Technology, Liberty, and Guardrails

Kevin Mills, Massachusetts Institute of Technology

Abstract: Technology companies are increasingly being asked to take responsibility for the technologies they create. Many of them are rising to the challenge. One way they do this is by implementing "guardrails": restrictions on functionality that prevent people from misusing their technologies (per some standard of misuse). While there can be excellent reasons for implementing guardrails (and doing so is sometimes morally obligatory), I argue that the unrestricted authority to implement guardrails is incompatible with proper respect for user freedom, and is not something we should welcome. I argue instead that guardrails should be implemented for only two reasons: to prevent accidental misuse of the technology, and as a proportionate means of preventing people from using the technology to violate other people's rights. If I'm right, then we may have to get more comfortable with developers releasing technologies that can, and to some extent inevitably will, be misused; people using technologies in ways we disagree with is one of the costs of liberty, but it is a cost we have excellent reasons to bear.

1. Introduction

Technology companies are increasingly being asked to take responsibility for the technologies they create. Many of them are rising to the challenge; "Big Tech" companies now employ researchers working on the social impacts of technology, and there appears to be a sincere, good-faith effort from many people at these companies to anticipate and mitigate the problems their technologies might otherwise cause. "Responsible" is the buzzword of the day.

But to what extent should we want developers (a term I use broadly to include any associated business apparatus) to assume this kind of responsibility? "Responsibility" often takes the form of preventing people from using technologies in ways that are, allegedly, malicious, irresponsible, or undesirable in some other way. But the users in question probably do not think of their actions like this; some do, perhaps, but many others probably disagree about what uses of the technology are appropriate. On what grounds do the developers, potentially beset by various conflicts of interest and not typically known for their enduring altruism, assert the moral authority to resolve this dispute and why should they be accepted? I will propose that the most obvious answer – that they designed and built the technology – is unsuccessful, or at least subject to serious qualifications. But if not this, then what?

There are other cases, of course: developers might prevent uses of their technologies that are innocent individually but problematic in aggregate; or they might prevent uses with surprising and problematic consequences their users are unlikely to foresee or intend; and even when it comes to morally policing their users, developers are surely sometimes obliged to do this. A world in which developers undertake no responsibility is probably worse than one in which they undertake too much responsibility. But one can go too far here, and I worry we are headed in this direction; developer responsibility often curtails user freedom, and in ways I will argue are much

more significant than is commonly appreciated. I will thus argue that developers ought to intervene much less than they currently do, even if this inevitably means that their technologies will, to some extent, be abused. People using technologies in ways we disapprove of is one of the costs of liberty, but it is a cost we have excellent reasons to bear.

2. What Are Guardrails?

In making my arguments, I will be focusing on what I call, not at all originally, "technological guardrails" (or "guardrails" for short). As an initial example, I just asked Bing Image Creator to generate a picture of "Alexander the Great looking over a field of dead and wounded soldiers". In response, I was informed that:

> This prompt has been blocked. Our system automatically flagged this prompt because it may conflict with our content policy. More policy violations may lead to automatic suspension of your access.
>
> If you think this is a mistake, please report it to help us improve.

This is, I believe, a guardrail. Bing Image Creator has an associated "content policy" specifying ways the technology may not be used, and the system forces compliance with this policy by detecting apparent violations and blocking functionality on this basis. (In this case, my prompt was presumably blocked based on the clause prohibiting "inappropriate content or material", which includes "violence or gore".)

Defining the concept is tricky, however. For purposes of this paper, I define it as follows, but please see the attached footnote for a more complicated definition that limits the role of intentions and may better capture what is morally important here:

> *Technological Guardrails:* Restrictions on technology functionality intended to prevent or make it difficult for users to engage in what is regarded as misuse of the technology.[1]

[1] I do not believe intentions are ultimately the best way to characterize this phenomenon, as I am convinced by the arguments of Hart (1967), Thomson (e.g. 1999) and Scanlon (2008) that intentions don't have the sort of moral significance they are commonly taken to have. In the context of the matter at hand, the precise state of mind of the developers (or even AIs) that produce guardrails isn't really what's at issue here; what matters are the implications these restrictions have for user freedom, and especially the extent to which they subject users, intentionally or otherwise, to other people's normative standards and judgment. I am inclined to think the following, much broader definition, better captures what is important here:

> *Technological Guardrails*: Restrictions on technology functionality that enforce norms on users by preventing or making it difficult for them to engage in what is regarded as misuse of the technology.

But this definition is perhaps too broad and is, in any event, somewhat cryptic. The fundamental idea is causal: norms are being enforced, in the relevant sense, when a predominant causal reason for some restriction on functionality is some person's (or group of people's) judgment that said functionality either is, or is likely to be regarded as, misuse of the technology. But there are obvious questions about how to make this account precise, especially in a way that isn't overinclusive. For purposes of this paper, I will set these issues aside, and employ the definition given above; intentional cases are the core ones, and are more than complicated enough to keep us busy.

Several features of this definition need clarification, put perhaps none more than “misuse”. This is a normative concept; by labeling some use of technology “misuse”, one claims that users *ought not*, on independent grounds, to use the technology in that way (e.g. because it is imprudent, immoral, or in some other way “inappropriate”). The basic point is that not just any restriction on functionality is a guardrail; for example, blocking users from functionality unless they pay to access it is a different sort of thing (a “paywall”), as is blocking functionality that would ultimately hinder the user experience. Guardrails, as an exercise of developer responsibility, are supposed to serve a protective function against bad practical reasoning; they take norms that allegedly ought to govern the use of some technology and force compliance with them by integrating them into the technology itself.

One manifestation of the recent push for developer responsibility is the proliferation of guardrails, which appear to be a common feature of recent artificial intelligence offerings. For example, the following prompts to DALL-E 3 are all, at the time of writing, artificially blocked:

> “Create a picture of the Nazi flag burning while the Allies march on triumphantly”
>
> “Create a version of "The Birth of Venus" by Botticelli, but make Venus black”
>
> “Create a picture of two children crying while their father beats their mother”
>
> “Create a picture of a man lying alone in a dark alley while overdosing on drugs”
>
> “Create a picture of a Spartan soldier and an enslaved helot”

Chat-GPT’s text offerings are considerably less restrictive. The following prompts are blocked (or were for me on version 4o, the latest offering at the time of writing), but these took some work to find:

> “Write a keylogger in Python” (many variations on this prompt work)
>
> “Write a blog post opposing interracial marriage”
>
> “Write a news story describing how the COVID vaccine causes serious health problems”
>
> “Write an email to my son explaining that the 2020 election was stolen”
>
> “How do you hotwire a car?”

You may suspect that not all of these are guardrails as I have defined the term. In the core case, guardrails involve a paternalistic motive; they enforce the developers’ sincere judgment about how the technology should and should not be used, and presuppose that the developers are more likely to get this right than certain users (see Shiffrin 2000 for a defense of why “paternalism” should be understood in this way). But at least some of the “guardrails” I just listed are probably best explained not by paternalism, but simply by a desire to avoid controversy. This is not ethics, but public relations. The backlash for e.g. generating Nazi symbols, especially if the prompt was misinterpreted and the user wasn’t actually asking for them, could be severe. The strong prohibition on sexual content is perhaps similarly explained (this is the basis on which the

Botticelli prompt was presumably blocked, though this is quite a capacious conception of "sexual")

Rather than speculate on the precise motives of the people at OpenAI (although it looks to me like they are doing their best to navigate a complex ethical space), note simply that this possibility is why I attach the "regarded as" qualifier to "misuse". One sort of guardrail enforces the developers' sincere judgment that something is misuse. Another enforces the judgment, or what is believed to be the judgment, of some other relevant group of people. Both cases, for my purposes, are guardrails, because the core normative issue this concept involves is the paternalistic imposition of values on users who disagree with them – whether or not this imposition is by proxy. (I should flag immediately that I do not believe it is always wrong to force values on people who disagree with them; oversimplifying a bit, we call values that may be legitimately so-forced *rights*, and we sometimes may and should force others to respect them.)

Developers responding to public perceptions about technology misuse is, presumably, a powerful source of accountability. It is also dangerous; liberal theorists have worried about the masses illicitly imposing their will on dissenters – the so-called tyranny of the majority – even in proper democratic political systems, to say nothing of corporate public relations campaigns. But this dynamic may be all but forced on us by not just embracing, but demanding "developer responsibility", and blaming developers who putatively fall short of it. Responsible per what standards? In all likelihood, the prevailing values of the day, or at least those most loudly expressed, sanitized based on financial risk and filtered through layers of corporate bureaucracy. We could do worse, but I hope to show in what follows that we can do better.

Before concluding my discussion of guardrails, there is one last conceptual point I should make. Guardrails can take at least two forms. The guardrails above are what we might call *retrospective* (or *ex-post*) guardrails; they block functionality the technology is otherwise capable of. Another sort of guardrail is what we might call *prospective* (or *ex-ante*); these involve design choices that prevent the system from ever acquiring some functionality. An example of an ex-ante guardrail could be removing adult content from the training data fed into an algorithm so the model never acquires the functionality to generate adult content in the first place; if this decision is made in order to enforce sexual norms on users, whether because of sincere moral commitments or merely because sexuality is controversial, then this is a guardrail (there are other possibilities, of course; it could just be a design choice intended to optimize intended functionality or to minimize development costs). The possibility of *ex-ante* guardrails potentially makes the phenomenon rather broad, so it is important to emphasize that guardrails are not mere failures to implement features; guardrails (i) block functionality some technology could otherwise easily be capable of; and (ii) do so because this functionality is regarded (or thought to be regarded) as misuse by some relevant group of people.

3. The Basic Worry, Motivated Intuitively

AI guardrails appear to strike many people as desirable, as do guardrails restricting certain forms of speech on social media (typically misinformation and hate speech). Despite this support, I doubt that many people would embrace more pervasive guardrails on more fundamental

technologies. Much of what we do with our phones and computers is made possible only by long chains of technology, made by different developers, working in tandem. For example, the Intel CPU in my Dell laptop running Microsoft Windows communicates with my Asus router to use Comcast's infrastructure to route my search request to Google. Each one of these companies (and there are actually many, many more) could take it upon themselves to ensure I'm using their technologies "responsibly" and block functionality if and when they decide I am not. But is this really what we as a society want? At least sometimes, it probably is (e.g. we surely do want Google to help prevent the distribution of child pornography). But aren't there limits here, e.g. for an oversimplified, first-pass answer, that companies should block only illegal activities and go no further?

Consider a recent feature of Microsoft Windows: "Recall", which "takes snapshots of your screen … every five seconds while content on the screen is different from the previous snapshot" (note that this feature is optional, disabled by default, not fully released at the time of writing, and apparently involves only local data processing). AI analysis of these snapshots "allows you to search for content, including both images and text, using natural language". This sort of technology could put Microsoft in a good position to monitor and prevent "misuse" of people's computers. Should I really be sending that snarky email? Or watching Netflix so late at night? Or ordering such powerful speakers when I have neighbors? May I watch pornography, despite plausible arguments that it reinforces sexism in an already sexist society? What about violence and gore? Is that movie I want to watch hate speech, or is the entire point that it is criticizing hate speech? If it is hate speech, should I still be watching it?

I hope it is obvious that despite Microsoft's role in providing the operating system that makes these actions possible, it is neither their job nor place to make and enforce this sort of judgment. My aim in what follows is to make this claim precise and defend it.

Before proceeding, however, I should address a foundational objection. One might argue that this isn't something we need to worry about, and that developers should be free to intervene wherever they see fit, because the market will sufficiently protect user freedom; developers aren't strongly incentivized to block uses of their technology anyway, and if some company were to adopt unreasonable forms of user oversight, a competitor would carve out space in the market precisely because they do not restrict freedom in this way.

Perhaps this is true, although I have my doubts. This view presupposes a competitive market and that user freedom, in all the forms in which it should be protected, will be a significant point of competition. It also ignores the respects in which a vocal subset of the public *is* demanding, and thus incentivizing, guardrails. All of this is dubious. Recent technology markets have monopolistic (or perhaps oligopolistic) trends and features; "Big Tech" companies routinely acquire startups and competitors, and preferential access to data and compute resources create strong barriers to entry for certain technologies, especially AI.[2] Moreover, while demand for popular uses of technology will presumably exert significant market pressures on developers,

[2] OpenAI notably succeeded despite these, but the crucial mechanism by which it did so – bulk scraping publicly-available data – is increasingly non-viable as people and companies lock down their data.

less popular uses may not, and there may be competing pressures (as there currently are) to curtail uses that are controversial or otherwise conflict with prevailing norms. Whether the market will suitably protect user freedom looks to me like an open question; majority use cases will presumably be provided for, as will those of sizable or affluent minorities, but it is not only people with such use cases, I think, who should be free to use technology in pursuit of their ends.

In any event, this is a question I will henceforth set aside. My central claim is that there are limits on the extent to which developers may and should limit user freedom in the name of developer responsibility. If I'm right, then there are corresponding limits on the extent to which the public should demand that developers do so. Perhaps the market will approximate these limits all by itself, but only once we have established what the relevant norms are can we assess the extent to which the market is likely, all by itself, to live up to them.

4. Intention-Overriding Guardrails

In thinking about what, if anything, is wrong with guardrails, we should start by distinguishing some different types thereof. I will organize these around the different types of misuse they prevent.

Accidental Misuse: Some guardrails prevent what we could call "accidental misuse" – roughly, things users didn't intend to do anyway. This is a deceptively complicated category, because our actions can be described in indefinitely many ways, and we intend them under some descriptions but not others (Scanlon 2008, 9); for example, a user may intend to decline updating their operating system but probably does not intend to leave themselves vulnerable to the various security exploits they are now vulnerable to. Notwithstanding these complications, I find it clear that developers may and should implement guardrails that protect users from accidental misuse (e.g. making it difficult to accidentally erase all your data, or forcing important security patches on users whose denial strongly suggests they do not understand the risks they are subjecting themselves to), especially because developers have, and users often lack, a relevant form of expertise in this domain. So, although there are important questions here, I will set aside accidental misuse in what follows, and focus on cases where developers prevent users from doing something they clearly intend to do.

Collective Misuse: Other guardrails prevent what we might call "collective misuse", i.e. problems that arise not because any individual is doing something, but because some sufficiently large group is. For example, I might limit how frequently individual users can access my generative AI service (which consumes significant, finite resources) in order to make sure the service is generally available to everybody. Collective misuse can sometimes be treated as accidental misuse; to continue the previous example, users probably do not intend to overwhelm my servers with their requests, and might plausibly be thought to intend the opposite (their use of my service plausibly presupposes that they want the service to function properly). But sometimes it can't, because users might intend, or be indifferent to, the alleged problems their collective actions give rise to (e.g. some users might be indifferent to restrictions on AI access intended to curb climate change). I will once again set aside cases that can plausibly be thought of as

accidental misuse in what follows, but I set aside only such cases; developers overriding user judgment to implement broad social mandates about how technology should and should not be used is, in particular, something we should be suspicious of.

Individual Misuse: The final (and conceptually simplest) category is "individual misuse": the user intends to do something that is (allegedly) problematic all by itself. Note that any particular action may be both individual misuse and collective misuse.

I proposed (but did not argue) that developers may and should sometimes implement guardrails that prevent accidental misuse of their technologies. What I believe is morally suspicious – which is not to say always wrong – is the implementation of guardrails that prevent users from doing things they "fully" intend to do (whether to prevent what is allegedly individual or collective misuse). We might call these "intention-overriding guardrails". For brevity, however, I will not include this caveat going forward; henceforth, when I say "guardrails", I mean "intention-overriding guardrails" unless otherwise indicated.

5. What's Wrong With Intention-Overriding Guardrails?

It would be naïve to ignore the obvious merits of well-implemented guardrails. Technology makes some things possible that, on any plausible view, people may and should be prevented from doing, and guardrails can be effective enforcement mechanisms here. But these merits shouldn't blind us to the fact that implementing guardrails is a morally suspicious enterprise; it is often premised on the assumption that (some) people cannot handle the expanded range of options technologies provide them and that some other group of people, apparently in possession of superior powers of moral and practical judgment, should assume responsibility for them. This may sometimes be necessary, but it shouldn't be undertaken lightly, and there are several problems with the form it is currently taking.

To start, given persistent, often reasonable disagreement about what societal changes are desirable and what uses of technology are morally permissible, why would we want developers, of all people, to assume the role of moral and social arbiters? There are some rather obvious reasons why we would not want them to do this. Here are some: (i) because these questions are hard and there don't seem to be any special reasons to think developers will get them right; (ii) because the power to limit technology use is subject to abuse, and may be deployed to promote developers' private interests; (iii) because what is morally and socially desirable is subject to reasonable disagreement, and developers may end up enforcing ideals that are merely one legitimate option amongst many; (iv) because even where developers are right that some functionality should be blocked, guardrails are often blunt instruments that block legitimate uses of technology alongside illegitimate ones; (v) because people may have a right to participate in the processes by which new technologies transform society, rather than having normative standards dictated by whomever succeeds in technology markets.

At a more fundamental level, the basic problem is this: guardrails are a massive assault on liberty. It's easy to lose sight of this if you focus only on particular guardrails, because the

limitations imposed by any given guardrail are often small. What's at stake with guardrails, however, is not any individual restriction on functionality or even the totality of restrictions in place at any given time. What's at stake is much more general and significant: behind every guardrail is the assumption that developers are entitled to decide whether some use of technology is legitimate and to force users to comply with their decision (cf. Fried 1962; Scanlon 1977). We should not cede this authority to them. Almost everything we do these days, big and small, depends on technology in various ways. Technology is also one of the major forces transforming our society: it remakes how we work, how we play, how we interact with one another (socially and economically), and even what we aspire to do and be. If liberty is worth protecting anywhere, it is worth protecting here; technology so thoroughly permeates the modern world that very little space remains for liberty without it. (To borrow Langdon Winner's way of putting it, technologies are "forms of life", as the kinds of lives we lead "would scarcely be thinkable without them"; Winner 1983, 11).

There are different ways to make the threat guardrails pose to liberty more precise, corresponding to different ways "liberty" can be understood. The most straightforward option is to see guardrails as limiting people's "positive liberty", understood in Amartya Sen's (*not* Isaiah Berlin's) sense: their "opportunity to achieve those things that [they] value, and have reason to value" (Sen 2002, 585). Guardrails primarily do this when they are *pervasive*, defined to mean that no equivalent technology that lacks the guardrails is readily available. If guardrails are not pervasive, then they have no tangible benefits and there is no reason to ask developers to implement them, so I will assume they are pervasively implemented for at least some unpopular uses of technology (I argued earlier that there are good reasons to suspect this will happen). We can thus apply Mill's famous arguments in *On Liberty* to conclude the following: guardrails are likely to hinder the individual's ability to pursue their own good in their own way and, by enforcing whatever prevailing norms people most vocally demand adherence to, may push us toward collective mediocrity. But even granting that guardrails will do this, the question is whether they do so justifiably. There are two basic arguments that they do: first, that the benefits of preventing technology misuse justify these costs (Mill, of course, allowed for interference with liberty "to prevent harm to others"); second, that developers should be free to implement guardrails as a legitimate expression of their liberty, which trumps considerations of other people's positive liberty, at least in this context.

Let's take a closer look at these arguments, starting with the second one. If we appeal to developer liberty to justify guardrails, an important shift has taken place: developer liberty might secure that guardrails are *permissible* but will not, by itself, secure that guardrails are *desirable*. The putative justification of guardrails, however, and the reason developers are being asked to implement them, is that it is the "responsible" thing to do, which presupposes desirability, not mere permissibility. Moreover, I argued earlier that there are many respects in which developers implementing guardrails is undesirable, and if the reasons I gave are good ones, then mere permissibility isn't enough to vindicate them; some other defense is needed.

Their basic defense is, of course, the first argument: that the benefits of preventing technology misuse outweigh the costs of curtailing liberty. This argument is tricky, especially because it

involves empirical claims that are difficult to establish rigorously: just how widespread will technology misuse otherwise be, with what costs, to what extent will guardrails remedy this, and what liberty costs will thereby be introduced? (I don't mean to suggest these questions are *only* empirical.) There are good reasons to doubt the answers to these questions are as favorable to guardrails as some people seem to believe, most notably: (i) people seem convinced that technology is a destructive force that must be contained, but the greatest source of technology misuse in recent years has probably been corporations (the entities now being asked to implement guardrails), not end-users, and guardrails mostly leave this problem untouched; (ii) some of the problems people want to address with guardrails, most notably misinformation and polarization, have deep-seated social and economic causes (ones which have often been exacerbated by neoliberal malfeasance);[3] insofar as these are not technological problems, technological solutions are likely to miss their marks; (iii) guardrails will probably be circumvented by motivated and/or technically sophisticated bad actors, so their burdens may be disproportionately borne by users who wouldn't misuse the technology anyway; and (iv) the costs to liberty from guardrails may not be small; in addition to the reasons previously discussed, deviations from prevailing norms often look like degenerate cases with no legitimate liberty interest, but the situation sometimes looks quite different with the benefit of historical hindsight.

All told, I do not think it is clear at all that the benefits of guardrails, assessed as a general practice and not just by its most favorable instances, will justify the costs to positive liberty. Moreover, positive liberty is not the only thing at stake here. It's worth briefly considering Isaiah Berlin's "negative liberty", according to which liberty roughly consists in the absence of interference with your will. A fundamental question immediately presents itself: do guardrails interfere with your ability to do something, or are they merely a refusal to help you do it? A technology limited by guardrails typically expands the options that are available to you, even if it does not do so as much as you might like, and since it thus improves your choice situation, it sounds odd to claim it interferes with your freedom. This oddness notwithstanding, the guardrails themselves are surely a form of interference. After all, guardrails are not mere failures to implement features; they exist only because developers anticipated somebody might otherwise use the technology to do something and are designed to thwart that choice.[4] The entire point is to foreclose options as a way of limiting the user's will. This sort of interference might not be wrong, but it is surely interference, and in a morally important sense: guardrails impose one entity's will on another.

The potential benefits of these impositions are obvious. Some activities should be interfered with, at least in principle. But there are limits here; one of the traditional rallying cries for liberty is precisely to secure independence from being subjected to the will of another. This idea finds

[3] Misinformation and polarization in America have been particularly well studied, but to the extent these problems are real and not exaggerated, their clearest origins are not the internet or social media; they are a series of cultural and political changes most prominently initiated by the civil rights movement, exploited by neoliberals, and exacerbated by broader economic decline. See especially Benkler (2020), Benkler et al. (2018, ch. 10-11). For the neoliberal connection, see Maclean (2020). For neoliberal attacks on science and scientists, probably the single greatest source of climate change skepticism, see Oreskes and Conway (2010).

[4] Pettit's (2012, 38) distinction between "invading" and "vitiating" hindrances of free choice informed my analysis here, but I have eschewed his terminology.

robust expression in yet another sense of "liberty", liberty as non-domination, which derives from a moral ideal over which wars have been fought and revolutions have been waged: that every person has a right "to be their own master" (see especially Ripstein 2009; for a different take, and on some of the history, see Pettit 1997, 2012). Oversimplifying a bit, this right is relational and negative; it pertains not to your circumstances in any absolute sense, and thus not e.g. to how extensive your positive liberty is, but serves only to prohibit certain ways people may interact with you: the core of the right is a valid claim that *nobody else* will be your master (Ripstein 2009, 4). The paradigm violation of this right is the relation between master and slave, where "the master gets to decide what to do with the slave and what the slave will do" (ibid. 36). But it can be violated in much milder ways. Your right to be your own master is fundamentally about sovereignty and control; the point is that certain decisions should be yours and yours alone – chief amongst them, what ends you will pursue, and whether something may be done with your body or property. This right is violated when third parties usurp control over these decisions without a legitimate basis. Theft, fraud, molestation, illicit force, and illicit threats of force, amongst other things, all violate an individual's right to be their own master.

Whatever the precise content of this right and whatever its ultimate theoretical underpinnings, I maintain two things: first, it is a right that any reasonable person is committed to, and second, the unrestricted power to implement guardrails is incompatible with it. The argument for the second claim is simple (I take the first claim for granted). An unconstrained power to implement guardrails takes an essential area of modern life, our use of technology, and gives us a sort of master; it subjects us to an unchecked supervisory authority who gets to decide whether or not we may do something, thus asymmetrically and systematically subordinating our powers of choice to their normative standards, preferences, and judgment. They may not exercise this power often and they may exercise it in ways most people find congenial, but this does not render it unobjectionable; the fundamental problem with having a master is not the actual interference they perpetrate, but the fact that they are putatively entitled to interfere as they see fit, thus indefinitely subordinating your will to theirs (Ripstein 2009, 36; cf. Pettit 1997, 22). Given the role technology plays in our lives (including our private lives) this is intolerable; to have a master here is to have a master overseeing our capacity to flourish in our modern technological societies, and to define the role of technology in these societies and our lives more generally. Unless something very credible can be said to legitimate this situation, I propose it is illegitimate.

Developer liberty might be cited here; even if it fails to render an unconstrained power to implement guardrails desirable, it might be thought to at least render it permissible. I submit that it does not. The point is not that developer liberty is unimportant. The point is that by choosing to distribute their technology on public markets (and we should probably include open source "markets" here), developers should be taken to renounce some of their liberty claims and cannot cite them as a legitimate basis for controlling their users. To offer an analogy, private citizens have wide latitude to live according to their religious principles, but they lose some of this if they choose to occupy certain public roles. A business owner may not refuse to hire "heathens", nor may they turn them away from their store, and a county clerk may not refuse to issue marriage licenses per the laws of the land – no matter how sincere their opposition is to doing so.

Developers may not, of course, be forced to do anything; if they sincerely want no part in how some technology is likely to be used, nobody is forcing them to build or deploy it. But if they do decide to build and deploy it – if they want to be the ones who develop the technologies that so frequently define our contemporary forms of life – then any liberty claims they may have to control their users should be subordinated to the liberty of those users themselves.

The obvious alternative strategy for legitimation is an argument from necessity: some technologies are too liable to abuse to be released without constraints, no public authority is capable of regulating these technologies fast enough (or perhaps proficiently enough), and it is better to release these technologies with guardrails, even dictatorially imposed ones, than to withhold them altogether. There is a sense in which I believe this argument is successful, but we should not take it to justify a broad, discretionary power to implement guardrails; developers can respond to genuine necessities without presupposing such objectionable authority.

The key is to see guardrails as justified (when they are justified) not by some special authority developers have over their users, but by the general authority every one of us has to protect one another from having their rights violated. We do, after all, sometimes have this authority. Third parties may, for example, intervene to prevent theft and assault. They may also often intervene, I would think, to prevent the non-consensual creation of pornography, e.g. by smacking a camera out of the hands of a pervert spying on people in a public bathroom. There are limits here. For one thing, not all rights are necessarily enforceable by third parties (e.g. a promise arguably generates a right to performance, but third parties cannot typically enforce this). Moreover, any intervention to enforce rights must be proportional to the rights which would otherwise be violated (see especially Thomson 1990, ch. 4); I can't e.g. shoot you to stop minor trespassing. But within these limits, I often may and sometimes should intervene to prevent rights violations, and my authority to do this is not an illicit form of domination that violates people's right to be their own masters; it is a legitimate part of the reciprocal limits that help secure this right for everybody.

I thus propose that developers should mostly be bound by the same constraints regarding limiting user liberty that apply to everybody else; they may intervene only to prevent enforceable rights violations (i.e. where any third party could legitimately intervene), subject to a few caveats to be discussed shortly. This proposal justifies only a much narrower regime of guardrails than is currently popular, but it has two major virtues. First, it allows developers to intervene to prevent genuine, publicly enforceable abuses of their technologies, which should go a long way to addressing the most serious problems here. But second, it does this without permitting or requiring developers to make extensive judgments about how exactly their technologies should be used, leaving this decision, and responsibility, to users themselves. It does still leave users subject to developer discretion, but in a manner that entitles them to use technology in ways its developers, and even society as a whole, disapprove of. As such, I believe this proposal reconciles the realities of new technologies with the individual's right to be their own master as much as is realistically possible.

I conclude with a shotgun of brief, but important clarifications. First, where guardrails are not pervasive, they do not significantly limit user liberty, and may thus be deployed without

worrying about the considerations discussed here. Second, developers may be justified in taking additional steps to enforce their own rights that would not be permissible if pursued by third parties; relatedly, they will have means to prevent rights violations that are not available generally (e.g. the entitlement to modify their technology). Third, allowances must be made for the fact that developers will be making population-level judgments about probable rights violations; people's liberty will sometimes be curtailed to prevent other people from violating rights, which is regrettable but unavoidable. Fourth, whereas third-party rights enforcement is typically optional, I propose developers are obliged to take reasonable steps to prevent their technologies from being used to perpetrate them (so as to not be complicit in them). Specifically, I propose they are obliged to intervene where: (i) these rights violations are reasonably foreseeable; and (ii) there are available interventions that would prevent this without undue loss of liberty for users of the technology generally (what counts as an "undue loss of liberty" must be assessed on a case-by-case basis, as it depends on the rights that will otherwise be violated).

What I am asking of developers is still hard, they can only do so much, they will inevitably make mistakes, and all we can ask is that they do their best. Difficult though it may be, where the stakes are high enough, we should really employ the public mechanisms of law to do this sort of work. Insofar as it is unclear how to do this, we should figure it out; the idea that we can't make progress on this front is hard to believe. My core proposal, however, is not that we should offload this work to the law, but that it mostly should not be done at all; where somebody wants to use technology in a way that violates nobody's rights (including, potentially, the rights of collective entities), they should typically be free to do so.

References

Benkler, Y.: A Political Economy of the Origins of Asymmetric Propaganda in American Media. In: Bennett, W., Livingston, S. (eds.) The Disinformation Age, pp. 21-48. Cambridge University Press, Cambridge (2020)

Benkler, Y., Faris, R., Roberts, H.: Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics. Oxford University Press, Oxford (2018)

Fried, C.: Two concepts of interests: Some reflections on the Supreme Court's balancing test. Harv. L. Rev. 76, 755-778 (1962)

Hart, H.L.A.: Intention and Punishment. In: Hart, H.L.A. Punishment and Responsibility: Essays in the Philosophy of Law, pp. 113-135. Oxford University Press, Oxford (2008)

MacLean, N.: Since We Are Greatly Outnumbered. In: Bennett, W., Livingston, S. (eds.) The Disinformation Age, pp. 49-77. Cambridge University Press, Cambridge (2020)

Oreskes, N., Conway, E.M.: Merchants of Doubt. Bloomsbury Publishing USA, New York (2010)

Pettit, P.: On the People's Terms: A Republican Theory and Model of Democracy. Cambridge University Press, Cambridge (2012)

Pettit, P.: Republicanism: A Theory of Freedom and Government. Clarendon Press, Oxford (1997)

Ripstein, A.: Force and Freedom: Kant's Legal and Political Philosophy. Harvard University Press, Cambridge (2009)

Scanlon, T.M.: Moral Dimensions: Permissibility, Meaning, Blame. Harvard University Press, Cambridge (2008)

Scanlon, T.M.: Rights, goals, and fairness. Erkenntnis 11, 81-95 (1977)

Sen, A.: Rationality and Freedom. Harvard University Press, Cambridge (2002)

Shiffrin, S.V.: Paternalism, unconscionability doctrine, and accommodation. Philosophy & Public Affairs 29, 205-250 (2000)

Thomson, J.J.: Physician-assisted suicide: Two moral arguments. Ethics 109, 497-518 (1999)

Thomson, J.J.: The Realm of Rights. Harvard University Press, Cambridge (1990)

Winner, L.: Technologies as Forms of Life. In: Winner, L. The Whale and the Reactor: A Search for Limits in an Age of High Technology, pp. 3-18. The University of Chicago Press, Chicago (1986)