this trade-off decision. This acts as a subject-centred, context-specific stopgap that any intervention must pass to be implemented. Although it does not guarantee innocuousness – subjects, after all, might be wrong about their own needs – it makes it much more likely that boosts are indeed interventions that are innocuous to those not targeted.

Boosts' engagement with and reliance on existing motivations also facilitate a third strategy for paternalistic interventions that avoid the knowledge-deficit problem: the

motivations that are engaged by the boost incorporate subjects' specific, situational, contextual knowledge. Because it is left to subjects to choose to be boosted or not, this specific knowledge – which critics argue policymakers typically do not have access to – is provided by the subjects themselves. Boosting policymakers therefore need not worry about whether specific individuals already have certain competences, whether boosting a specific competence would interfere with their existing competences, or whether a certain competence is useless to them. They simply leave those particular worries to the involved individuals, who can draw on their privileged specific knowledge. The boosting policymaker instead only has to ensure that for the population on average, such a boost would be useful, desirable, and not undermining; and such general pattern knowledge can typically be acquired through standard social science research. Boosts thus offer paternalistic interventions that avoid the knowledge-deficit criticism, simply because they do not require the kind of specific knowledge that the critics claim is typically not accessible to behavioural policymakers.

How boosting promotes equality

Behavioural interventions typically are aimed at a population of people, often a large one at that. Consider Zurich's green default, in which all households were offered a choice of energy providers, with a green but expensive provider set as the default (i.e. the option consumers would get if they did not make a choice by the deadline). About 2/3 of consumers ended up subscribing to the default green provider (Ghesla, Grieder and Schubert 2020). Or consider a San Diego experiment, in which 1,600 households were provided with personalized feedback about their water consumption, coupled with normative information about similar households in their neighbourhood (Schultz et al. 2016). Given that all such policies are claimed to be beneficial, the question arises whether this good is fairly distributed – both within the relevant population, but also with respect to other populations. Are all members of the population treated equally? Do they all receive the same benefit from the policy? These are some of the issues regarding distributive justice within the population. Why does this population (e.g. the citizens of Zurich) receive this benefit, but others (e.g. the residents of another Swiss canton) do not? This is an issue regarding distributive justice across populations. Both types of questions are driven by a moral concern for equality, although experts differ on whether this equality should concern material goods and services, opportunity, or respect.[16]

Let's look at some principles of distributive justice that express such equality concerns and see how behavioural policies fare with regard to them. The simplest principle might be *equality of treatment* – an intervention is distributively fair only if it treats all members of a population equally. *Within* a population, this is typically satisfied, as behavioural interventions are rarely tailored to individuals due to cost and knowledge limitations.[17] However, policymakers often determine the population for their interventions opportunistically – for example, the employees of a cooperating company or all patients of a particular health authority facing a particular choice –

leading to notable inequalities of treatment *across* populations. Both boosts and other behavioural policies seem to equally satisfy the within-population principle but are equally troubled by the across-population one, so I will not focus more on that.

But it isn't obvious that equality of treatment is even desirable for most policy contexts. People differ from each other, and thus under equal treatments often benefit in unequal and thus unfair ways. Imagine a new vaccine that provides immunity against a disease for all except Oliver. Oliver suffers from a rare genetic mutation, which makes the vaccine highly toxic for him, and taking it would cause him a lot of pain. An equal treatment mandate here seems morally wrong: it will produce a vastly worse outcome for Oliver, and it isn't Oliver's fault that this is so. People like Oliver are worse off due to *bad luck* – i.e. some random event that gave them opportunities that are worse than those of their peers and that they had no power to control. Therefore, it seems intuitively just that they should be provided with additional treatments (thus violating the principle of equal treatment) in order to ensure *equality of opportunity* for all in the population.

Heterogeneity in populations, including being born with a particular body and particular talents, or growing up in a poor or less educated family, often gives rise to inequality of opportunity. But at least for assessing specific behavioural policies, the relation is constrained in two ways. First, equality of opportunity only concerns properties that affect how beneficial the policy is for different individuals – inequality in other independent factors doesn't count. What matters is the unequal benefit that the vaccine has on Oliver and others; whether Oliver is otherwise disprivileged (e.g. that his family is poor) plays no role in the inequality this policy might produce. Second, equality of opportunity only concerns effects that those affected have no control over. Oliver cannot control his genetic makeup and thus cannot benefit from the vaccine through no failure of his own. Otto, in contrast, fails to adhere to the dietary restrictions necessary for the vaccine's effectiveness. He had the opportunity to benefit but blew it due to his own failures – hence here the equal opportunity criterion is not violated.

Despite these constraints, the equality of opportunity rule is more demanding of behavioural policies than the equality of treatment rule: it mandates that policymakers should compensate for differences in opportunities to benefit from interventions between subjects. Many behavioural interventions are, of course, motivated by such a principle – because some people suffer more than others from 'cognitive biases' or lack certain competences, many behavioural policies exactly aim to level the playing field, so that all individuals have more equal opportunities to live the life they want to lead. But such policies, even when aiming to level the playing field, might inadvertently create unequal opportunities of their own.

Let's apply these considerations to the Zurich green default and the San Diego norm messaging cases mentioned above. In Zurich, it turned out, poor consumers were more likely to end up with the green default than wealthy consumers, and poor consumers of the green provider also were more likely to regret this choice than their wealthier counterparts (Ghesla, Grieder and Schubert 2020). Being poor somehow made the behavioural intervention more effective, making poor consumers carry an overproportional burden of reducing pollution. In San Diego, consumers with a low

personal moral obligation to save water were more affected by the norm messaging intervention (reducing their water consumption more) than those who had a high personal moral obligation (Schultz et al. 2016).

While both of these cases exhibit differences in treatment effects, caused by heterogeneous personal characteristics, they are not equally morally problematic. Both concern non-paternalistic goals: they aim to reduce an externality (pollution and water waste, respectively) that concerns the population as a whole. But the unequal treatment effects put the burden disproportionately on some members of the population. This violates *fairness* if the personal characteristic is *not* normatively relevant for the behavioural outcome. Arguably, in the Zurich case, being poor does not give a reason to impose stricter pollution reduction on you than on the rich; poverty is normatively irrelevant for the distribution of the pollution reduction burden. Consequently, a policy like the green default is unfair if it generates such an unequal distribution. In the San Diego case, in contrast, if someone has a weaker sense of moral obligation to save water typically means that this person saves less water than others, *ceteris paribus*. Your having a weak sense of moral obligation thus is a reason to impose stricter water reduction on you by clarifying the social norm to you. In this sense, a weak sense of obligation is normatively relevant for the distribution of the water-saving burden. Consequently, a policy like the San Diego norm messaging is fair, other things being equal, even when it generates such an unequal distribution (Grüne-Yanoff 2024).

Boosts, because they seek to help people acquire new competences, must build on the assumption of some already existing competences. For example, improving risk literacy by showing the relation between frequency and relative probability formats requires basic arithmetic. Training caregivers to replace sympathy with compassion demands a basic grasp of one's emotions. Helping people to control their food intake by monitoring their satedness necessitates a basic sensitivity to their bodily signals. Teaching people anything will require their attention, which requires some ability to concentrate, a minimum of attentiveness and a learned ability to learn. In any population, the degree to which people will have these prerequisites will vary: some will have more and others less. A skilled boosting policymaker must consider the population average in the boost design; otherwise, the policy will not be effective. It may also, as a result, be less just (Kamphorst and Anderson 2024). If the population is quite heterogeneous, and targeting different subgroups with different boosts is not possible, then the boost will not help people 'in the tails': in the higher tail, they don't need those competences; on the lower end, they cannot acquire those competences because they lack prerequisites. In such difficult circumstances, the policymaker needs to decide which members of the population are most in need of the boost and design the intervention with their basic competences in mind. Mind you, not all boosts make high demands on the boosted. There are many boosts that do not exclude anyone, nor do they disadvantage those with lesser cognitive abilities.

The upshot is that both nudges and boosts can increase inequality when not properly employed by policymakers. Neither seems to have an advantage over the other in this regard. No general ethical conclusion can be drawn. Instead, the assessment depends on the details of the target population and the context in which it is employed. Users

must carefully consider the possibility of such negative side effects, seeking to avoid them altogether or at least balancing them against other advantages of their preferred intervention option.

Conclusion

When it comes to interventions designed to change behaviour – even very problematic behaviour – sometimes the cure is worse than the disease. In this chapter, I've used four ethical principles – freedom, autonomy, paternalistic modesty and equality – to investigate the potential side effects of policy interventions, side effects that are also relevant to the design and implementation of behaviour change technologies. Considering these principles is crucial, especially when an intervention is effective for some but potentially harmful to others. The four principles elucidate the ways that policies could make things worse, and for whom. They serve as both a general warning and a toolbox that policymakers can use systematically to assess proposed interventions.

My analysis has shown that boosts generally fare better than nudges and traditional legislative prohibitions across these ethical dimensions:

- Freedom: Boosts never violate freedom of choice, while nudges sometimes do. By enhancing competences rather than restricting choices, boosts often preserve or even increase freedom.
- Autonomy: Boosts never reduce autonomy, while nudges often do. In fact, boosts can often increase autonomy by encouraging reflection and deliberation. For example, a boost teaching critical thinking skills for evaluating online information not only preserves autonomy but enhances it by empowering users to make more informed decisions.
- Paternalistic modesty: Boosts are always modestly paternalistic, while nudges often are not. The self-selective nature of boosts addresses the knowledge-deficit problem that plagues other forms of paternalism. Users engage with boosts based on their own assessment of their needs, reducing the risk of policymakers imposing their values on others.
- Equality: Both boosts and nudges can – but don't always do – preserve equality of opportunity. While neither has a clear advantage, boosts that focus on widely accessible competences may have an edge in promoting equality.

In conclusion, while boosts are not without their ethical challenges, they generally have fewer ethically problematic side effects than either nudges or legislative regulation. Their focus on enhancing competences rather than manipulating choices aligns well with the values of freedom, autonomy, modest paternalism and equality. As we navigate the complex landscape of behaviour change in the digital age, boosts offer a promising approach that respects individual agency while still working towards

positive behavioural outcomes. However, continuous ethical scrutiny and refinement of all behaviour change approaches, including boosts, remain essential as technology and our understanding of human behaviour continue to evolve. By focusing on empowerment rather than control, on education rather than manipulation, we can create behaviour change technologies that not only effectively address societal challenges but also uphold our highest ethical standards.

Notes

1 While some authors have defended behavioural interventions by challenging the relevance or validity of such principles – Skinner's *Beyond Freedom and Dignity* (1971) and Conly's *Against Autonomy* (2013) come to mind – I do not. Instead, I argue that boosts satisfy these principles, and furthermore, that they satisfy them better than other behavioural policies.

2 This concept of negative liberty goes back to Hobbes and Hume but was explicitly delineated by Berlin (1969). Subsequently, many different versions have been proposed, differing on what counts as an option (all physically possible alternatives or only eligible ones) and what the attribution to others involves (e.g. causing, intending or being morally responsible for the reduction) (Kristjánsson 1996).

3 For an extended discussion of motivations as 'inner' obstacles, see Kramer (2003) and Garnett (2007). Other authors have taken the difficulties with 'inner' motivations and liberty as reason to replace negative liberty with the concept of *positive liberty* that centres on *actually doing* certain things in certain ways – for example, in realizing one's true self or in acting on the basis of rational and well-informed decisions (Christman and Anderson 2005; Velleman 2000). I discuss such notions in the next section as 'autonomy' but maintain that boosts should be evaluated both with respect to negative liberty *and* positive liberty, which I refer to by the more traditional term autonomy. I explain the reasons for this in the next section.

4 Unless of course one considers knowledge and competences as something one would wish to be free from. But freedom *generally* is not a value; only particular liberties are worth protecting (Dworkin 1977, 2011 and the later Rawls 1991). For example, freedom to be ignorant, freedom from knowing about alternative ways of doing things, freedom from realizing that one does not do things in one's own interest or optimally according to one's own criteria do not fall under the same special protection that freedom of choice, freedom of expression, freedom of association, etc. Boosts interfere with people's 'freedoms' of the former, but not the latter, kind: they help them see their environment and their own motivations, feelings and beliefs in new, more nuanced and often more accurate perspectives, thus facilitating more competent reasoning, judgements and practical conclusions, that – *if the individual so desires* – can lead to different choices. But boosts do not limit the freedom to choose differently, again, if the individual so desires. In particular, note that I am not denying the value of freedom to commit errors or to refuse learning (which are all part of the freedom of choice)!

5 There is of course the issue of uncertainty: what seemed to be true or a skill, based on given evidence turns out not to be so. But that is for the boosted subject to decide, not the policymaker.

6 For an extended argument on why nudges sometimes violate negative as well as republican liberty, see Grüne-Yanoff (2012).
7 The *coherentist* account has been defended by e.g. Bratman (2004). . The *reasons-responsive* account has been defended by e.g. Wolf (1990). The *responsiveness-to-reasoning* account has been defended by e.g. Anderson and Honneth (2005). The *incompatibilist* account has been defended by e.g. Kane (1996).
8 Whether real or not, the prospect that freedom and autonomy might be played against each other in this way has been the bane of various liberal authors (e.g. Hayek, Berlin). Aiming to increase autonomy, they argued, might lead to authoritarian forms of government. They concluded that policy, therefore, should be evaluated with reference to liberty, *not* autonomy. I think it should be assessed with respect to both.
9 Autonomy of acts consequently is determined by the autonomy of desires driving these acts, and the autonomy of a person consists in most of their acts being considered autonomous.
10 For a systematic comparison of boosts and nudges in terms of their impact on the personal autonomy of their targets, see Grüne-Yanoff (2023).
11 Sunstein and Thaler (2003) for example that individuals affected by a changed default lack clearly defined preferences, and thus policymakers need not worry about satisfying their genuine preferences. Others have argued that some people will be less able to judge for themselves what their own good is and hence be more susceptible to justified paternalistic intervention (Conly 2013).
12 That incommensurability of values can be upheld by reasoning, reflected, rational agents has been argued among others by Raz (1986).
13 There are more demanding definitions, which insist that something is a paternalistic intervention on agent *X* only if (i) it is an interference with the liberty or autonomy of *X*, (ii) lacks *X*'s consent, and (iii) the intervener intends to promote the interests, values, or good of *X* (Dworkin 2013). Such a definition would automatically exclude boosts from being paternalist (given our arguments from the last two sections), and I do not want to use this easy excuse here. I therefore use paternalist in the wider sense of satisfying condition (iii) only.
14 A classical critique of paternalism along these lines is offered by Mill (1859).
15 For the full argument, see Grüne-Yanoff (2021).
16 For equality of opportunity, see Rawls (1991), Dworkin (2011), for equality of respect, see Anderson (1999).
17 Note, however, the call for behavioural intervention personalization, e.g., by Sunstein (2012).

References

Abrahamse, Wokje and Nadja Contzen. 2025. 'Psychology of Behaviour Change.' In Joel Anderson, Lily Frank and Andreas Spahn (eds), *Ethics of Behaviour Change Technologies*. London: Bloomsbury Academic.

Anderson, E. 1999. 'What Is the Point of Equality?' *Ethics*, 109: 287–337.

Anderson, J. and A. Honneth. 2005. 'Autonomy, Vulnerability, Recognition, and Justice'. In J. Christman and J. Anderson (eds), *Autonomy and the Challenges to Liberalism: New Essays*, 127–49. New York: Cambridge University Press.

Berlin, I. 1969. 'Two Concepts of Liberty'. In I. Berlin (ed.), *Four Essays on Liberty*. London: Oxford University Press. New ed. in Berlin 2002.
Bratman, M. 2004. 'Three Theories of Self-Governance'. *Philosophical Topics*, 32: 21–46.
Camerer, C., S. Issacharoff, G. Loewenstein, T. O'Donoghue and M. Rabin. 2003. 'Regulation for Conservatives: Behavioral Economics and the Case for "Asymmetric Paternalism"'. *University of Pennsylvania Law Review*, 151 (3): 101–44.
Christman, J. and J. Anderson, eds. 2005. *Autonomy and the Challenges to Liberalism: New Essays*. New York: Cambridge University Press.
Conly, Sarah. 2013. *Against Autonomy: Justifying Coercive Paternalism*. Cambridge: Cambridge University Press.
Dworkin, Gerald. 2013. 'Defining Paternalism'. In Christian Coons and Michael Weber (eds), *Paternalism: Theory and Practice*, 25–39. Cambridge: Cambridge University Press.
Dworkin, R. 1977. *Taking Rights Seriously*. London: Duckworth.
Dworkin, R. 2011. *Justice for Hedgehogs*. Cambridge, MA: Harvard University Press.
Garnett, M. 2007. 'Ignorance, Incompetence, and the Concept of Liberty'. *Journal of Political Philosophy*, 15: 428–46.
Ghesla, C., M. Grieder and R. Schubert. 2020. 'Nudging the Poor and the Rich: A Field Study on the Distributional Effects of Green Electricity Defaults'. *Energy Economics*, 86: 104616.
Gigerenzer, G., W. Gaissmaier, E. Kurz-Milcke, L. M. Schwartz and S. Woloshin. 2007. 'Helping Doctors and Patients to Make Sense of Health Statistics'. *Psychological Science in the Public Interest*, 8: 53–96.
Gollwitzer, P. M. 1999. 'Implementation intentions: strong effects of simple plans'. *American psychologist*, 54 (7): 493.
Grüne-Yanoff, T. 2012. 'Old Wine in New Casks: Libertarian Paternalism Still Violates Liberal Principles'. *Social Choice and Welfare*, 38 (4): 635–45.
Grüne-Yanoff, T. 2021. 'Boosts: A Remedy for Rizzo and Whitman's Panglossian Fatalism'. *Review of Behavioral Economics*, 8 (3–4): 285–303.
Grüne-Yanoff, T. 2023. 'Boosting Agency: Autonomy-Promoting Behavioral Public Policy'. *Social Philosophy and Policy*. Accepted for publication 19 July 2023.
Grüne-Yanoff, T. 2024. *Unfair Behavioral Public Policies*. Unpublished manuscript.
Grüne-Yanoff, T. and R. Hertwig. 2016. 'Nudge Versus Boost: How Coherent Are Policy and Theory?" *Minds and Machines*, 26: 149–83.
Hertwig, R. and T. Grüne-Yanoff. 2017. 'Nudging and Boosting: Steering or Empowering Good Decisions'. *Perspectives on Psychological Science*, 12 (6): 973–86.
Kamphorst, B. A. and J. H. Anderson. 2024. 'E-Coaching Systems and Social Justice: Ethical Concerns About Inequality, Coercion, and Stigmatization'. *AI and Ethics*. https://doi.org/10.1007/s43681-024-00424-7.
Kane, R. 1996. *The Significance of Free Will*. New York: Oxford University Press.
Kramer, M. H. 2003. *The Quality of Freedom*. Oxford: Oxford University Press.
Kristjánsson, K. 1996. *Social Freedom: The Responsibility View*. Cambridge: Cambridge University Press.
Mill, John Stuart. (1956) 1859. *On Liberty*. Indianapolis: Bobbs-Merrill.
Oinas-Kukkonen, Harri and Eunice E.Y.F. Agyei. 'Technologies of Behaviour Change.' 2025. In Joel Anderson, Lily Frank and Andreas Spahn (eds), *Ethics of Behaviour Change Technologies*. London: Bloomsbury Academic.
Pettit, Philip. 1997. *Republicanism: A Theory of Freedom and Government*. Oxford: Clarendon Press.

Rawls, J. 1991. *Political Liberalism*. New York: Columbia University Press.
Raz, J. 1986. *The Morality of Freedom*. Oxford: Clarendon Press.
Rizzo, M. J. and G. Whitman. 2020. *Escaping Paternalism: Rationality, Behavioral Economics, and Public Policy*. Cambridge: Cambridge University Press.
Schultz, P. W., A. Messina, G. Tronu, E. F. Limas, R. Gupta and M. Estrada. 2016. 'Personalized Normative Feedback and the Moderating Role of Personal Norms: A Field Experiment to Reduce Residential Water Consumption'. *Environment and Behavior*, 48 (5): 686–710.
Skinner, B. F. 1971. *Beyond Dignity and Freedom*. New York: Vintage Books.
Skinner, Quentin. 2002. 'A Third Concept of Liberty'. *Proceedings of the British Academy*, 117: 237–68.
Spiegelhalter, D., M. Person and I. Short. 2011. 'Visualizing Uncertainty About the Future'. *Science*, 333: 1393–400.
Sunstein, C. R. and R. H. Thaler. 2003. 'Libertarian Paternalism is Not an Oxymoron'. *The University of Chicago Law Review*, 70 (4): 1159–202.
Sunstein, C. (2012). *Impersonal Default Rules vs. Activate Choices vs. Personalized Default Rules: A Triptych. Harvard Public Law Working Paper*. posted 8 Nov 2012. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2171343.
Thaler, R. H. and S. Benartzi. 2004. 'Save More Tomorrow™: Using Behavioral Economics to Increase Employee Saving'. *Journal of Political Economy*, 112 (S1): S164–87.
Thaler, R. H. and C. R. Sunstein. 2009. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. London: Penguin.
Velleman, J. D. 2000. *The Possibility of Practical Reason*. Oxford: Clarendon Press.
Weinert, F. E. 2001. 'Concept of Competence: A Conceptual Clarification'. In D. S. Rychen and L. H. Salganik (eds), *Defining and Selecting Key Competencies*, 45–65. Ashland, OH: Hogrefe and Huber Publishers.
Wolf, S. 1990. *Freedom Within Reason*. New York: Oxford University Press.