5

Freedom, Agency, and Information Technology

So far, we have given some arguments about what we owe to each other, and we have supported these arguments in part by relating them to some of the ways algorithmic systems and technologies can conflict with our autonomy. One of our key tasks has been to examine the various complaints and criticisms of algorithms and other information technologies in terms of *wrongs* rather than harms. To this end, we have argued that these systems can create circumstances people cannot reasonably endorse and they can preclude people from information they are owed.

In Chapter 4, we discussed two forms of agency that go beyond the mere manifestation of intentional action and its natural causes.[1] One sort is practical agency, which involves making decisions, formulating plans, and executing strategies. The other is cognitive agency, which involves exercising evaluative control over our mental attitudes, in the form of personal consideration of our beliefs and values.[2] Autonomy, meanwhile, involves something above mere agency: self-government. Our view is that autonomy requires both procedural independence (requiring that an agent be competent and that their beliefs, preferences, desires, and values be authentic) and substantive independence (requiring that an agent be supported by their social and relational circumstances).

Agency is a broader concept than autonomy, in the sense that it is possible for one to act, decide, or plan without those actions or attitudes exemplifying the relevant sorts of self-government. However, we will not take a strong position here on the metaphysical conditions distinguishing autonomy from agency, because our focus is on both cases of diminished agency and diminished autonomy. Our argument in this chapter is that both sorts of case result in a shortfall of freedom, properly understood.

We argue in Section 5.1 that freedom has two fundamental conditions: that persons be undominated by others and that they have an adequate degree of autonomy and agency. However, we will argue in Section 5.2 that algorithmic systems can threaten

[1] The standard theory of agency holds that the agency involves intentional action and its causes. For the classic versions of this sort of account, see Davidson, *Essays on Actions and Events*; Goldman, *Theory of Human Action*; Bratman, *Intention, Plans, and Practical Reason*.

[2] See also Hieronymi, "Two Kinds of Agency."

https://doi.org/10.1017/9781108895057.005 Published online by Cambridge University Press

both the domination-based and the agency-based requirements, either by facilitating domination or by exploiting weaknesses in human agency. We will explicate these types of threats as three sorts of challenges to freedom. The first we discuss are "affective challenges," which involve the role of affective, nonconscious processes (such as fear, anger, and addiction) in human behavior and decision-making. These processes, we argue, interfere with our procedural independence, thereby threatening persons' freedom by undermining autonomy. The second type of challenge is what we call "deliberative challenges." These involve strategic exploitation of the fact that human cognition and decision-making are limited. These challenges also relate to our procedural independence, but they do not so much interfere with it as they exploit its natural limits. A third sort of challenge, which we describe as "social challenges," involves toxic social and relational environments. These threaten our substantive independence and thus, our freedom. In Section 5.3, we sketch a policy agenda aimed at combating these challenges and promoting the conditions of freedom.

We have two main goals in this chapter. One is to extend our analysis to the affective, deliberative, and social challenges that algorithmic systems pose. The other is to relate our overall project to the notion of freedom. Our understandings of freedom and autonomy are closely linked, and it would be possible to consider affective, deliberative, and social challenges of algorithmic systems solely in light of autonomy. However, such an analysis would fail to connect algorithmic systems to a good that is on many views basic. Exploring these issues from the vantage point of freedom allows us to make that connection. In Chapter 6 we will draw on this conception in our discussion of epistemic paternalism.

5.1 FREEDOM AS UNDOMINATED SELF-GOVERNMENT

5.1.1 *The Forms of Freedom*

The concept of freedom is often discussed in terms of one's "negative" freedom; that is, in terms of noninterference, nonaggression, noncoercion, or in general, the absence of external constraints on one's ability to act.[3] A person enjoys negative freedom when they are not being actively interfered with or coerced. Negative freedom is often considered primary, for a couple of reasons. One is that limitations on negative freedom are intuitively obvious; if one is prohibited by law or physically prevented from taking an action, their freedom has been restricted. Another reason is that some shortfalls in negative freedom are especially restrictive, which makes negative freedom appear to be the most important form of freedom. For instance, if one has been kidnapped or forced at gunpoint to hand over one's possessions, one's current political freedoms are of secondary concern.

[3] On noninterference, see Berlin, "Two Concepts of Liberty"; Butt, *Rectifying International Injustice: Principles of Compensation and Restitution between Nations*; Rothbard, *For a New Liberty: The Libertarian Manifesto*; Hayek, *The Constitution of Liberty*.

https://doi.org/10.1017/9781108895057.005 Published online by Cambridge University Press

Yet, looking beyond the obvious importance of negative freedom to human life, the value of noninterference does not swamp the value of all other forms of freedom one could value. Freedom must be understood as in two ways extending beyond the mere absence of interference. To see why, consider two cases. First, consider the case of a woman living under an oppressive political regime in which she is permitted to vote only with her husband's permission. Supposing that her husband does, in fact, allow her to vote, she would have negative freedom. After all, no one has coerced her, interfered with her, or physically prevented her from voting. But her husband could have chosen otherwise and could have interfered. Depending on someone else's approval to vote is a deeper sort of unfreedom underneath her thin, negative freedom to vote. Second, consider the case of a person who is not dependent on the approval of someone else to vote, but who *does* require a wheelchair for mobility. If there are stairs that they must navigate to access their polling place, and there are no ramps or alternative means of casting a ballot, they are not free to vote, their ample negative freedom (insofar as no one coerces, prohibits, or physically blocks access to voting) aside. They are unfree because they cannot effectively do what they wish to do, namely vote.[4]

One conception of freedom that extends beyond negative freedom, corresponding to the case of the first woman, is called "republican freedom." Beyond mere noninterference, republican freedom requires non-domination, which is to say, the absence of arbitrary power or domination. On this conception, even the possibility of interference counts as unfreedom. Noninterference that occurs at the whim of a benevolent ruler does not count as genuine freedom. For republican freedom, it is not enough that there is no interference. Rather, as Philip Pettit puts it, interference must be "robustly" absent, meaning that interference is "absent over variations in how far others are hostile or friendly."[5] For the first woman to be free, her husband's benevolence cannot play a role in her ability to vote; she must be able to do it without needing his permission.

The republican theory of freedom does not emphasize that the fullest sort of freedom requires having not just choices, but autonomous choices. Pettit, for instance, argues that his account "does not require that resources [. . .] must be present over variations in your personal skills, the natural environment, or the structure of society," because "[i]n order to choose freely between the options in a certain choice, it is enough that you actually have such assets available."[6] The republican theory holds people to account for the choices they make, no matter why they made them or what challenges they faced leading up to the choice. Yet, if the

[4] One might object here that having stairs, while lacking ramp access or alternative means of voting, is indeed a way in which one is physically prevented from voting. That is true, and the line between negative freedom (lack of external constraints on one's actions) and positive freedom (the ability to carry out one's intentions) is blurry. See Cohen, "Freedom and Money."

[5] Pettit, *Just Freedom*, 50.

[6] Pettit, 50.

https://doi.org/10.1017/9781108895057.005 Published online by Cambridge University Press

affective and deliberative challenges to human agency are real, it must be in part through their influence on our choices. To fully understand our freedom, we must take into consideration not just our choices, but how those choices get made.

Another conception of freedom that is distinct from negative freedom, corresponding to the case of the second voter, is "positive freedom." It involves not only the absence of external constraints, but the ability to effectively carry out one's interests. While negative freedom is defined in terms of (and, to some extent, presupposes) the capacity for autonomy,[7] positive freedom requires the ongoing exercise of self-government. Republican freedom, moreover, cannot by itself guarantee positive freedom. This is because the republican account is primarily oriented toward the alleviation of subjugation, or "defenseless susceptibility to interference,"[8] rather than to the positive development of persons' talents and capacities. But this glosses over the fact that one might be undominated by all others but nonetheless have interests that one cannot pursue because they lack the capacities to effectuate their desires or realize their interests. For the second voter to be free, they must not only be undominated by all others, their polling station must also have the relevant accessibilities, allowing them to actually vote rather than merely have the (empty) right to do so.

"Unfreedom" in the positive sense can be tricky to diagnose or discern, in part because it can become fully internalized. Consider, for instance, a third person who has been raised in an insular religious community. Suppose that in this community, people adhere to a narrow set of religious precepts, their access to information is restricted by their religious leaders, and their social roles are defined in terms of highly gendered categories. By the time the person reaches adulthood, they are likely to have internalized their community's standards about the nature of personhood and about what counts as an adequate range of options. Unlike the first woman, this person is not externally constrained, in the sense that someone could prevent them from acting according to their desires, and unlike the second person, their choices have not been undermined or frustrated by inaccessibility. Indeed, they may even possess all the material resources necessary to exit the community. Yet there is an important sense in which they lack freedom, owing to how their space of possibility has been circumscribed; their agency itself has been short-circuited in a way that they cannot repair or (possibly) even recognize. To be clear, this is a generic, hypothetical example. Gesturing at this kind of example does not justify any inference about the authenticity or "false consciousness" in any real cases. Rather, it is an example to show that freedom is not reducible to persons' abilities to act on their preferences, as preferences can be formed in response to oppressive conditions, which are themselves antithetical to freedom.

It's worth pausing to explain this a bit more. We have distinguished negative freedom (freedom from external constraints) and positive freedom (ability to

[7] Raz, *The Morality of Freedom*; Feinberg, "Autonomy."
[8] Pettit, "Freedom as Antipower," 577.

effectively carry out one's interests). One might argue that these apparently distinct facets of freedom are reducible to a single conception. Gerald MacCallum, for example, argues that we can reconcile negative and positive freedom under a single, tripartite conception of freedom: between (1) a person, (2) their goals (their "doings" or "becomings"), and (3) the relevant constraints.[9] From this perspective, understanding freedom as primarily being about external constraints or as primarily being about individuals' effective capacity to act is a mistake. Rather, what matters is a person, some action they wish to take or way they would like to live, and whether there are constraints on the person's ability to take the action or to live the way they would like. Hence, a person driven by an addiction might be free of external constraints, but unfree insofar as their addiction prevents them from acting or living in a way that comports with their higher-order desires.

But even a conception that collapses positive and negative freedom cannot adequately explain our third case. The problem there is that despite the fact that there are no external constraints and despite the fact that the person is able to act upon their values, it is odd to say that they are free tout court. Being raised in oppressive circumstances may preclude them from developing reasonable sets of values and preferences. The limitations of freedom in this case are, as Christman puts it, to their "quality of agency." That is, a person who has had unreasonable constraints on their ability to develop their own sense of value has diminished "effectiveness as an agent."[10] More is needed to secure their effectiveness as an agent than the mere absence of external constraints and ability to act upon their preferences. Quality of agency requires that their beliefs and desires be both authentic and competently acquired for an adequate range of options. It also requires social and political structures that support those beliefs and desires and that one has sufficient affordances to act on those beliefs and desires. This quality of agency view can explain how challenges to authenticity, such as in cases of severe addiction, limit freedom. For instance, we might imagine a person who is driven by addiction but who also has enough resources and opportunities to sustain their habits in perpetuity.[11] Such a person lacks freedom in the fullest sense because their addiction limits their efficacy as an agent.

Our view of freedom aims to capture both freedom as quality of agency (which itself encompasses both negative and positive freedom) and non-domination (which encompasses republican freedom). In this sense it is as ecumenical as our conception of autonomy from Chapter 2. Specifically, we recognize both non-domination and autonomy as vital facets of freedom.[12] An agent's freedom requires what we will call "ecological non-domination." It is ecological in the sense that one's freedom encompasses both facts

[9] MacCallum, "Negative and Positive Freedom."
[10] Christman, "Saving Positive Freedom," 80.
[11] Eric Clapton, for instance, has estimated that during the worst periods of his addiction, he was spending $16,000 per week. See "A Life in Twelve Bars."
[12] To some degree we are taking our cue from Anderson, *Private Government*, chapter 2.

about oneself (quality of one's agency) and their environment (absence of external constraints and non-domination). One is not free without both.

Notice that it is possible to enjoy some aspects of autonomy without full republican freedom (or even negative freedom). Consider, for instance, the "Russian oligarchs"; the group of wealthy Russian kleptocrats who rapidly acquired enormous wealth in the wake of post-Soviet privatization. They enjoy unfettered positive freedom, in that they have at their disposal the means to purchase world-class football clubs,[13] record-breaking superyachts,[14] and former royal estates.[15] Yet, at the same time, their enjoyment of noninterference is not all that robust: Their personal safety depends on remaining in good favor with the regime. Mikhail Khodorkovsky's oil company (Yukos) came to control a sizable fraction of Russia's oil supply. Yet when Khodorkovsky engaged in a power struggle with the government, he was imprisoned for nearly a decade over trumped-up fraud charges. Similarly, the chairman of the country's biggest telecom, Vladimir Yevtushenkov, was placed under house arrest by Russian authorities under suspicion of money laundering.[16] Indeed, to some extent even the president of Russia, Vladimir Putin,[17] faces this sort of precarity: Both his wealth and safety depend on retaining political power, and this in turn requires at least minimal cooperation with the oligarchs. For him, interference is absent, but perhaps not robustly absent. That is, it may not be absent over variations in the friendliness of others.

5.1.2 *The Value of Freedom*

It is still an open question as to the relation between freedom and morality. Why, in other words, is freedom good or morally significant (if it is)?

As in our discussion of autonomy earlier in this book, we can understand the value of freedom by considering its boundaries. We have already seen that autonomy is not good without qualification. For example, using one's autonomous capacities to undermine others' freedom is bad. Consider the oligarchs just mentioned: not only does their freedom permit them to capture a sizable proportion of the wealth and national product of Russia for their personal benefit; they also have a long history of using their freedom as a tool for promoting the domination and diminished quality of agency of others. With this sort of example in mind, John Danaher argues that freedom is an "axiological catalyst" – that it "makes good things better and bad things worse."[18] Yet even this seems too strong when we consider cases that raise the "paradox of choice."[19] This paradox is that although we often believe that

[13] BBC Staff, "Russian Businessman Buys Chelsea."
[14] Segal, "A Russian Oligarch's $500 Million Yacht Is in the Middle of Britain's Costliest Divorce."
[15] Cramb, "Scotland's Most Expensive Sporting Estate Bought by Russian Vodka Billionaire."
[16] *Guardian* staff, "Sistema Boss Arrested in Russia on Money-Laundering Charges."
[17] Aslund, *Russia's Crony Capitalism*.
[18] Danaher, "Freedom in the Age of Algocracy"; Kagan "The Additive Fallacy."
[19] Schwartz, *The Paradox of Choice*.

having more options is good, in fact having a large number, wide diversity, or substantial nuance in our choices leads to anxiety and stress rather than happiness or satisfaction. A greater range of choices is simply not always better, for either the chooser or those who depend on their choices.

These cases suggest that it will be impossible to coherently explain either freedom or autonomy as intrinsically good. Of course, we need not ignore the fact that freedom and autonomy are almost always good. An easier proposition to defend is cast in terms of unfreedom: shortfalls in freedom (that is, either shortfalls in self-government or instances of domination) are prima facie bad. This commitment is compatible with freedom being outweighed by any number of other values in any number of cases, such as when we take the freedom of oppressors to be outweighed by the value of human rights. There may be cases where the protections licensed by the assumed value of freedom are morally and politically outweighed by the public interest, such as public health mandates. The shift in emphasis from freedom to unfreedom is motivated by the fact that in practice, it is easier to discern what is objectionable (or not) about a given freedom deficit than to discern how freedom (conceived of as an intrinsic good) must be limited.

This account has substantive implications. It holds that freedom deficits are usually morally bad and that unfreedom can be thought to serve morally good purposes only in certain circumstances. Figuring out how to best promote human freedom, then, is not as simple as settling debates about which sorts of interference would be objectionable to rational actors. Unlike freedom, which can be defined in the abstract, unfreedom can only be defined in terms of the specific foibles of human agency, specific architecture of human environments, and specific implications of social choices. As we show in the next section, the effects that shortfalls in freedom can have on our freedom are well illustrated in terms of the challenges to freedom mentioned earlier.

5.2 THREE CHALLENGES TO FREEDOM: AFFECTIVE, DELIBERATIVE, AND SOCIAL

At the start of this chapter we introduced three challenges to freedom: the affective challenge, the deliberative challenge, and the social challenge. In this section we will examine each of these challenges by considering several different algorithmic systems. We will argue that, to the extent that these technologies can undermine either people's autonomy or their freedom from domination without good reason, they can objectionably undermine our freedom.

5.2.1 *Affective Challenges to Autonomy*

The first sort of challenge to freedom is that human behavior is driven by affective influences, such as fear, anger, resentment, or addiction, as opposed to purely being driven by rational attitudes. These affective states, we will argue, can undermine the

https://doi.org/10.1017/9781108895057.005 Published online by Cambridge University Press

authenticity of people's preferences and desires, threatening our freedom by threatening our autonomy.

People are so driven by affective states that it might seem difficult or impossible to clearly pinpoint any kernel of rational attitudes that lie underneath. Humans, it might seem, experience affective influences "all the way down," in the sense that our authentic attitudes cannot be distinguished from any other attitudes one might have in a principled way.[20] But in some cases, the influence can be so dramatic that it is hard to accept that the resulting behavior could be freely chosen or even independently motivated. The AAA Foundation for Traffic Safety has found, for instance, that around 4 percent of drivers each year have gotten out of their cars to angrily confront another driver, and 3 percent have purposefully run into another car for the same reason.[21] Anger is not the only source of authenticity-undermining affective influence: It is difficult to believe that a person could have an authentic desire or preference to go to an internet cafe and play a game for fifty straight hours before dying of exhaustion, yet this has happened on several occasions.[22]

As has been known since B. F. Skinner, people's susceptibility to operant conditioning can be used by third parties (such as casinos) to reinforce certain patterns of behavior over others. Many digital platforms employ conditioning strategies in the same sort of way, engendering affective states in their users that undermine the authenticity of their choices. The targets of optimization are different, of course; mobile developers optimize in terms of clicks, views, watch times, and so on, rather than lever pulls or revenue benchmarks. However, the underlying strategy – engendering some sort of artificial dependency – is the same.

As Nir Eyal points out, engendering artificial dependencies is a crucial element of the general strategy of keeping users "engaged." He outlines a four-step method for effectively "Skinner boxing" a digital platform.[23] The first step involves a trigger, which draws the user's attention to the app or platform. The quintessential modern trigger is, of course, the "push notification," which literally forces you to pay attention to it enough to dismiss it. The second step of Eyal's method involves getting users into acting on the trigger. The best way of getting users to do this is to get them to anticipate some sort of reward. The third step involves tying user behavior to these rewards but making the rewards variable, in the sense that they are sometimes highly rewarding but at other times mundane, in a way that is unpredictable. Dopamine surges whenever the brain has been conditioned to expect a reward, but without variability, the experience becomes unsatisfyingly predictable. The fourth step, finally, involves getting users to make some sort of investment in or commitment to the app, to maintain their engagement with it in the future.

[20] Thanks to Sarah Worley and Dana Nelkin for drawing our attention to this point.
[21] AAA Foundation for Traffic Safety, "Aggressive Driving | AAA Exchange."
[22] BBC Staff, "S Korean Dies after Games Session."
[23] Eyal, *Hooked: How to Build Habit-Forming Products*, 8–9.

The Duolingo platform offers an illustration of how Eyal's model can be employed. The app uses push notifications to draw the user's attention to the app, sometimes coupled with a guilt trip. When a user has stopped using Duolingo for a period, they will receive a notification depicting the very cute Duolingo mascot in tears, stating that "Language Bird is crying," that a failure to go back to learning a language will lead the mascot to "eat a poison loaf of bread," and that the next email to the user will be an e-vite to Language Bird's funeral.

Skill badges, which represent user achievements, are displayed at the end of rounds in a pleasingly unpredictable way. Finally, the platform has a number of mechanisms for investment: users' earned skills decay over time, requiring ongoing practice, and the platform itself also contains a microtransaction market for premium avatars, allowing users to invest in the app using real money.

In several ways, then, Duolingo uses its users' affective states to undermine their autonomy and, thus, limit their freedom. For the most part,[24] its freedom-impinging tactics do not raise anyone's hackles, because there is a background presumption in any discussion of it that language acquisition really *is* valuable for its users (thereby serving their global autonomy if not quite their local autonomy). The same tactics, however, are more troubling in the context of employment, where people can be compelled to act in dangerous ways to protect their basic livelihood. Consider the practices of the ride-hailing company Uber.

Uber has come under substantial criticism on the basis of how its algorithms have negatively affected its drivers.[25] The company uses these systems for a variety of purposes: to track passengers, to anticipate market demand, and (at one point) even to identify and deceive regulators, media, and law enforcement. As we discuss in Chapter 7,[26] the company employs several practices to keep its drivers working longer than they might otherwise. Here we will discuss two sorts of practice, arguing that one serves to undermine the freedom of drivers while the other is more morally benign.

Both sorts of practice can be linked to the strategy outlined by Eyal. The first involves Uber's user interface design choices. The driver app, for instance, is configured by default to remind drivers of their goals through push notifications, and it is also configured by default to queue another rider before the current rider has been delivered. Presumably, the motivation for Uber to employ this particular set of defaults rather than the alternative sets is that their chosen configuration is statistically more likely to induce drivers to accept more riders than the others. (In fact, it is likely that Uber has discerned the precise overall effect of this design choice through massive-scale A/B testing.)

Nevertheless, Uber's design choices do not necessarily undermine drivers' freedom in this case. It is plausible that drivers have their own reasons for such settings

[24] However, see Lee, "Duolingo Redesigned Its Owl to Guilt-Trip You Even Harder."
[25] Rosenblat, *Uberland*, 98–100.
[26] Rubel, Castro, and Pham, "Agency Laundering and Information Technologies."

being one way or the other. It is plausible that someone might want to work a shift determined by a particular monetary goal or period of time and, thus, to simply queue as many customers as fast as possible without having to continue manually making choices at each possible choice point. Uber would certainly not increase the freedom of users by doing away with the choice altogether. So long as Uber does not covertly reset the defaults, or design the interface in a predatory fashion, the company's UI design choices do not strike us as an issue of serious moral concern (or even, for that matter, an issue of freedom at all).

The second practice for keeping drivers working longer than they would otherwise, however, is less apparent to the drivers and much more troubling from the standpoint of autonomy. It involves such practices as "surge pricing," which is a cost-multiplier that raises the price of a ride (sometimes dramatically) in a location at a time when the demand for drivers is high or the supply of drivers is low. Some of the additional revenue from this cost markup is passed to the drivers, so the surge-pricing system gives drivers a financial incentive to work more than they would otherwise, because they will be able to make more money than they would have otherwise. The "surge-pricing" mechanism is a familiar device in modern retail – it's just a form of dynamic pricing – but essential to Uber's use of it is an element of uncertainty and variability that is not readily apparent to the drivers at first: when drivers accept a ride at a "surge priced" rate, they do not know whether the ultimate payment for the ride will be surge priced.[27]

The fact that these practices are opaque to drivers might lead one to question Uber's motives. Our concern here, though, is the drivers' autonomy. A recent report describes the phenomenon of "chasing the surge" at length.[28] Since 2016, Uber has employed the "boost" system, which grants automatic surge pricing to drivers who have completed a certain number of rides the previous week, in a tiered arrangement from "Bronze" to "Platinum." One driver described his relationship with Uber in the following way:

> I had some days off from work. So I was on the road. I mean, I just had coffee upon coffee, and I'm just on the road. So I ended up doing about, I thought it was over 100, but I did 94 rides . . . It was 94 rides in essentially 3 days . . . After you do 90 rides, I think it's 90 or 95, they bump you to platinum . . . And then basically you're always chasing platinum.[29]

The authors of the report note that "[t]he stress of these games, be it chasing a surge or platinum, was a refrain in 50% of our interviews in surveys."[30]

What is the moral upshot of the Uber case? According to our view of the value of freedom, the verdict is mixed. The nudges Uber employs in its user interface, which

[27] Rosenblat, *Uberland*, chapter 4.
[28] Wells and Cullen, "The Uber Workplace in Washington, D.C."
[29] Wells and Cullen, 47.
[30] Wells and Cullen, 47.

for the most part serve obvious functions for its users (and which are also relatively low stakes and transparent), are unproblematic as a class. From the perspective of alleviating the domination of its users or promoting their autonomy, it is not clear that Uber could have done better.

Uber's Skinner-boxing of its drivers to "chase platinum," in contrast, is considerably more troubling and gives rise to a freedom-based complaint. The authenticity condition of autonomy precludes users being in affectively compromised states; if people's actions lack authenticity – in the sense that those actions serve to alienate the person from their basic values – then those actions will not be autonomous. However, the boost aspect of the Uber surge-pricing system exists precisely to place users in such a state and, in so doing, urges them to "chase" overwork even at the expense of their health. The same freedom-impinging features that seem somewhat innocuous in the context of Duolingo and language acquisition become more troubling when a dangerous degree of overworking is a potential effect.

5.2.2 *Deliberative Challenges to Agency*

Exploiting our emotional vulnerabilities is not the only way to hack human agency. Even if we set aside the fact that human agency can be undermined by noncognitive affective states, human cognition itself is limited in ways that can be exploited to limit our quality of agency.

Our actual deliberative process is both *bounded* and *inaccurate*; the information we receive as reasoners is limited by our computational capacities and fitted to our choice environments, and our processing of the information we do get is subject to a variety of errors and biases.[31] Many of these biases – the availability bias, anchoring bias, conjunction fallacy, base-rate neglect, and so on – are well known.[32] To this end, in the next chapter we discuss problematic search suggestions and algorithmically curated news feeds, which allows people to expose themselves only to information that reinforces their existing beliefs. Here, we focus more on the boundedness of human agency.

Algorithmic systems, particularly digital platforms, introduce new sources of economic value – namely personal and consumer datasets at big data scale[33] – as well as new sources of risk – data breaches and other forms of exposure.[34] These systems can be so complex both for developers and for end users that it is impossible to live up to the Reasonable Endorsement Test in practice. To be sure, the choice environment of our digital daily lives does not live up to this moral standard: most people do not and

[31] Simon, *Models of Man: Social and Rational-Mathematical Essays on Rational Human Behavior in a Social Setting*.

[32] Kahneman, Knetsch, and Thaler, "Experimental Tests of the Endowment Effect and the Coase Theorem."

[33] Pham and Castro, "The Moral Limits of the Market: The Case of Consumer Scoring Data."

[34] Yaffe-Bellany, "Equifax Data-Breach Settlement"; Abrams, "Target to Pay $18.5 Million to 47 States in Security Breach Settlement."

https://doi.org/10.1017/9781108895057.005 Published online by Cambridge University Press

could not always read all of the fine print to which they claim to consent and the idea of "common terms" has been stretched thin by massive swathes of boilerplate that are unreadable by humans in practice.[35] In practice, we face an avalanche of digital "pseudo-contracts" – contracts that are so convoluted that they assume impossible levels of human competence and thus fail to represent any actual "meeting of the minds" between the parties.[36]

Our constant subjection to these "contracts" violates our procedural independence, by exploiting the limits of our epistemic competence and thereby undermining our consent. One study, for instance, found that "only one or two of every 1,000 retail software shoppers" actually access the contract "and that most of those who do access it read no more than a small portion."[37] In 2014, Europol conducted an experiment in which they offered access to a Wi-Fi hotspot behind a contract that granted access only if the recipient signed a so-called Herod clause, in this case, to agree "to assign their first born child to us for the duration of eternity."[38] Six people still signed the contract. In 2015, a *Guardian* journalist resolved to read all the terms of conditions for all his services and wound up reading 146,000 words in a single week.[39]

Most modern web services now have end-user licensing agreements (EULAs), but Brainly, a peer-to-peer learning platform in which users can provide homework help to one another, offers an especially striking example of a pseudo-contract. The platform's end-user licensing agreement is expansive. First, the company is permitted to force users into arbitration. Second, it is permitted – despite its claim to protect users with a privacy policy – to license user content to third parties, distribute it through any media, or even sell personal data outright as part of bankruptcy, *even after its users have terminated their accounts*. Third, it is permitted to change the terms of service at any time without notice.

Each of the components of the Brainly pseudo-contract can be seen to exploit the undermined capacity of its users to reflect on their values, motivations, and decision-making in the digital environment. The company's ability to automatically force users into arbitration allows it to settle disputes with its users on legal terrain that is broadly unfavorable to them; its privacy policy serves as a smokescreen that obscures the extent to which it can freely trade or sell user data; its right to change its terms of service at any time prevents and discourages users from reading any one version of it and contributes to the overwhelming deluge of information users must cope with.

This pseudo-contract exploits its users' limited quality of agency. By holding them responsible for self-government in a context where this is humanly impossible, the contract exploits their lack of understanding of their circumstances for corporate

[35] Benoliel and Becher, "The Duty to Read the Unreadable."
[36] Kar and Radin, "Pseudo-Contract & Shared Meaning Analysis," 1140.
[37] Bakos, Marotta-Wurgler, and Trossen, "Does Anyone Read the Fine Print?"
[38] Fox-Brewster, "Londoners Give up Eldest Children in Public Wi-Fi Security Horror Show."
[39] Hern, "I Read All the Small Print on the Internet and It Made Me Want to Die."

benefit. In many cases, the stakes are sufficiently low: the opacity of Brainly's EULA aside, it is difficult to argue that anyone has been harmed by having their content unknowingly open-sourced to the educationally minded public. Yet users have proven eager in the past to allow their likenesses to be sold by Instagram,[40] and (as we discuss in Chapter 8) they have more recently allowed their (and their friends') personality profiles to be mined by Cambridge Analytica, even if they later come to regret these choices. In the context of complex algorithmic systems, epistemic competence – and therefore, freedom – is impossible to achieve.

5.2.3 *Social Challenges to Freedom*

So far in this section, we have argued that one way that human agency can be undermined is if its processes come to be directed by affective states rather than by conscious deliberation, violating the authenticity condition. Another way agency can be undermined, we have argued, relates to the fact that our cognitive capacities themselves are bounded, violating the competence condition. In this subsection, we explore a third way that agency can be undermined: through the influence of the affective states and erroneous deliberative processes *of other people*, in the context of an epistemologically toxic social environment.

Consider, for instance, YouTube's recommendation algorithm, whose autoplay mechanism drives users toward echo chambers and other ideologically extreme and conspiratorial content. By YouTube's own analysis, almost three-quarters of viewing time spent on the site is driven by this recommendation system as opposed to their viewers' independent actions,[41] so the operation of this system is of overriding significance in determining what content people are exposed to. When left unsupervised by humans, it fosters an unsafe environment for children: it allows bad actors to expose them to distressing parody content, such as a fake Mickey Mouse cartoon depicting eye-gouging[42] or a fake Paw Patrol episode where one of the main characters attempts suicide.[43]

YouTube's recommendation system has also arguably been a source of violent self-radicalization. The reason has at least in part to do with YouTube's business model and with the evolution of its incentive structure. Originally, the YouTube recommendation algorithm had been designed simply to maximize the number of clicks,[44] but this proved to offer content creators an incentive to post "clickbait" videos, ones that entice users to initially click on the video but that do not necessarily hold their interest through the duration of the video. In light of this, YouTube

[40] Arthur, "Facebook Forces Instagram Users to Allow It to Sell Their Uploaded Photos." Instagram reversed this policy after public outcry.
[41] Neal Mohan, YouTube's Chief Product Officer, claimed this at the 2018 CES convention.
[42] Orphanides, "Children's YouTube Is Still Churning Out Blood, Suicide and Cannibalism."
[43] Maheshwari, "On YouTube Kids, Startling Videos Slip Past Filters."
[44] Roose, "The Making of a YouTube Radical."

changed its algorithm in 2012 to optimize more for *watch time* than for number of clicks so that "creators would be encouraged to make videos that users would finish, users would be more satisfied and YouTube would be able to show them more ads."[45] The change worked – watch time increased by 50 percent each year from 2012 to 2015 – but it also conferred an advantage to those content creators who produce naturally engaging videos: conspiracy theorists.

Guillaume Chaslot, a former Google engineer who left the company over issues with the development of the YouTube algorithm, wrote a piece of web software that examines which videos follow others via the recommendation system. When the *Guardian* examined the one thousand top-recommended videos, it found that YouTube had an 85 percent chance of recommending a pro-Trump video rather than a pro-Clinton video. The newspaper also interviewed several conspiracy theorists, whose videos normally receive only a few hundred views but whose traffic increased dramatically right before the 2016 election, and it found that most of these content creators got their traffic from the YouTube recommender system rather than from external links.[46] These videos, which almost always have titles like "WHOA! HILLARY THINKS CAMERA'S OFF … SENDS SHOCK MESSAGE TO TRUMP" and "Irrefutable Proof: Hillary Clinton Has a Seizure Disorder!" are more engaging than they are factual. The natural result of this state of affairs is a social environment in which divisiveness and sensationalism are in themselves advantageous traits for content, quite apart from whether or not that content reflects misinformation or authoritarian aims.

Given the affective and deliberative challenges to agency that we have already discussed, we should not be surprised that sensational and conspiratorial content grounded in fear, anger, and resentment turns out to deliver greater engagement than the messy realities of legitimate news reporting. But it is worth noticing how toxic social environments can aggravate the other challenges to agency: When users who are already prone to fear-conditioning are fed an escalating diet of misinformed, radicalizing content, their autonomy is *socially* comprised. For such users, both the possibility of autonomy and the possibility of high-quality agency are ruled out from the start.

5.2.4 *Summarizing the Challenges to Freedom*

As we have seen in this section, drivers who have become caught up in the gamified ridesharing ecosystem, information consumers who have become overwhelmed by a deluge of labyrinthian pseudo-contracts, and conspiracy theorists who have become radicalized all have been made less free in a broadly similar way: either their autonomy has been undermined or their quality of agency has been

[45] Roose.
[46] Lewis, "Fiction Is Outperforming Reality."

diminished.[47] However, the traditional conceptions of freedom (negative, positive, republican) that we canvassed at the beginning of the chapter do not provide an adequate explanation for these shortfalls of freedom. That is because drivers, information consumers, and conspiracy theorists all seem to satisfy the traditional conceptions. They are not subject to external constraints, they are not subject to arbitrary exercises of power, and they often even have the capacity to make choices to realize their preferences. The challenges to freedom we have outlined here make clear that freedom requires autonomy, high-quality agency, and non-domination.

Taking these three challenges seriously motivates an important shift in how we think about freedom. Recall that our account of freedom is ecological non-domination, which includes quality of agency and republican freedom. That account captures the emotionally volatile, cognitively limited, and fundamentally social nature of human agency. It is also consistent with the fact that it remains possible to coherently act even under some degree of individual, psychological disunity, such as when one is internally conflicted about a course of action. Surely the authenticity of Uber drivers, the competence of overwhelmed service users, and online radicalization of conspiracy theorists are not all-or-nothing affairs. People in those circumstances are neither fully free nor acting purely on reflex. When we reconsider human agency in light of its distinctive challenges, any plausible account of freedom must cope with this complexity.

We are finally able to make more precise the sense in which these users of technology have been made unfree: their beliefs, preferences, constraints, and values have been formed in a way that conflicts with agency. Freedom requires, in addition to non-domination and effective choice, that one's beliefs, preferences, constraints, and values themselves have been formed without undue influence. It is only when people are both undominated *and* unstricken by such malformations that they can be said to be free in the fullest sense.

5.3 ECOLOGICAL NON-DOMINATION, POLICY, AND POLESTAR CASES

The next question to consider is this: what kinds of moral claims does ecological freedom ground? Both the new cases we have examined in this chapter and the cases we have examined in previous ones can offer some guidance.

Preserving our autonomy – and thus, our freedom – requires managing the possible sources of compromising affective states. However, at least in the United States, policymakers have discussed the issue only where the analogy between the digital and non-digital addiction is sufficiently robust.[48] Consider, for instance, "loot boxes," which are certain economic transactions within digital games that

[47] Rubel, "Privacy and Positive Intellectual Freedom," 399–401.
[48] See the Protecting Children from Abusive Games Act.

involve a randomized reward structure. These transactions bear a clear resemblance to traditional forms of gambling, such as slot machines and lotteries, including the fact that the model is primarily funded by a tiny proportion of "whales."[49] There is, therefore, reason to think that the "loot box" market is noxious and, thus, apt for public regulation.[50] One bill has been introduced to ban the use of loot boxes in games marketed to children, but as of 2020, there are effectively no regulations in place about them. In the European Union, in contrast, the practice has already been banned completely by the authorities in at least two member states (Belgium and the Netherlands).[51] Our account offers, on relatively minimal and ecumenical grounds, support for regulation that nudges users to be self-critical about their agency.

Preserving our quality of agency might also require designing law and policy in a way that is mindful of human cognitive limitations. Robin Kar and Margaret Radin, for instance, offer a heuristic for courts: They should imagine that all text exchanged during the formation of the contract "be converted into oral form" and then imagined to occur "in a face-to-face conversation between the relevant parties."[52] Courts then are not obliged to accept all boilerplate as meaningful from the start. Instead, they are enabled to scrutinize whether the text genuinely conforms to "the cooperative norms that govern language use to form a contract."[53]

Finally, we might consider ways of combating toxic online social environments, which serve as threats to all sorts of freedom, republican freedom included. Platforms have implemented some light-touch solutions on their own, at least for the most egregious problems. For instance, YouTube searches for content associated with the Islamic State, now attract banners promoting skepticism about their aims.[54] However, broader and more general solutions remain elusive. How can would-be political radicals be nudged toward content that promotes gentler aims and methods than violence, when the business models of the platforms hosting this content depend on the engagement produced by the more radical content? The depth of this problem suggests that the engagement-centric business model itself might need to be abandoned to see progress.[55]

[49] The name draws its source from casino slang and refers to those gamblers who are known to bet large amounts of money. And just as casinos compete for the largest high rollers, app developers depend heavily on those users who spend the most money: A recent Swrve survey showed that about 0.15 percent of mobile users contributed approximately half of all in-app purchases in "freemium" games.

[50] Satz, *Why Some Things Should Not Be for Sale: The Moral Limits of Markets*; Castro and Pham, "Is the Attention Economy Noxious?"

[51] Netherlands Gaming Authority, "Study into Loot Boxes: A Treasure or a Burden?"; Belgian Gaming Commission, "Loot Boxes in Three Video Games in Violation of Gambling Legislation."

[52] Kar and Radin, "Pseudo-Contract & Shared Meaning Analysis," 1167.

[53] Kar and Radin, 1167.

[54] Alfano, Carter, and Cheong, "Technological Seduction and Self-Radicalization," 25.

[55] Lanier, *Ten Arguments for Deleting Your Social Media Accounts Right Now*.

We can also reconsider some of the polestar cases in light of the techniques we have used to address the new cases. Do Eric Loomis or the teachers have a reason to think that their freedom was undermined?

For *Loomis*, the argument that his freedom was wrongly or unjustly curtailed is difficult to get off the ground. It is hard to argue that the use of COMPAS somehow undermined the authenticity of Loomis's desires, so his autonomy-based arguments would rest on claims about how the proprietary nature of the system exploited his diminished quality of agency. However, as the Wisconsin Supreme Court noted, most of the information used to generate his risk assessment report was either static or under his control, so agency-based concerns are also somewhat implausible. Now, he certainly had his freedom curtailed by the court, but it is hard to argue that he was somehow dominated by it in a way that defies sensible criminal justice; as the court confirmed, he likely received the same sentence he would have otherwise. So whatever autonomy- and agency-based affronts he was unreasonably subjected to, he did not face a morally objectionable affront to his freedom.

For the teachers from *Wagner* and *Houston*, the freedom-based argument is more persuasive. As we also discussed in Chapter 4, the EVAAS system produces results that are fragile (and, as they acknowledge, unreproducible) and it produces those results without any specific explanation of its method or independent oversight. And – unlike in *Loomis* – were the two teachers to have gained the necessary understanding of the system, they would have been able to act differently (and a lot more directly): They could have anticipated their problems and been prepared to impugn the results to future employers, or they could have mounted a comprehensible public campaign against the system. But they were denied such choices and therefore had no choice but to litigate the issue. To the extent that litigation against technology companies is prohibitively costly for individual litigants such as Teresa Wagner or Jennifer Braeuner (and, for that matter, Catherine Taylor and Carmen Arroyo), such sources of domination constitute objectionable affronts to freedom.

There are general lessons to be learned even for those who do not find themselves on the wrong side of such systems. Each of us has an individual obligation to accept that we are driven partly by affective states, boundedly competent, and highly influenced by others. We can be influenced to agree to transactions we might not otherwise, we are unlikely to be aware of the finer details of most of the contracts we sign, and we are especially susceptible to radical ideas – good or bad – when we encounter them from within the community in which they arose. We need not take these features of our nature to undermine the legitimacy of every agreement or transaction we make through these apps and platforms, but we must accept that these features undermine the legitimacy of some of those agreements and transactions.

5.4 WHY NOT MANIPULATION?

Manipulation is a common element of many of the cases we consider in this chapter and the next. The "blind-draw" loot boxes resemble lotteries and slot machines in all ways other than that they are digital rather than physical, luring their users to play through classic tricks of the advertising trade. Facebook's Click-Gap metric, which we discuss in the next chapter, offers the platform's developers the ability to invisibly curate individual News Feeds. Cambridge Analytica's targeting of Facebook users for the purposes of political advertising, too, involved influencing that was tailored to users' deep tendencies. Even in cases where there does not appear to be any direct manipulator, such as the case of the YouTube recommendation system, manipulation is still involved, in virtue of involving someone who seems to have been manipulated. Considering these cases, one might wonder whether our analysis of the ethics of algorithmic systems could be subsumed under the rubric of manipulation. And if not, then what *does* make manipulation objectionable as a practice?

Before answering questions relating to our analysis to manipulation, we would do well to specify the concept itself, at least loosely. However, there is no consensus in the philosophical literature on what exactly constitutes manipulation, nor is there a consensus on how manipulation relates to autonomy and freedom. Moreover, there are problematic implications related to each possible conception. Some view manipulation in terms of threats to rationality. Raz, for instance, understands manipulation in terms of undercutting rational decision-making; as "pervert[ing] the way that a person reaches decisions, forms preferences or adopts goals."[56] Yet manipulators need not always threaten rationality; in some cases rational implications can themselves be used manipulatively.[57] Others, such as Anne Barnhill, Karen Yeung, and Daniel Susser, Beate Roessler, and Helen Nissenbaum, conceive of manipulation in terms of forms of deception or hiddenness that undermine people's self-interest and autonomy.[58] This, too, seems right, but many of our key cases, such as that of Catherine Taylor (Chapter 4), reflect vulnerabilities that do not include elements of deception, hiddenness, or trickery. At times, such as in the COMPAS, EVAAS, and CrimSafe cases, the practices we scrutinize shade closer to coercion than to deception. To this end, Joel Feinberg and Allen Wood understand manipulation as lying on a "spectrum of force" between compulsion and enticement.[59] But this analysis is also incomplete: It does not cover the cases of manipulation that seem to involve deception. So none of the analyses cover all of the cases that seem to fall under the concept of manipulation.

[56] Raz, *The Morality of Freedom*, 377–378.

[57] Gorin, "Do Manipulators Always Threaten Rationality?"

[58] Barnhill, "What Is Manipulation?"; Yeung, "'Hypernudge': Big Data as a Mode of Regulation by Design"; Susser, Roessler, and Nissenbaum, "Online Manipulation: Hidden Influences in a Digital World"; Lanzing, "'Strongly Recommended' Revisiting Decisional Privacy to Judge Hypernudging in Self-Tracking Technologies."

[59] Feinberg, *Harm to Self: The Moral Limits of the Criminal Law*, 3:189.

We do not ground our moral analysis in any specific conception of manipulation. Rather, we will only aim to highlight two salient points. First, a point about method: the moral permissibility of manipulation depends on how manipulation is construed. In considering the disputes mentioned earlier about how to identify manipulation, for instance, it might seem reasonable to adopt a broad or disjunctive conception, under which *either* deceptive or coercive practices count as manipulative. However, this leaves unresolved what unifies the concept of manipulation, that is, what all instances of the phenomenon have in common. Moreover, since entirely avoiding both deception and coercion seems impossible in practice, this broad sort of conception does not justify categorical condemnation of manipulation.

Our second point is that manipulation is probably best understood in a non-moralized way. Instead of viewing manipulation as constitutively wrong, we should view it as identified by "objective facts about a situation that give us good reasons for condemning or approving certain things."[60] Insofar as reasons can be good yet nonetheless be undercut or outweighed, this view about the nature of manipulation is compatible with a moral evaluation of it as prima facie or pro tanto wrong – that is, as "not always wrong" but "generally wrong."[61] So, just as we can accept the defeasible badness of unfreedom, in light of examples where paternalism serves the public interest, here we accept the defeasibility of the badness of manipulation in light of examples where manipulation serves those interests. These examples might be few and far between, but they remain important. As we argue in the next chapter, the managers and developers of some technological systems might have obligations to influence their users in ways that are non-persuasive, deceptive, or perhaps even coercive. In these cases, such as in the case of the Click-Gap metric (discussed in the next chapter), the obligation not to manipulate simply gives way.

In general, we think that the problems with algorithmic systems are closely related to, but not reducible to, the affronts to autonomy wrought by manipulation. To this end, Marjolein Lanzing discusses the manipulative aspects of information technologies in terms of affronts to informational and decisional privacy. Yet, for her, as for us, the fundamental lesson is about affronts to autonomy, not manipulation per se: she writes that "[s]ince informational and decisional privacy protect autonomy, autonomy is under threat."[62] From this broader perspective, it is not only clear why manipulation is morally problematic, it is clear why manipulation is wrong to the extent that it engenders or reflects an affront to autonomy.

[60] Wood, "Coercion, Manipulation, Exploitation," 19–20.
[61] Baron, "The Mens Rea and Moral Status of Manipulation," 108.
[62] Lanzing, "'Strongly Recommended' Revisiting Decisional Privacy to Judge Hypernudging in Self-Tracking Technologies," 565.

5.5 CONCLUSION

Human agency and autonomy are, as we have seen, difficult to pin down. As we have argued, understanding these important ideas as they apply to actual people requires examining human behavior in depth. Here, we have conceptualized human agency in terms of three "natural challenges" to human agency, and we have argued that each of these challenges influences the nuances of a third important idea: freedom. Human freedom is fundamentally ecological: it is shaped by our capacity to act on our emotions, the limits of our cognition, and the centrality of our social relations to our decision-making.