What's Wrong with Automated Influence

Canadian Journal of Philosophy, Accepted 24 June 2021

Claire Benn[1]* and Seth Lazar[1]

[1]Australian National University
*Corresponding author. Email: cmabenn@gmail.com

Abstract: Automated Influence is the use of AI to collect, integrate and analyse people's data in order to deliver targeted interventions that shape their behaviour. We consider three central objections against Automated Influence, focusing on privacy, exploitation, and manipulation, showing in each case how a structural version of that objection has more purchase than its interactional counterpart. By rejecting the interactional focus of 'AI Ethics', in favour of a more structural, *political* philosophy of AI, we show that the real problem with Automated Influence is the crisis of legitimacy that it precipitates.

Keywords: artificial intelligence; surveillance; privacy; exploitation; manipulation; legitimacy; power

1. Introduction

After decades of slumber, the world is awaking to the extraordinary power that we have vested in the custodians of our digital infrastructure. 'Big Tech' is under attack from regulators worldwide seeking to wrest that power back. CEOs have been dragged (by video) before Congress; antitrust cases have launched; the GDPR is in force, the EU

Digital Services Act in preparation.[1] Even smaller countries like Australia have squared up.[2] Societies go through few such 'constitutional' moments—when we collectively recognise that we are subject to illegitimate power structures, and determine that they may not stand. Political philosophers should be well placed to help at these moments (think Hobbes, Paine, Rousseau). We can diagnose the moral flaws of existing power structures, and, using that diagnosis, recommend alternatives. And yet, political philosophy's engagement with this digital revolution is in its infancy. The normative analysis of our digital infrastructure has been led by other disciplines, in a tidal wave of critique known as the 'techlash', in which there is considerable normative agreement, and little sustained focus on unpacking the conceptual foundations of that agreement.

This should give us pause. We need to be sure the tsunami of critique is aimed at the right targets. And we need arguments for it that do not presuppose antecedent agreement. Most importantly, we need to know not only *that* some practice is morally objectionable, but *why* it is. Only then can we know *how* problematic it is, and so calibrate our concern to its seriousness, and craft positive proposals that address the root cause of our moral concern.

In this paper, we introduce and offer a moral diagnosis of one of the primary engines of our contemporary digital infrastructure: *Automated Influence*, the use of automated systems to collect and analyse user data, and then target interventions aimed at changing their behaviour. Ultimately the tech titans' power relies on their revenues, and those depend on Automated Influence, encompassing online behavioural advertising,

[1] Hearings to examine Section 230 immunity focusing on Big Tech (https://www.congress.gov/event/116th-congress/senate-event/328200); FTC sues Facebook for Illegal Monopolisation (https://www.ftc.gov/news-events/press-releases/2020/12/ftc-sues-facebook-illegal-monopolization); GDPR (https://gdpr.eu/); EU Digital Services Act (https://ec.europa.eu/digital-single-market/en/digital-services-act-package)

[2] Treasury Laws Amendment (News Media and Digital Platforms Mandatory Bargaining Code) Bill 2021 https://parlinfo.aph.gov.au/parlInfo/search/display/display.w3p;query=Id%3A%22legislation%2Fbillhome%2Fr6652%22.

recommender systems, and newsfeed and search algorithms. Automated Influence has also driven Artificial Intelligence (AI) research and development, whether finding new modalities for the exercise of influence (e.g. digital personal assistants operationalising advances in natural language processing) or optimising existing methods (e.g. tweaking a recommendation algorithm to increase user engagement) (Hao 2021). More perhaps than any other discrete practice of the leading digital platforms, Automated Influence has inspired popular concern, from *New York Times* editorials to Netflix documentaries (Zuboff 2019).[3]

In the moral critique of any social practice, we can adopt at least two broad perspectives, which we will call *interactional* and *structural*.[4] These are of course archetypes; most work includes some combination of the two. The interactional approach considers the interactions between agents that make up a social practice. It aims to identify adverse effects for individuals directly caused by those interactions. Its normative critique is grounded exclusively in the self-authenticating claims of persons with moral standing. A 'claim' is a fact about a person that can potentially ground pro tanto duties in others—that is, give others moral reasons that it can be wrong to breach. A self-authenticating claim is sufficient on its own to ground such duties.

The structural approach evaluates the emergent social structures of which those interactions are the leading edge.[5] It considers how those social structures directly and indirectly impact people's lives, as well as their relational properties, such as how they influence distributions of power, knowledge, and resources; and their aggregate

[3] The documentary is 'The Social Dilemma'.
[4] Political scientists and other social theorists commonly describe the interactional approach as fundamentally liberal, and its inadequacy as having to do with the perceived demise of some species of liberalism (Yeung 2017; Benthall and Goldenfein 2020).
[5] We understand social structures as the intended or unintended products of interaction among people in society, which reliably program for particular kinds of social and individual outcomes. We will focus primarily on formal and informal institutions, and on incentives (Jackson and Pettit 1990; Haslanger 2016; Fedoseev 2021).

effects—cumulative social impacts that are significant at scale, but relatively insignificant for each person affected. The structural approach can be motivated by showing how these structures have downstream impacts on people's self-authenticating claims. But it can also be motivated by these fundamentally relational goods (C. Taylor 1995; Waldron 1987; Griffin 2008).[6] Individuals do not have self-authenticating claims to a particular distribution of power, knowledge, or resources, or to one particular cumulative outcome over another.

Interactional critiques of social practices have a compelling kind of free-standing moral authority. One has instrumental reason to win others' support for one's cause, but the claims at stake are self-authenticating, so do not depend on that support. For example, think of abolitionists campaigning against slavery. Structural critiques that focus on relational and aggregate social goods are more deliberatively demanding. Since we do not have individual claims to social goods, we must collectively decide on the right path to take. Winning others' support for your cause is not just instrumentally important, it is constitutive of the value of your cause. Think, here, of campaigners for national self-determination of a cultural group (Margalit and Raz 1990).

The prevailing critique of Automated Influence, especially in public discourse but also in academic research, emphasises its interactional shortcomings. Although this lends normative clarity and motivational force—*you* should oppose Automated Influence, because it is undermining *your* self-authenticating claims—we think an exclusively interactional approach misses crucially important structural dimensions of the problem with Automated Influence. And this presents us with a more demanding challenge: to

[6] To be clear, a critique showing that target phenomenon X contingently precipitates a social structure Y, and that Y undermines people's self-authenticating claims to some good Z, is a structural critique because it focuses not on the interactions that are constitutive of X, but on the impacts of the social structure Y that X precipitates. For an example of this, see section 3.5.

decide how we want distributions of power, knowledge, and resources to be shaped by our digital infrastructure. That decision cannot be made by a 'moral vanguard'—it requires a genuine rethink of our social institutions writ large.

We begin by precisifying Automated Influence, then consider three central objections against it. In each case we show how a structural version of that objection adds something crucial to its interactional counterpart. Our paper therefore makes a case for political philosophers giving greater weight to structural arguments in their moral diagnoses of social phenomena. We recommend the emerging field of 'AI Ethics' turn away from its present interactional focus, and towards a more structural agenda: a genuinely *political* philosophy of data and AI.

2. Automated Influence

> Automated Influence: The use of Artificial Intelligence to collect, integrate and analyse people's data, and to deliver targeted interventions based on this analysis, intended to shape their behaviour for exogenous or endogenous ends.

Many first become concerned by Automated Influence through online behavioural advertising. An ad seems to follow one around the web; we begin to realise that we are being tracked online, and targeted accordingly. But online behavioural advertising is just the most explicit, and crudest, face of Automated Influence. Most of our digital services—from search, to social media, to online shopping—rely on user direction to secure our engagement and attention (and so show us more ads), as well as to help us navigate the functionally infinite space of our digital infrastructure, analysing our preferences to suggest complementary content, products, and services.

Automated Influence is driven by AI; but it has also driven epochal advances in AI.[7] The revenues generated by Automated Influence have sustained research and development in AI; the data gathered has made possible great leaps forward in computer vision, natural language processing, and other fields using machine learning (ML). Reciprocally, AI has also enabled a speed, scale, and personalisation of influence that would never have been possible without it.

Our definition highlights the role of AI in *collection*, *integration*, and *analysis* of user data, and its operationalisation by targeting a particular intervention.[8] We cannot morally assess Automated Influence without considering the pipeline of data that makes it possible: both to train predictive models and to target particular interventions.

Automated Influence makes it possible in principle to target behavioural interventions at an audience of one.[9] This targeting comes in two broad forms. First, matching people with products, services and content they may find appealing.[10] This means differentiating 'persuadables' from 'sure things', 'lost causes', and 'do not disturbs'—people whom targeting would actively put off. Second, tailoring the message to the individual, based on their inferred susceptibility to that method of persuasion (Calo 2014, 1018). Experimental results show the viability of such 'persuasion profiling', but there is little publicly available information about how widespread it is (Kaptein and Eckles 2012, 2010).[11]

These interventions aim to shape the user's behaviour—that is, they aim to raise the

[7] Interestingly, the newsfeed algorithm at Facebook is a descendant of their first algorithms for targeting advertisements (Hao 2021; Graepel et al. 2010).
[8] Of course, Automated Influence is not a wholly automated process, and can include more or less human involvement at each of these stages.
[9] For an excellent overview, see Turow and Draper 2012, 138.
[10] And indeed with particular 'price and feature packages' for those products, services and content: Cohen 2018, 229. See also Calo 2014, 123.
[11] For a recent review of some relevant empirical literature, see Susser and Grimaldi 2021.

probability they will ultimately take some particular course of action—in order to realise some goal. Behaviour is, minimally, a function of one's beliefs and desires given one's option set. Automated Influence can shape each element. Search and newsfeed algorithms shape what we believe; ads and recommender systems prompt and direct our desires; platforms make some options available and attractive, while hiding others. Each modality of influence can be either covert or explicit.

Automated Influence can shape user behaviour in their own (endogenous) interests, and/or in the (exogenous) interests of others. Typically the goal is to do both: to provide the user with a benefit while also extracting profit for the influencer—for example, hold the user's attention on the platform, in order to serve them more ads.

Presented in this light, Automated Influence does have a benign face, and may to some extent be necessary. The internet is as good as infinite; without some means to navigate it, we would be lost. Automated Influence enables us to discover relevant products, services, and content. Developing the infrastructure of Automated Influence requires significant investment; that investment is possible because tech companies optimise for profit as well as for user functionality.

But there is a malign face too. Critics of Automated Influence argue that it relies on invasive inferences from data that is illicitly acquired, thereby delivering excessively targeted interventions that covertly shape people's beliefs, desires, and behaviour, for exogenous ends. From this general anxiety we extract three objections to Automated Influence, focusing on privacy, exploitation and manipulation. We discuss each in turn.

3. Privacy[12]

We'll call data collected to train predictive models 'training data', and that used for targeting 'targeting data'. We also distinguish between sensitive and non-sensitive data points. 'Sensitivity' is a functional term, intended to identify data about a person which that person might reasonably not want others to know.[13] Our key distinction is between data that is intrinsically and extrinsically sensitive. A data point is intrinsically sensitive if it is sensitive when considered on its own—that is, if you would reasonably not want others to know that data point alone. It is extrinsically sensitive if it is sensitive only when considered in combination with other data points.

This is the basic paradigm of Automated Influence. An 'influencer' has training data including sensitive and non-sensitive information about a population. They train a model on that data, revealing a link between intrinsically non-sensitive properties P, Q and R, and intrinsically sensitive property S, such that if [P, Q, R] obtain for a user, the probability of S obtaining increases (Barocas and Nissenbaum 2014, 55). Suppose P, Q and R have to do with the user's music, podcast, and browsing patterns, while S is their sexuality, for example. The model is then applied to a user who has revealed P, Q, and R, but not S; this enables the influencer to infer that S likely obtains, and target the user with interventions aimed at S-people.

[12] The sociolegal study of privacy is a vast field, to which we cannot hope to do justice in this paper, but which has largely developed independently of the debate on privacy among philosophers (to the detriment of the latter). We think the following are particularly illuminating entry-points for philosophers: Barocas and Nissenbaum 2014; Cohen 2000, 1396; 2018, 220; Solove 2004. In the policy context, see also Hildebrandt and Gutwirth 2008. For a philosophical approach that engages with (and adds to) that literature, see Véliz 2020. Arguably the shift from privacy law to data protection regulations captures the essence of our concern with the shift from interactional to structural approaches, however the public justification of data protection regulations like the GDPR does often seem to assume an interactional/individualist justification. Thanks to Jake Goldenfein for his helpful discussion with respect to this literature.

[13] What one wants others to know depends on the context; we distinguish, however, between sensitivity due to context (you don't want your boss to know something you allow your partner to know), and extrinsic sensitivity (you don't mind if someone knows P on its own, but you don't want them to know P, Q and R given that they together entail S) (Nissenbaum 2004). Thanks to Selim Berker.

3.1 Control of Data About You

In the public discourse on Automated Influence, a prominent objection claims that using people's data in this way undermines their privacy.[14] More specifically: influencers have no right to use people's data to train their predictive models; and it is wrong to make invasive inferences about people's sensitive information.

This objection can be developed in interactional or structural terms. We start with the interactional approach. This is most compelling if we can identify an underived self-authenticating claim against our privacy being undermined by Automated Influence. One can also argue for a derived claim, grounded in privacy's utility in protecting other interests—such as in not being exploited or manipulated—but since that argument is really grounded in those other interests, we return to it below.

The internet's first decades have seen many egregious invasions of individual privacy, on any reasonable interpretation.[15] However, these are now widely acknowledged as being obviously wrong, so we set them aside to focus on practices that are central to the ongoing business model of Automated Influence.

We are sceptical about grounding the critique of Automated Influence on its undermining an underived self-authenticating claim to privacy, because we think that you do not have a weighty underived claim to unilaterally control your intrinsically non-sensitive behavioural data, because that data is the product of your interaction with a digital infrastructure, over which the creators of that digital infrastructure also have some antecedent claim.[16] This behavioural data is about you. But it is also about the

[14] See e.g. https://www.nytimes.com/interactive/2019/opinion/internet-privacy-project.html, Véliz 2020, 33.
[15] For extensive catalogues, see Zuboff 2019; Véliz 2020.
[16] For a similar view (reached independently) see Benthall and Goldenfein 2021. Note that covert third-party trackers have no such claim to access this information. However, typically such trackers operate by agreement with the digital service provider, so the real question remains whether they are entitled to provide others with access to your data when you use their service.

site you have navigated, and the service you have used. You have some claim over it, to be sure. But so does the service provider.

There has long been a struggle over who should control people's 'data exhaust', or 'behavioural surplus' (the very terminology is the site of this struggle) (Zuboff 2019). The conventional wisdom now is that this is *your* data—the user has unilateral rights over it (Véliz 2020). While we might endorse this as the conclusion of a political argument, grounded in aggregate, relational, and structural considerations, we deny it as an underived moral premise in a critique of Automated Influence.[17] For you to have a natural, underived claim to unilateral authority over some data point, it should be either intrinsically sensitive, or else you should otherwise have some kind of special claim to it—for example, perhaps, because you unilaterally generated it (think of intellectual property as an example). If you make something together with another person or organisation, then both you and that organisation have some natural claim to control the fruits of your joint labour. If it is not intrinsically sensitive, the mere fact that a data point so generated is *about you* is not sufficient to give you unilateral authority over it.

One could counter, here, that it's a mistake to place too much weight on whether the data point is *intrinsically* sensitive. If S is a sensitive attribute, and knowledge of [P, Q, R] raises the probability of S, then can *that* ground a claim to unilateral control over P, Q and R?

We think this argument is worth exploring. We can develop it in at least two ways. First, one might argue that you have a claim to unilateral control over P, Q and R, just in case they are *necessary* to infer that the probability S obtains is above some threshold. Or,

[17] To her credit, Véliz emphasises just this kind of collectivist political argument.

second, the claim could be grounded in P, Q and R being *sufficient* to make that inference.

The first approach seems unlikely to generate robust privacy protections. The redundant encoding of sensitive attributes in large datasets typically means that many different subsets of the data enable the same inference, so no particular subset is necessary. As a result, on this view we would have limited if any rights to control the data that enables sensitive predictions.

The second approach is worth exploring in more depth. P, Q and R entail a higher probability of S only given that the model has also been trained on data about many other people. If data point X being part of a set of data points that are jointly sufficient for S to be inferred grounded a claim to your having unilateral control over X, then you would have a claim to unilateral control over data about others, which you do not. After all, the set of data sufficient for S to be inferred about you will also include data that is part of a set that is sufficient for S to be inferred about many other people, and you cannot all have unilateral control over the same data points.

Could we then supplement the sufficiency approach, by arguing that if X is about you, *and* is part of a dataset that is sufficient for a high probability of S to be inferred about you, then you have a claim to unilateral control over X? However, we think this is likely to be overly inclusive—it is hard to imagine a piece of intrinsically non-sensitive information about you that is *not* part of a set that is sufficient for making sensitive inferences. On this approach, you would have a right to unilateral control over literally every data point that is about you. But much of the data that is about you is also about other people—it is relational data, such as that A and B are spouses, or that A and B were communicating together on a messaging platform (Salome Viljoen 2020).

Probably the best version of this argument, then, would say that you have a right to unilateral control over any data point that is *exclusively* about you, that is part of a set that is sufficient for inferring a high probability of some sensitive attribute S about you. This raises some interesting questions, which we cannot settle here, about what it takes for a data point to be exclusively about one person. And as we noted above, data that you generate by using some digital service is not *exclusively* about you. It is also about that digital service. We therefore think this argument is likely to be significantly underinclusive—though we think it deserves further consideration.

3.2 Control over Inferences

Rather than appeal to our claim to unilaterally control P, Q and R just because they enable an inference to S, one might instead simply argue that others who licitly know P, Q and R should not infer S from it. Although there are instrumental reasons to prohibit such inferences in particular cases, we deny an underived claim that others not put two and two together. There can be nothing wrong (we think) with the mere fact of making a warranted inference.

Objection: does our scepticism derive from irrelevant assumptions about human psychology? We generally lack a claim that others make inferences from what they licitly know, because we could never prevent such inferences in practice, and even if we could, it would egregiously constrain their freedom of thought. We can, however, easily prevent people from using predictive models, and doing so does not obviously undermine their freedom of thought.

However, we think that if there is a basic objection to drawing inferences by predictive

models, then it should also be at least somewhat wrong to infer S when you licitly know P, Q and R. But we think it cannot be. Identifying patterns and making inferences from licitly acquired data is not in itself wrongful. *Acting on* those inferences might be wrongful because of the consequences of doing so. But that is a separate matter.

3.3 The Role of Consent

We are somewhat sceptical about the force of appealing to individual privacy to ground opposition to Automated Influence. But suppose we *could* show either that you have a self-authenticating claim that others not make certain inferences from what they licitly know, or that you have a claim to unilateral control over any data that is exclusively about you, and that is part of a set that is sufficient for inferring some sensitive attribute (and that the set of data exclusively about you is meaningful and large). Even then, we presumably would not think that either of those claims were inalienable. If you *want* to let companies know P, Q and R, even knowing that this will enable them to infer S, then there are seemingly few interactional grounds for denying you the right to do so. It is therefore unsurprising that consent looms so large in discussions of individual privacy and Automated Influence.

We can use consent to criticise Automated Influence on the grounds either that it involves breaching actual agreements between users and digital service providers, or that the agreements that license it are themselves invalid. We set aside the former objection—there is no mystery about why breach of contract is wrong. The second objection has more promise, and over the last two decades, scholars have exhaustively demonstrated the inadequacy of individual consent to legitimate the collection and use of individuals' behavioural data in the era of ML (Barocas and Nissenbaum 2014).

Instead of revisiting these arguments, we will argue that the best reasons for thinking these contracts invalid refer to the structural, aggregate effects of managing behavioural data by individual consent.

A predictive model does not need to access *everyone's* data to make reliable predictions. Its training data could be a sample as small as 20% of the whole (Barocas and Nissenbaum 2014, 62). It can then make sensitive predictions based only on targeting data, which can be significantly less comprehensive than training data, and indeed can include only the data that you cannot avoid sharing in order to use a digital service, such as your hardware and browser metadata, and your IP address.[18] In these cases, your only hope for avoiding having sensitive inferences made about you is to avoid using the digital service entirely.

Assume that consent in the absence of a reasonable alternative is not morally effective (that is, does not change what others are permitted to do [Wertheimer 1987]). How then should we assess the consent to share behavioural data with a digital service provider, in light of these externalities? You have three options: A, use the service and share behavioural data that can be used to train a predictive model, perhaps with some modest incentive to do so; B, use the service, share only the minimal targeting data that you cannot avoid sharing; C, do not use the service at all. Suppose that if 1 in 5 people choose A, then there is little to no difference between the inferences that can be made about you, whether you choose A or B. In that case, you gain no real advantage by choosing B, and you miss out on the incentive to choose A. So, if enough people choose A, then B is no longer a reasonable alternative to it. Everything then depends on whether

[18] Even if you use a VPN, you can still be identified to an alarming degree of precision just by your browser metadata: see https://coveryourtracks.eff.org.

C, not using the service at all is a reasonable alternative to A.

In the present digital environment, we think that option B is equivalent to using the new (putatively) privacy-preserving digital services, which have been launched in response to growing concern about Automated Influence. Many users try to protect their privacy by using a Virtual Private Network, searching on sites like DuckDuckGo, browsing on Safari, or deleting their Facebook accounts, to prevent some kinds of cross-site tracking. Almost invariably these privacy-preserving techniques impose some cost on the user (most privacy-preserving search engines license Bing's search results—try using those for a week). And the reality is that, *given the choices of others to use the more popular, more invasive services*, your privacy-preserving choices make little to no difference to the ability of online advertisers to profile you, and target you with advertisements (and other interventions). Hence the only meaningful choice is between not using the internet at all, and submitting to being profiled and targeted. Given how many of us are dependent on the internet for our professional and personal lives, this is not the kind of choice that can generate morally effective consent.[19]

The obvious alternative to the lens of individual consent—as has been recognised by privacy scholars for some time, and with particular force in a forthcoming paper by Salome Viljoen (Salome Viljoen 2020)—is that we must instead work out a collective approach to allocating and using behavioural data.[20] We think this is the right answer—but it entails focusing on the relational and aggregate effects of the data practices of Automated Influence, rather than considering individuals' claims to privacy first and foremost. Privacy claims, on this view, are the product of a negotiation over how we, as

[19] This, incidentally, helps to explain the 'privacy paradox', i.e. the thesis that people profess to value privacy seriously, but are wiling to trade it for relatively trivial benefits (Acquisti, John, and Loewenstein 2013; Boerman, Kruikemeier, and Zuiderveen Borgesius 2017, 372).
[20] For other recent collective approaches to privacy, see Véliz 2020; L. Taylor, Floridi, and van der Sloot 2017.

societies, should govern the flow of data, rather than a crucial input into those negotiations.

There is a further problem with grounding our critique of Automated Influence in individuals' privacy claims, and so in our practices of notice and consent. For there is a way to improve those practices, and make them much more tractable for users. But it may involve centralising authority in a few trusted platforms, which then automatically manage the user's preferences with respect to third parties.[21] The larger platforms have long recognised the opportunity in taking charge of the enforcement of privacy norms online (Clark 2021). But while they constrain third parties' access to users' behavioural data, their own access is practically unconstrained. And while they might solve one problem with Automated Influence, they do so by exacerbating another—the concentration of power in too few unaccountable hands.

3.4 A Structural Approach

Instead of focusing on individuals' voluntary decisions whether to share their data with digital service providers, we need to emphasise the aggregate effects of the broader institutions of data governance. This shifts us from an interactional perspective to a structural perspective. Continuing in the same vein: the problem with Automated Influence is not just that automated systems access and make inferences from intrinsically non-sensitive behavioural data, but that they create standing economic incentives to turn *everything* into behavioural data, steering us ever closer to ubiquitous

[21] An alternative would double down on the decentralised approach of the internet, for example using blockchain, but this raises serious privacy and feasibility concerns which we lack space to address here.

surveillance.[22] Instead of having just our online behaviour recorded, we increasingly find it impossible to escape being continually recorded, wherever we are. What's more, we are often complicit in this mass mutual surveillance, wilfully filling our lives with devices that record both ourselves and others.

But what is actually wrong with ubiquitous surveillance? We think it encroaches on the basic, self-authenticating claim to have some significant space free from being observed, and on the social good of living in free and equal societies.

3.5 Surveillance and Sovereignty

We can readily imagine a world with Automated Influence, but without ubiquitous surveillance. However, in the actual world, Automated Influence creates a standing economic incentive to turn everything into behavioural data, so that it can be used to target advertisements, products, services and content. We have both interactional and structural reasons for objecting to ubiquitous surveillance, but invoking ubiquitous surveillance contributes to the structural critique of Automated Influence, because only by attending to the social structures enabled by Automated Influence can we see its contingent downstream impacts on other aspects of our lives. An interactional approach that focused on Automated Influence, without attending to these social consequences, would not hold it accountable for the ubiquitous surveillance that it incentivises.

Our first objection to ubiquitous surveillance is grounded in our sovereignty over our

[22] 'There is no logical endpoint to the amount of data required by such systems—no clear point at which marketers or the police can draw the line and say no more information is needed. All information is potentially relevant because it helps reveal patterns and correlations.' (Andrejevic 2012, 94). See also Pridmore 2012, 323.

own persons and our claim to a reasonable sphere of action free from observation by others. We need not take advantage of this sphere if we choose not to. But the basic licence to retreat from the gaze of others is as fundamental to our sovereignty over our persons as is our similar authority over our bodies.

Suppose we could achieve some non-trivial benefit for others by cutting off some of your hair while you are asleep, without your ever knowing it had happened. Even though you would never knowingly be affected, and the objective effect would be trivial, it is still wrong to do this without your consent—it's your body, and you are sovereign over it (Quinn 1989). To be sovereign over your person, you must have a morally authorised sphere of freedom in which you are at liberty to decide what to do, without penalty or censure (Lazar 2019).

Just as you are entitled (to a point) to refuse others the use of your person for the sake of fulfilling overarching goals, you are also entitled (to a point) to refuse them the observation of your person. For this to be possible, you must be able to withdraw from others' gaze without undue penalty. Increasingly ubiquitous surveillance raises the costs of withdrawing, since it shrinks our sphere of freedom. So it undermines your capacity to be sovereign over your own person.

Much rests, here, on the idea of 'observation'. Some think that one's basic interest in privacy is activated only when data about one is *accessed* by others, so that merely being recorded is not sufficient to set back that interest (Macnish 2020). We think that you lack sovereignty over your person if some other person or group is *able to* observe you without adequate limitation.[23] This means that the problem is not merely that we

[23] By 'able to' we mean, roughly, 'sufficient probability of success conditional on trying' (Southwood and Brennan 2007).

are always susceptible to being recorded by different devices, but that it is possible to integrate those different streams in order to build a comprehensive picture of each person. If your whole life (or close enough) *could* be observed by some other person or group, should they choose to, then you are not properly sovereign over it. If you were recorded every waking moment of your life, but it was impossible to integrate those recordings, then your sovereignty over your person would be less seriously contravened, since no other person or group would be able to surveil your every moment—each would have only a snapshot.

3.6 Surveillance, Freedom, and Equality

The next two arguments focus on structural, relational social goods: the value of living in societies that are *free* and *equal*. This value is not simply reducible to the instrumental benefit for each person of society being free and equal: free and equal societies are good in themselves, over and above how they contribute to the well-being of each person.[24]

Ubiquitous surveillance, together with the power of the modern state, makes for an unfree society. This point is often made, so we will not dwell on it at length.[25] States face many different challenges, real and imagined, and granular data about each of their citizens' behaviour can help solve some of those challenges. So our behavioural data exerts an irresistible pull on state authorities. For most of us, this comes to nothing. However, some have their basic privacy rights invaded, but never know it. Some suffer

[24] Compare Lazar 2010. For the welfarists sceptical of such impersonal values: on an extended understanding of welfare, we could describe these goods as 'public interests'. For those who think that welfare includes only self-regarding interests, but who also deny the existence of impersonal values, we have no response. For useful context, see C. Taylor 1995; Waldron 1987.

[25] See e.g. Richards 2013, 1941 for a discussion of the interplay between commercial and state surveillance, and Andrejevic 2012 for a prescient account of the rise of ubiquitous surveillance. For a brilliant description of how ubiquitous surveillance 'supercharges' the state, see Susskind 2018.

the direct consequences of the mistaken or unjust exercise of state power, supercharged by big data and AI, and lose their freedom. This is especially likely for those who lack the full protections of citizenship (for example, undocumented migrants in the US [Bedoya 2020]).

But the broader problem, independent of precisely who ends up suffering these severe incursions into their privacy and their freedom, is that a society in which we can be surveilled in this way by state authorities is one in which we are all unfree. Automated Influence provides the economic case for launching product after product that records our online and offline behaviour; these products are either expressly and transparently repurposed for state use (for example, Ring doorbell cameras transmitting data to police forces), or else surreptitiously accessed through back doors, or through ISPs (Harwell 2019). If democratic states tried to install this kind of surveillance equipment so pervasively, there would be massive uproar. Instead we are installing these gadgets ourselves.[26]

The obvious solution here would be to ensure that our behavioural data is genuinely secure against all third parties, including the state, by preventing it from being aggregated at all, keeping it on encrypted devices, or else aggregating only after encryption has been applied. However, this again ends up putting a lot of power in the hands of tech companies, which still have access to identified data, and which are in this scenario entrusted with protecting our data against the might of the state. As we will discuss in more detail below, in some ways the problem is that digital technologies enable *too much* power, making the challenge of identifying a legitimate authority still

[26] Besides the Snowden revelations, in 2021 we learned that many states were buying 'Pegasus' software, from Israeli company NSO, and using it to turn journalists and political activists' smart devices into remote cameras, microphones, location trackers, and so on. See https://www.theguardian.com/news/series/pegasus-project.

more daunting.

Final argument: ubiquitous surveillance threatens equality as well as freedom. Those who can access a comprehensive picture of our online and offline behaviour have undue power over us. This obviously undermines our freedom, but also places us in unequal social relations. Consider the Uber founder and one-time CEO's 'party trick' of turning on 'God View', a display revealing the location of everyone using an Uber (troubling enough in itself, but all the more so when explicitly used to track individuals [Hill 2014; Véliz 2020, 37]). They call it God View because it gives them a supernatural level of insight about and power over mere mortals like us. A society in which some people can have this kind of access to the behavioural data of others is to this extent and for this reason unequal (it may also be unequal for many other reasons, of course).

The central problem here is that contemporary computing power and data management and analysis capabilities enable us to integrate vast amounts of disaggregated data into a coherent whole. A mishmash of different devices—smart TVs, smart speakers, smart doorbells, smartphones—can be integrated into a single effective network for realising some objective. The net is not created at once, and then thrown over us all, so that we can see it coming and resist. Instead we are each stitching our own little piece of it, and data management companies like Palantir are drawing it all together.

This is a general feature of the political problems raised by big data and AI, and of the central contribution that they can make to society: seemingly disconnected and ineffectual individual elements come together in the aggregate to realise something astonishingly powerful. One net result is that some people are placed in an extraordinarily asymmetrical position relative to others: we each know only our piece of

the patchwork; they have a view of the whole. For most of us this makes little practical difference. The data gathered to facilitate Automated Influence will only ever be used for that purpose. But we now live in a society where some people are subject to unjust or mistaken intervention on the basis of this data, and in which some people have access to awesome power. We live in a society that is pro tanto less free and equal than it would be without the ubiquitous surveillance that Automated Influence has incentivised.

4. Exploitation

The privacy-based argument against Automated Influence is most compelling if it is either developed into a structural critique of the social relations enabled by big data, or else pinned on the downstream implications of affording people inadequate control over their behavioural data: for example, that without this control, we will be subject to exploitation by digital service providers.[27] We think the interactional version of this argument is insufficient. Individual users do not in general have a strong complaint that they are being exploited by influencers. But when we consider users as a group, and influencers as a group, and when we consider the overarching infrastructure of Automated Influence rather than individual interactions, the argument becomes more compelling.

We adopt the following (stipulative) understanding of exploitation. Exploitation occurs when one party to an ostensibly voluntary agreement intentionally takes advantage of a relevant and significant asymmetry of knowledge, power or resources to offer the

[27] This argument is made throughout Zuboff 2019. See also Cohen 2000, 1390; 2018, 223, and Noble 2018, 36.

other party terms of exchange to which they agree, but would never accept were they more symmetrically situated in that respect.[28] Applied to Automated Influence, this would imply that the apparently voluntary agreement to share our behavioural data for access to digital services is made against a backdrop of a significant asymmetry of power, resources, or knowledge, and that we would reject these terms if we had a stronger bargaining position.

4.1 Unfavourable Transactions

As with the argument from privacy, we concede that many internet users have been gulled into deeply unfavourable transactions that they would never have accepted had they known what was really at stake. More than this, many data companies and Automated Influencers have simply deceived their users, using subterfuge to acquire data that was never intended to be shared. The actual practice of Automated Influence has been riddled with this kind of naked corruption. Individuals have a clear complaint against these corrupt practices.[29] However, even when these are set aside, some have argued that Automated Influence is still objectionably exploitative. Let's look at why.

The purveyors of Automated Influence have indeed made a tidy profit from it. Advertising has proved extraordinarily lucrative. Even companies whose traditional profit centres were in software or retail have recently seen more and more profits come

[28] Does our view of exploitation imply that workers are typically exploited by their employers? After all, there is always an asymmetry of power between the firm and the person they hire. However, this asymmetry is relevant only if it significantly affects the bargaining position of the two parties; Google is obviously much more powerful than any particular engineer they might hire, but if the engineer has adequate alternative options to working with Google, then that asymmetry is not relevant to this transaction. There is an asymmetry of power *in general* between them, but with respect to *this transaction*, they are symmetrically positioned. Our thanks to Stephen Campbell for pressing this question.

[29] indeed there has been pressure on the principals purveyors of Automated Influence to condemn and foreswear them, which has had some effect at least in public declarations if not in practice, see https://privacy.google.com/.

from this one stream (Graham 2021).[30] And users are severely asymmetrically positioned relative to the major digital service providers. Their level of power, and their knowledge of user behaviour, are jointly extraordinary. Ours, not so much (Calo and Rosenblat 2017).

And yet the case that individual users are exploited by these practices rests on a weak foundations. For a start, the argument presupposes that each party to the transaction has a right to unilateral control over what they are exchanging. As argued above, that is contentious for our behavioural data. It is generated by our use of the digital infrastructure, so is part of the cooperative surplus that we must agree to divide, rather than something of ours that we bring to the bargaining table.

Even so, the division of that surplus could be an unfair one, to which we agree only because of a radical asymmetry in our respective bargaining positions. Users typically believe their behavioural data trivial, while Influencers know that, with enough data to train their predictive models, they can reap significant benefits. One might compare them to an unscrupulous art collector, who knowingly buys a priceless masterpiece for a song from its ignorant owner. This would be a kind of exploitation—taking advantage of the other's ignorance. But it is not an accurate analogy here, because any given individual's data is effectively worthless.[31] Predictive models depend on massive datasets; the marginal individual is a drop in the ocean. A better analogy would be if millions of us each owned a piece of a priceless jigsaw puzzle, but all of the pieces are multiply duplicated, and assembling the puzzle requires tremendous investment and ingenuity. The art collector buys up a full set, without explaining their composite value

[30] Even companies like Apple that derive considerable revenues from hardware also rely on their platform to attract users to that hardware, and their platform is defined and structured by Automated Influence.

[31] In this sense analogies between data and oil are mistaken: like oil, data needs to be extracted and refined, and can then be used in multiple ways; but even a bucket of oil is valuable, whereas one person's data is worthless on its own (Srnicek 2017, 40).

to any of the sellers, but then has to invest considerable resources in assembling it. This does not seem so obviously exploitative.

The insights (and profits) generated by behavioural data require considerable investment and ingenuity to extract, and any individual's contribution to the end result is typically trivial. In that light, being paid for our data with free access to digital services does not seem to be exploitative. It also has a progressive cast: providing digital services free at the point of use enables everyone to take advantage of them, rather than keeping out those with less disposable income.

An interactional, individualist version of the exploitation objection seems at best incomplete. But when we focus not on individuals, but on communities, and not on individual interactions but on the broader infrastructure of Automated Influence, the picture is different.

4.2 Dividing the Cooperative Surplus

One of political philosophy's central questions is how we should distribute the productive surplus made possible by cooperation with one another in society (Rawls 1999). The cooperative surplus generated through our use of the new digital infrastructure has been divided to give digital service providers a disproportionate share of the benefit, and more importantly a disproportionate share of the power. They get to decide how our digital cooperative surplus is distributed, and what to do with it. In an adaptation of Julie Cohen's phrase, we have allowed them, to our detriment, to have

unilateral control over the 'means of prediction' (Cohen 2000, 1406).[32] We think this can provide the basis for a more structural, collectivist version of the objection from exploitation. Let's go through this in more detail, focusing in turn on the new resources, knowledge, and power enabled by this cooperative surplus.

At the crudest level, this is about *resources* (Cohen 2018, 216). Each individual's data is near value-less. But in the aggregate, it is an extraordinary resource that has generated untold wealth for the most prominent tech companies, their owners and employees. Though individually irrelevant, we are together essential for the creation of this collective surplus. But because we do not control the means of prediction, access to digital services is our primary return. We can redress the balance by taxing these companies and imposing other imposts on them. But they are extremely adept at avoiding those costs, and even at mobilising the public to resist their imposition.[33]

More importantly, our cooperative behavioural surplus enables new kinds of *knowledge*. Even if the primary economic motivation for data collection and analysis is to facilitate the personalised delivery of products, services and content, massive datasets have extraordinary 'latent energy', and can generate insights on many different topics of social importance, as well as providing training data to enable vast leaps forward in AI. These insights and advances are accessible to those who control the means of prediction, but the rest of us, including our democratically elected representatives, are locked out. We cannot even know how effective Automated Influence itself is; we cannot gain first-hand knowledge of the functioning of the different recommendation

[32] Cohen actually talks about 'modes of prediction', but for our purposes we think that 'means of prediction' is a more apt adaptation of the Marxist terminology for our purposes. We might describe contemporary informational capitalism, and its use of datafied means of prediction, as a new mode of production.

[33] Compare the recent showdown between Google, Facebook and the Australian government over diverting advertising income back to traditional media companies. See also Culpepper and Thelen 2020.

algorithms that structure our online experiences. And we cannot decide the research agenda for how to use our cooperative surplus to generate insights about our offline lives that could play a vital role in improving public policy. For example, consider the COVID-19 pandemic. Tech companies have access to location and interaction information that could be invaluable to understanding specific transmission scenarios and broader trends, but governments of democratic polities are locked out of that information except on the companies' terms. The decision how to weigh values like privacy and public health is taken not by democratically elected officials, but by the executives of Apple and Google (Lazar and Sheel 2020).

This brings us to the deeper and more persistent problem. Our cooperative behavioural surplus enables new kinds and distributions of *power*. The tech companies' control of the means of prediction means that we can only indirectly infer the extent of that power. And as yet we have no viable way of legitimating these new power relations. In the previous section we discussed the power over individuals made possible by ubiquitous surveillance. Here we need to consider the power over populations enabled by the insights that can be generated by the means of prediction applied to aggregated user data. We have put all this power in the hands of tech companies, leaving to them the decision of how to use this data, what to try to gain insight into.

Maintaining the means of prediction, and the broader infrastructure of Automated Influence, requires digital platforms. Any constraints on the collection or use of data have to be implemented within those platforms. And in practice, the complexity and sheer volume of interaction on those platforms mean that they largely police themselves (consider the example of copyright enforcement [Suzor 2019]). But where does their authority to do so come from? What procedural standards should they

observe? Can we ensure that they will implement duly authorised laws, and won't oversimplify them in order to reduce the cost of enforcement (Suzor 2019)?

4.3 Refusal and Resistance

At the same time as collectively generating this new cooperative surplus of resources, knowledge and power, the systems of Automated Influence and the companies purveying it have worked to atomise individual consumers, reinforcing in us the mindset of individual choice and consent, and fragmenting our shared epistemic landscape (Salomé Viljoen, Goldenfein, and McGuigan 2020, 7). This is one of the great ironies of Automated Influence: it depends on an infrastructure that derives from a species of *unthinking* collective action, but which then enables a kind of personalisation, and an ideology of individualism, that fragments us such that we become worse at engaging in *considered* collective action to undertake collective bargaining with the tech companies.

This has three steps: two epistemic, one ideological. First, just as Automated Influence affords influencers unprecedented insight into our lives, their control of the means of prediction prevents us from seeing and understanding just how they are governing the digital infrastructure they have created, and the extent of the insights and influence our cooperative surplus can create.

Next, Automated Influence delivers us each a personalised experience of the internet, in which we see content tailored for our interests. As we become increasingly dependent on our digital infrastructure to inform our worldview, we are subjected to an increasingly fractured epistemic landscape, which militates against coordinated

collective action to wrest unilateral control of the means of prediction away from the tech companies.[34]

The last step is ideological. Tech companies have extensively promulgated the idea of individual agency and choice, framing our experience of our digital infrastructure so that we consider ourselves atomised individuals negotiating only on our own behalf.[35] It is their solution to every objection raised against Automated Influence, because it ensures that the only collective action we engage in works to their benefit; beyond generating the cooperative surplus, we leave everything else to them. The sense that we must navigate all the shortcomings of our digital lives alone is deeply disempowering to many of us; a sense of 'digital resignation' leaves us simply agreeing to various disclosures so that we don't have to spend our whole lives online policing the boundaries of our rights (Draper and Turow 2019, 1829).

4.4 The Exploitation Objection Restated

When we view Automated Influence through this lens, focusing on the social structures that we have collectively allowed to emerge over the last twenty years, rather than on individual transactions between users and tech companies, the argument from exploitation looks much more plausible.

The relevant transaction is between us, the users of the internet on the one hand, and the Automated Influencers on the other. We are exchanging our data—individually of

[34] Although some researchers are sceptical of the existing of 'filter bubbles' per se, there's surely little doubt that social media in particular is facilitating the spread of misinformation, and contributing to the formation of extreme interest groups with no interest in social compromise (Bruns 2019).

[35] '[W]hile individuals recognize risks to their information privacy, they also describe a lack of power over the situation. They define privacy cynicism as "an attitude of uncertainty, powerlessness and mistrust towards the handling of personal data by online services, rendering privacy protection behavior subjectively futile"… [C]ompanies, including online advertisers, benefit from learned helplessness insofar as people tend not to dramatically alter their behaviors when they learn about unwelcome data practices.' (Draper and Turow 2019, 1828).

little value, but precious in the aggregate—for access to digital services. And while our data is generated through interaction with a digital infrastructure that we did not create, at issue here is not only entitlement to proceeds from particular interactions, but how to divide the cooperative surplus of resources, knowledge and power that our data collectively makes possible. And as self-determining political communities we *do* have robust presumptive rights to set the terms for how that cooperative surplus is distributed.

The relevant asymmetry is between, on the one hand, the tech companies' understanding of the value of that data and their ability to act in a coordinated and purposeful way, and, on the other hand, our general ignorance of the aggregate value of our data, and our inability to act in a coordinated and purposeful way. We have therefore, by accident and without coordination, in effect collectively accepted terms of exchange that give the tech companies near unilateral control over the means of prediction; if we were better coordinated we should certainly demand more control and a greater share of the cooperative surplus of resources, knowledge and power. Worse still, the tech companies have used the very tools to which we have given them access to exacerbate the asymmetry between them and us, by using the methods of Automated Influence to further undermine our ability to coordinate, nudging us towards atomised individual decision-making by promoting an ideology of individual agency and control, while also fragmenting the shared epistemic foundations for collective action.

5. Manipulation

Recent years have seen a groundswell of opposition to Automated Influence, from bestselling books and Netflix documentaries to resolutions in the European parliament

(Zuboff 2019; Lomas 2020).[36] People are increasingly concluding that Automated Influence is undermining our autonomy: that we are all subject to 'remote control' (Zuboff 2020). This objection deserves serious consideration; if Automated Influence were inherently manipulative, then that might be reason enough to reform or reject it.[37] When thinking through this objection, however, we again think that considering only the individual manipulatory effects of Automated Influence does not adequately convey the seriousness of what is at stake. For a comprehensive picture, we must adopt a more structural and collective approach.

5.1 A Sufficient Condition for Manipulation

We start by offering a sufficient condition for manipulation. Manipulation involves (though may not be exhausted by) undermining an individual's decision-making power—for example preying on their emotions, their momentary whims, or their reliance on cognitive biases and heuristics—in order to change their behaviour.[38] Their 'decision-making power' is, roughly, their ability to select among their options, given their beliefs about the world, in ways that advance their goals. Some contend that only covert influence counts as manipulation; we deny this.[39] While manipulation can proceed by concealment or deception—for example, when casinos manipulate people to stay longer than they might otherwise intend, by not having any visible clocks in their gaming rooms—many of our cognitive shortcomings are equally decisive even when we know

[36]. The Netflix documentary 'The Social Dilemma' encapsulated some of this argument too. For further presentations of similar arguments see Becker 2019; Vold and Whittlestone 2019; Calo 2014, 999.

[37] As above, we think it is more productive not to focus on the most egregious cases of wrongful manipulation on the internet, because they are widely condemned even by the leading purveyors of Automated Influence, and our task is to assess the prevailing practices of Automated Influence, rather than to call out obvious outliers. On those outliers, see Susser, Roessler, and Nissenbaum 2018.

[38] This sufficient condition is inspired by Susser, Roessler, and Nissenbaum 2018.

[39] Here we depart from Susser, Roessler, and Nissenbaum 2018.

they are in play, so one can manipulate another entirely transparently.

The wrong of manipulation has two sides. First, it involves effectively suborning the will of others, and as such it undermines their autonomy. Second, it involves the manipulator placing themselves above the manipulated, treating the manipulated as a subordinate, one whose will can be suborned. This is an objectionable species of disrespect, and an affront to egalitarian social relations.

5.2 Tailoring the Message, Targeting the Product

Are the methods of Automated Influence manipulative? Let's start with online behavioural advertising. This involves two salient species of Automated Influence: tailoring the message, and targeting the product. Tailoring the message can certainly appear manipulative, especially if it relies on extracting and operationalising users' 'persuasion profiles'. Some psychologists have argued that we have a propensity to be swayed more easily by some tactics than others, which is constant across contexts (Kaptein and Eckles 2010, 2012). On some approaches this draws on quite specific features of individual psychology; on others, we target relatively crudely-drawn personality types with a kind of messaging known to resonate well with that type (Matz et al. 2017). We might thus advertise the same product to two different people in quite different ways, based on our estimation of the likely success of the specific method used for each.

Like many aspects of the infrastructure of Automated Influence, it's hard to say how widespread persuasion profiling is. However, a possibly less invasive analogue is common: A/B testing particular messages with particular target groups. One can soon

discover the effectiveness for each group, and continue to use the most persuasive message, without explicitly categorising anyone according to their persuasion profile.

Tailoring the message is manipulative if it involves identifying and targeting a weakness in the user's rational decision-making. But advertising in general makes a virtue out of identifying and operationalising cognitive biases and heuristics, so if tailoring the message is manipulative, it does not stand out much from other kinds of advertising.

We do think, however, that suasion can be morally problematic (whether we want to call it manipulative or not) when it involves concealing some fact that might, if known, make that suasion less effective. And tailoring the message plausibly does so. If you knew that the same product being advertised to you in one way was being advertised to another person quite differently, you might resist, especially if the messages were somehow conflicting.[40] If you knew that your persuasion profile was being inferred and operationalised, you would very likely refuse to do what you are being influenced to do on that basis alone (Boerman, Kruikemeier, and Zuiderveen Borgesius 2017; Baek and Morimoto 2012).

So there is some reason to think that tailoring the message is problematically manipulative, albeit arguably not a cardinal sin. However, the bulk of online behavioural advertising is not about tailoring the message, but about targeting the product. This concerns both audience selection, and the process of real-time auctioning of advertising spaces, driven in part by predictions of users' click-through rates based on their traits and history (He et al. 2014). In some extreme cases this might be unacceptably manipulative—the much-cited cases of identifying depressed users on

[40] Calo describes firms using personal information to 'extract as much rent as possible from the consumer' Calo 2014, 1029. Whether we conceive of this as manipulation or not, it's clearly a species of morally problematic suasion.

social media and targeting them with products tailored to their depression would perhaps be an example. These, however, are extreme cases. More commonly, targeting the product is a matter of using familiar methods of market segmentation. One might still object that if we knew why they were showing us this ad—not 'because of our browser history', but 'because your mouse hovered over this image on two separate occasions in the past', or 'because your frequent use of smart scales implies that you are dieting'—then we would be less likely to click through.[41]

5.4 How Effective is Online Manipulation?

The most compelling case for Automated Influence involving the manipulation of individual people requires us to look past online behavioural advertising towards the recommender algorithms that shape our experience of digital platforms more generally.[42] These work by shaping our options, as well as influencing our beliefs and desires, to hold our attention for longer and direct it towards products, services and content that we might ultimately be willing to spend our money on. Is *this* an autonomy-undermining form of suasion? On the one hand, perhaps our putative 'addiction' to the products of recommender systems is in fact bad for us; on the other, perhaps this kind of judgment about what makes a life go better or worse ought not be the basis for a broadly liberal critique of Automated Influence. Either way, the mere fact that digital platforms are addictive presumably does not make them much more manipulative than, for example, videogames and other forms of entertainment. It is possible, of course,

[41] In Aguirre et al. 2015, 43 the authors show that undisclosed personalisation is less effective than transparent personalisation by trusted brands; Kim, Barasz, and John 2018 shows that transparent personalisation without background trust of the brand leads to increased reactance.

[42] Again, it is remarkable to note that the very same algorithms first used to target advertisements began the evolution of the Facebook newsfeed algorithm that led to its acute propensity to promote misinformation (Hao 2021).

that the degree of information that social media companies have about their users enables them to more powerfully operationalise our propensity to addiction than is true for other platforms, which might, again, ground valid concerns.

While it might seem hyperbolic to say that Automated Influence has us under remote control, we have found some grounds for saying that it subjects individuals to manipulation. The next question, however, is: how morally serious is this? Manipulation is morally graver, in our view, if (a) it is more successful and (b) the option ultimately chosen by the manipulated is significantly worse than the option they would have chosen, had they not been manipulated. Unfortunately for the prophets of doom, Automated Influence, especially in the form of online behavioural advertising, is not *especially* effective on an individual level (Boerman, Kruikemeier, and Zuiderveen Borgesius 2017; Tucker 2014; Aguirre et al. 2015; Jones et al. 2017; Calo 2014, 1003; Kaptein and Eckles 2010, 2012; Matz et al. 2017; Hwang 2020).[43] It can be significant in the aggregate, as we discuss below. But from each individual's perspective, the probability that they will be successfully influenced by these different kinds of intervention remains small in absolute terms.[44]

Some might object here, that the very fact that the tech companies dominate the advertising market is evidence of their product's success. This would be too quick. Their success arguably comes primarily from their ability to monopolise our attention—to be our default site for search, or for idle browsing. This alone would make their platforms indispensable to advertisers, even if they entirely stopped using user data to target

[43] For a review of empirical literature, and identification of what research needs to be done, see Susser and Grimaldi 2021.
[44] It's important to remember 'the long click', and the fact that individuals might see the same advertisement many times. But these repeated exposures are not independent of one another; it's not like repeatedly rolling a dice such that, by the law of large numbers, they'll get you in the end.

advertisements.

The next question is how much is at stake. In a matter of choosing one product rather than another, the stakes seem pretty low. Of course, online behavioural advertising is also used to market much bigger, life-altering kinds of products, such as unsecured loans and job opportunities. But everything we know suggests that the higher the stakes, the less likely we are to be significantly swayed by advertising of any kind (Boerman, Kruikemeier, and Zuiderveen Borgesius 2017).

What about Automated Influence in political campaigning? Here again the stakes for any particular individual might be relatively low, and the higher the stakes, the less the role we would expect digital advertising to play in their decision. A targeted ad might generate a small donation. A series of such ads might even contribute to a decision not to vote, or (less likely) to switch sides. These might seem pretty significant outcomes, but at the individual level they really aren't, because whether you vote or not, and whether you vote for one side or the other, almost certainly makes no difference to the outcomes for you given the vanishingly small probability that your vote will be decisive.

However, while the effects of manipulation might fail to achieve the intended behavioural changes, they might still succeed in altering the subject's beliefs and desires, and so affecting other aspects of their lives. Automated Influence has clearly contributed to many people in highly digitised societies becoming relatively unmoored from political reality (Vosoughi, Roy, and Aral 2018; Paul 2021; Hills 2019; Törnberg 2018).[45] Properly understanding how Automated Influence has contributed to misinformation and the widespread adoption of conspiracy theories, however, requires

[45] See also https://www.theguardian.com/australia-news/series/web-of-lies.

zooming out from individual interactions to the broader structural implications of Automated Influence. We return to it below, but we acknowledge that the individuals whose worldviews have been significantly altered through content served to them by targeted advertising and rapacious recommender algorithms have suffered a morally serious species of manipulation.

We can draw an interim conclusion that, in general, online behavioural advertising is not significantly more effective than other forms of advertising; even the more nefarious methods don't seem to make that much difference, and anyway it's hard to get too riled up about being nudged into consuming a little more than your budget allows, or spending more time than you think you should staring at a screen.

5.5 Stochastic Manipulation

What happens, then, when we consider the infrastructure of Automated Influence through a wider lens? The magic of big data is in its aggregate effects, which are more than the sum of their parts. The same is true of the harms of big data. They might be relatively trivial for most of those who are adversely affected, while being serious in the aggregate. Even if Automated Influence only involves a modest degree of manipulation of individuals, it permits a more troubling species of *stochastic manipulation* of groups.

By 'stochastic' manipulation, we mean that the interventions of Automated Influence may have a relatively low probability of changing the behaviour of any particular individual, but in the aggregate may make non-trivial impacts on group behaviour as a whole. What's more, in keeping with our account of manipulation above, we think that stochastic manipulation preys on some pathologies of collective decision-making, in

particular our failure to coordinate our actions with one another, and our propensity to realise tragedies of the commons. This is most obvious in the context of political decision-making—not just in elections, but more broadly when mobilising public support for or against particular policy proposals. In these contexts the ability to sway a given group by a few percentage points, even a few fractions of a percentage point, can ultimately prove decisive (Heilman 2020).

Stochastic manipulation also impacts on non-political decision-making. From the perspective of each individual consumer, choosing one product rather than another may make little difference. But at the aggregate level, the inevitability that digital platforms will shape our purchasing choices can lead to serious anticompetitive results. For example, while the nudge we receive to buy products with the 'Amazon Prime' badge may benefit each user individually, each individual transaction contributes to the centralisation of power in the retail economy, putting Amazon's competitors out of business (Romm, Zakrzewksi, and Lerman 2020).[46]

The central moral concern of stochastic manipulation is less its effect on individuals whose decisions are swayed, and more that these new techniques enable small groups of savvy people to exercise a disturbing amount of power over groups and populations at large (Moore 2019). As individuals, we may not be subject to remote control, but the tools of Automated Influence seem to allow those who can wield them an outsized ability to influence populations to advance their goals.

Are individuals gravely wronged by stochastic manipulation? We think not, because an agent's subjective probability of success can affect the seriousness of the wrong they

[46] Indeed, buying from Prime is probably beneficial only because Amazon artificially inflates the price of non-Prime products.

commit, when they successfully manipulate the subject. In other words, if A attempts to manipulate B, and succeeds, then A wrongs B more severely the higher the probability, when A acted, that her manipulation would be successful (other things equal) (Lazar 2015). Recall that the wrong of manipulation consists both in the impact on the victim's autonomy, and in the disrespect shown by the manipulator to the manipulated, in violation of their equal social relations. The impact of being manipulated on B's autonomy is unaffected by A's probability of success when she acted. But the disrespect evinced by A in her action does vary with that probability, we think. A chancy attempt that happens to succeed involves a less egregious species of disrespect than does a sure thing.

To see why this must be so, note that if φing is wrong, then attempting to φ is typically also wrong. When the success of φing is chancy, we concede that successful φing is more seriously wrongful than an unsuccessful attempt. But the difference between them cannot be very great. Suppose then for reductio that A's successfully manipulating B1 with a high probability of success is no more seriously wrongful than her successfully manipulating B2 with a low probability of success. Suppose that A also unsuccessfully attempted to manipulate C2-Z2, with the same probability of success as for B2. If chancy unsuccessful attempts are not much less seriously wrongful than chancy successful harms, and if low probability successful manipulation of B2 is not less seriously wrongful than high probability successful manipulation of B1, then the low probability, unsuccessful attempt to manipulate each of C2-Z2 is not much less seriously wrongful than the high probability, successful manipulation of B1. But this is implausible. C2-Z2 have much weaker complaints against A than does B1. The way out of the reductio is to concede that successful high probability manipulations may be

substantially more seriously wrongful than successful low probability manipulations. Hence the impact of stochastic manipulation on individuals should carry less weight in our deliberations than would less chancy manipulation.

But stochastic manipulation can still pose serious problems. Automated Influence has surely played a significant role in the political upheaval of the last five years (Aral and Eckles 2019). The problem is less that we have ended up in one possible world rather than another, but that a few people have the means to reach and influence so many people, in terms tailored for their particular circumstances. This is especially clear when the tech companies want to get a particular message across to us. Their capacity to reach and influence political communities is extraordinary (Culpepper and Thelen 2020).

Stochastic manipulation concentrates power in too few hands. It also pollutes our capacity for, and willingness to commit to, collective deliberation and action. We tend to think that we are not susceptible to Automated Influence, but that others are (Ham and Nelson 2016, 689). The perception that others are being manipulated is corrosive to democratic deliberation, even if it is in fact overstated. Suppose, to illustrate, that you thought that some part of the population of your country might be Cylons—humanoid robots indistinguishable from homo sapiens without advanced biometric testing, but which can be reprogrammed by a central controller at any given time. Even if you don't know for sure how many Cylons there are, the mere fact that there might be some would be corrosive to public trust. How can we deliberate, debate, and decide in good faith, when some significant portion of our interlocutors might be immune to rational

argument, and effectively under the control of our implacable opponents?[47]

Even when Automated Influence is ineffective, it is perceived to be effective, which undermines trust in the authenticity of one's fellow citizens' deliberations.[48] It is also deeply objectionable that tech companies know how effective this influence is, while leaving the rest of us guessing. Imagine something in the water could be turning people into Cylons. To know whether it is, one needs to test the water at many different points. Only one private company can do so. But they don't make that data available to us, or reliably tell us whether and where the water is contaminated. That would surely be wrong. But it is similar to our situation now.[49]

Stochastic manipulation corrodes democracy, but it may not be the most serious manipulation enabled by Automated Influence. Instead, systems of Automated Influence are *accessories* to a more objectionable, more effective, and more traditional species of manipulation. Automated Influence has funnelled people towards *human* manipulators, because the recommendation algorithms that serve us products, services, and especially content are optimised to sustain user engagement; and content produced by manipulators is, by its nature, deeply engaging to the manipulated (Alfano et al. 2020). Automated Influence steers us towards manipulators, who then take advantage of our emotions, our prejudices and our fears, who lie to us, and who might ultimately incite us to do terrible things (Vaidhyanathan 2018). The worst kind of manipulation in our digital lives right now is being conducted by some of the people who use social media, and they are enabled and empowered by the newsfeed

[47] In Battlestar Galactica, the Cylons do eventually develop a measure of free will, so this hypothetical assumes that they, broadly-speaking, behave more as they did in the earlier seasons, or in general more like 1s and 5s than like 8s.
[48] Scepticism about whether filter bubbles really exist may be beside the point: if people believe they exist, then they have much the same pernicious effects.
[49] Calo 2014, 1006 rightly argued that 'society is only beginning to understand how vast asymmetries of information coupled with the unilateral power to design the legal and visual terms of the transaction could alter the consumer landscape'. Our worry is that this ignorance too is one-sided—we do not understand these effects, but the companies implementing these changes do.

algorithms that drive people towards more sensational, extreme, and polarising content (Hao 2021; Tufekci 2018).

5.6 Democratic Deliberation and Collective Decision-Making

As noted above, the victims of this kind of manipulation arguably have weighty individual complaints against the manipulators, and indirectly against the systems of Automated Influence that empower them (though they must also take *some* responsibility for their own susceptibility). But there are also larger-scale consequences. We all have a very weighty public interest in living in societies that are capable of meaningful democratic deliberation as a prelude to collective decision-making.

The greater the extent to which our public discourse is fragmented by misinformation and conspiracy theories, the less capable we are of reasonable, respectful, collective deliberation. Democratic success depends on norms of public discourse, in which we view one another as valid interlocutors, striving to realise our values in light of broadly accurate and shared beliefs about the world. When significant swathes of the population are simply unmoored from reality, and endorse radicalised values that are wildly out of step with not only the common good but also their own interests, it becomes impossible to have this kind of public forum. 'Democratic' politics becomes nothing more than a thinly veiled struggle for power, which undermines the legitimacy of the whole political process, and makes events such as the January 6 2021 insurrection in the US not just more probable, but all-but inevitable. Such events result from a corruption of public discourse enabled by systems of Automated Influence that serve people content that fires them up and keeps them engaged, at a speed and scale that content-moderation algorithms (and human content-moderators) cannot hope to

keep up with.

Though all the major social media platforms are now trying to redress these effects, we cannot set them to one side as incidental or outlying. The problem is much deeper. The entire business model of Automated Influence depends on optimising for engagement. The only recourse is to incorporate a measure of epistemic paternalism—giving people the information that is good for them, whether they want it or not. This goes beyond simply taking down unacceptable content, but ensuring that content *promotion* is regulated by epistemic ideals. Not only will this prove incredibly challenging to implement, but it aims to solve one problem with the infrastructure of Automated Influence by exacerbating another: the radical centralisation of power in the hands of a few unaccountable corporations. Once again, solving a core problem at the heart of the business model of Automated Influence requires *somebody* to exercise a significant degree of power; yet giving that power to tech companies simply increases our subjection to their unaccountable authority.

6. Conclusion: A Crisis of Legitimacy

We lack the space to do justice to all the plausible objections to Automated Influence.[50] Nevertheless, we see a clear common thread. Automated Influence is, at its heart, a novel mechanism for the exercise of power. It consolidates and adds to the power of the already-powerful, and it creates new agents of power. These new modalities for the exercise of power have emerged from the commercial, private sphere, and as such their sole claim to legitimacy lies in the consent of those affected by them. But, as we have seen, our individual consent does little to legitimate the new power structures of

[50] In particular, we have set aside the concern that it enables and exacerbates structural discrimination against marginalised groups, on which see e.g. Wachter 2020; Noble 2018.

Automated Influence. Indeed, assessing Automated Influence from the individual perspective at all largely misses the point. Instead, we must recognise that in the digital sphere, through our more-or-less uncoordinated voluntary choices, we have created a new set of social structures, which shape significant proportions of our lives. And our existing political institutions have proved distinctively ill-suited to governing those novel structures.

When we have to live together, we are driven to find ways of developing freely self-determining political communities so that we can be at home in the laws to which we are subject. But in our digital lives we are incapable of realising anything approaching this level of collective autonomy. Not only are we subject to the whims of a few extraordinarily powerful corporations, but we are immersed in fundamentally algorithmic governance, our experiences and our options shaped by authorities that are entirely opaque to us: we can't know how they work, or what effects they have, not only because we are precluded from knowing the facts by intellectual property laws, but because the algorithms themselves are inscrutable, and are little understood even by those who designed them (Selbst and Barocas 2018).

Unsurprisingly, this mixture of chaos and untrammelled power has led to seriously deleterious effects (as well as some good ones). The economic imperatives of Automated Influence have left us vulnerable to ubiquitous surveillance. A few corporations control the means of prediction, and the infrastructure that they have created work to fragment us: they reap the benefits of big data, while consigning us to the ideology and practice of small politics, undermining our capacity for collective action. And the mechanisms of Automated Influence allow too few people to subject too many people to stochastic manipulation—relatively trivial for many of the individuals

affected, but in the aggregate potentially changing the destiny of nations—and steer us towards the most adept manipulators of all: each other.

These problems all have more or less the same structure: they are collective action problems, the presumptive solution to which is *more* power, not less—a central authority that can hold the different players in our digital lives to common standards, which allow the market-lubricating aspects of Automated Influence while avoiding the costs. But unless *that* power is legitimate, we would just trade the feudal chaos of our digital lives for a kind of digital authoritarianism.

What's more, the only option less attractive than leaving this power with the titans of tech is giving the same kind of access to national governments, even democratic ones (to say nothing of quasi-democratic supranational organisations). Their power over us is already extreme; with unfettered access to our digital lives as well, the balance of power between us and them would be utterly and decisively skewed. What's more, national governments are by their nature territorial; our digital lives are not. Moreover, democratic governments are notoriously inept at implementing any kind of technological governance. At present, *only* the tech companies are able to implement and enforce reforms that might address some of the concerns in this paper. And they can do so effectively only if they remain, as they are now, large enough to stifle the kind of competition that leads to a race to the bottom. We are therefore at an impasse: subject to new kinds of power and reaping the whirlwind, with few appealing solutions for calming the storm without further empowering our digital masters. The task of all would-be self-governing citizens of the internet—political philosophers included—is to answer this crisis of legitimacy with new ways to realise collective self-determination

in our digital lives.[51]

Acknowledgements: For their helpful comments and advice on earlier drafts of this paper, we thank Annette Zimmermann, Alex Voorhoeve, Kate Vredenburgh, Max Fedoseev, Jake Goldenfein, Charles Evans, Selim Berker, Anne Gelling, the members of the HMI project at ANU, and the anonymous referees for this journal.

Claire Benn (PhD, University of Cambridge) is a postdoctoral researcher on the Humanising Machine Intelligence Grand Challenge project at the Australian National University. Her current research focuses on the ethics, politics and law of data and AI.

Seth Lazar (DPhil, University of Oxford) is Professor at the School of Philosophy and Project Leader of the Humanising Machine Intelligence Grand Challenge at the Australian National University. He works on the moral and political philosophy of data and AI.

[51] While this paper was under review, a number of major regulatory proposals were advanced in the US Congress, indicating an unusually bipartisan political will to curb the power of 'Big Tech'. We lack the space to assess these proposals here; however, they serve to reinforce our point that this is a critical constitutional moment. We suspect that the drive to break up the largest companies (itself unlikely to succeed) will ultimately prove in tension with the desire to solve the other collective action problems discussed in this paper.

Bibliography

Acquisti, Alessandro, Leslie K. John, and George Loewenstein. 2013. "What Is Privacy Worth?" *The Journal of Legal Studies* 42 (2): 249-274.

Aguirre, Elizabeth, Dominik Mahr, Dhruv Grewal, Ko de Ruyter, and Martin Wetzels. 2015. "Unraveling the Personalization Paradox: The Effect of Information Collection and Trust-Building Strategies on Online Advertisement Effectiveness." *Journal of Retailing* 91 (1): 34-49.

Alfano, Mark, Amir Ebrahimi Fard, J. Adam Carter, Peter Clutton, and Colin Klein. 2020. "Technologically scaffolded atypical cognition: the case of YouTube's recommender system." *Synthese*.

Andrejevic, Mark. 2012. "Ubiquitous surveillance." In *Handbook of Surveillance Studies*, edited by Kirstie Ball, Kevin D. Haggerty and David Lyon, 91-98. New York: Routledge.

Aral, Sinan, and Dean Eckles. 2019. "Protecting elections from social media manipulation." *Science* 365 (6456): 858-861.

Baek, Tae Hyun, and Mariko Morimoto. 2012. "Stay Away From Me." *Journal of Advertising* 41 (1): 59-76.

Barocas, Solon, and Helen Nissenbaum. 2014. "Big Data's End Run around Anonymity and Consent." In *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, edited by Julia Lane, Victoria Stodden, Stefan Bender and Helen Nissenbaum, 44-75. New York: Cambridge University Press.

Becker, Marcel. 2019. "Privacy in the digital age: comparing and contrasting individual versus social approaches towards privacy." *Ethics and Information Technology* 21 (4): 307-317.

Bedoya, Alvaro M. 2020. "The Cruel New Era of Data-Driven Deportation." *Slate*, Sept 22, 2020. https://slate.com/technology/2020/09/palantir-ice-deportation-immigrant-surveillance-big-data.html.

Benthall, Sebastian, and Jake Goldenfein. 2020. "Data Science and the Decline of Liberal Law and Ethics." *Unpublished MS*.

---. 2021. "Artificial Intelligence and the Purpose of Social Systems." AAAI/ACM Artificial Intelligence, Ethics, and Society, Virtual Event.

Boerman, Sophie C., Sanne Kruikemeier, and Frederik J. Zuiderveen Borgesius. 2017. "Online Behavioral Advertising: A Literature Review and Research Agenda." *Journal of Advertising* 46 (3): 363-376.

Bruns, Axel. 2019. *Are filter bubbles real?Digital futures*. Cambridge: Polity Press.

Calo, Ryan. 2014. "Digital Market Manipulation." *George Washington Law Review* 82 (4): 995-1051.

Calo, Ryan, and Alex Rosenblat. 2017. "The Taking Economy: Uber, Information, and Power." *Columbia Law Review* 117: 1623-1690.

Clark, Mitchell. 2021. "Google promises it won't just keep tracking you after replacing cookies." *The Verge*, Mar 3, 2021. https://www.theverge.com/2021/3/3/22310332/google-privacy-replacing-third-party-cookies-privacy-sandbox.

Cohen, Julie E. 2000. "Examined Lives: Informational Privacy and the Subject as Object." *Stanford Law Review* 52: 1373-1438.

---. 2018. "The Biopolitical Public Domain: the Legal Construction of the Surveillance Economy." *Philosophy and Technology* 31 (2): 213-233.

Culpepper, Pepper D., and Kathleen Thelen. 2020. "Are We All Amazon Primed? ." *Comparative Political Studies* 53 (2): 288-318.

Draper, Nora A, and Joseph Turow. 2019. "The corporate cultivation of digital resignation." *New Media & Society* 21 (8): 1824-1839.

Fedoseev, Max. 2021. "Understanding climate change as a social structural problem." *Unpublished MS*.

Graepel, Thore , Joaquin Quiñonero Candela, Thomas Borchert, and Ralf Herbrich. 2010. "Web-Scale Bayesian click-through rate prediction for sponsored search advertising in Microsoft's Bing search engine." In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, edited by Johannes Fürnkranz and Thorsten Joachims, 13-20. Haifa, Israel: Omnipress.

Graham, Megan. 2021. "Amazon and Goole reaped big rewards from a rebound in Q4 ad spend." *CNBC*, Feb 3, 2021. https://www.cnbc.com/2021/02/03/amazon-and-google-earnings-showed-big-rewards-rebound-in-q4-ad-spend.html.

Griffin, James. 2008. *On Human Rights*. Oxford: Oxford University Press.

Ham, Chang-Dae, and Michelle R. Nelson. 2016. "The role of persuasion knowledge, assessment of benefit and harm, and third-person perception in coping with online behavioral advertising." *Computers in Human Behavior* 62: 689-702.

Hao, Karen. 2021. "How Facebook got addicted to spreading misinformation." *MIT Technology Review*,

Mar 11, 2021. https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/.
Harwell, Drew. 2019. "Doorbell-camera firm Ring has partnered with 400 police forces, extending surveillance concerns." *Washington Post*, Aug 29, 2019. https://www.washingtonpost.com/technology/2019/08/28/doorbell-camera-firm-ring-has-partnered-with-police-forces-extending-surveillance-reach/.
Haslanger, Sally. 2016. "What is a (social) structural explanation?" *Philosophical Studies* 173 (1): 113-130.
He, Xinran, Junfeng Pan, Ou Jin, Tianbing Xu, Liu Bo, Tao Xu, Yanxin Shi, Antoine Atallah, Ralf Herbrich, Stuart Bowers, and Joaquin Quiñonero Candela. 2014. "Practical Lessons from Predicting Clicks on Ads at Facebook." In *ADKDD'14: Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*, 1–9. New York, NY, USA: Association for Computing Machinery.
Heilman, Steven. 2020. "The Electoral College is surprisingly vulnerable to popular vote changes." *The Conversation*, Jul 15, 2020. https://theconversation.com/the-electoral-college-is-surprisingly-vulnerable-to-popular-vote-changes-141104.
Hildebrandt, Mireille, and Serge Gutwirth. 2008. *Profiling the European citizen : cross-disciplinary perspectives*. New York: Springer.
Hill, Kashmir. 2014. "'God View': Uber Allegedly Stalked Users For Party-Goers' Viewing Pleasure (Updated)." *Forbes.com*, Oct 3, 2014. https://www.forbes.com/sites/kashmirhill/2014/10/03/god-view-uber-allegedly-stalked-users-for-party-goers-viewing-pleasure/?sh=4ac5008a3141.
Hills, Thomas T. 2019. "The Dark Side of Information Proliferation." *Perspectives on Psychological Science* 14 (3): 323-330.
Hwang, Tim. 2020. *Subprime attention crisis: advertising and the time bomb at the heart of the internet*. New York: Farrar, Straus and Giroux.
Jackson, Frank, and Philip Pettit. 1990. "Program explanation: a general perspective." *Analysis* 50 (2): 107-117.
Jones, Jason J., Robert M. Bond, Eytan Bakshy, Dean Eckles, and James H. Fowler. 2017. "Social influence and political mobilization: Further evidence from a randomized experiment in the 2012 U.S. presidential election." *PLOS ONE* 12 (4): e0173851.
Kaptein, Maurits, and Dean Eckles. 2010. "Selecting Effective Means to Any End: Futures and Ethics of Persuasion Profiling." Berlin, Heidelberg.
---. 2012. "Heterogeneity in the Effects of Online Persuasion." *Journal of Interactive Marketing* 26 (3): 176-188.
Kim, Tami, Kate Barasz, and Leslie K John. 2018. "Why Am I Seeing This Ad? The Effect of Ad Transparency on Ad Effectiveness." *Journal of Consumer Research* 45 (5): 906-932.
Lazar, Seth. 2010. "A Liberal Defence of (Some) Duties to Compatriots." *Journal of Applied Philosophy* 27 (3): 246-257.
---. 2015. "Risky Killing and the Ethics of War." *Ethics* 126 (1): 91-117.
---. 2019. "Moral Status and Agent-Centred Options." *Utilitas* 31 (1): 83-105.
Lazar, Seth, and Meru Sheel. 2020. "Contact tracing apps are vital tools in the fight against coronavirus. But who decides how they work?" *The Conversation*, May 12, 2020. https://theconversation.com/contact-tracing-apps-are-vital-tools-in-the-fight-against-coronavirus-but-who-decides-how-they-work-138206.
Lomas, Natasha. 2020. "EU Parliament backs tighter rules on behavioural ads." *TechCrunch*, 2020. https://tcrn.ch/3kinyhn.
Macnish, Kevin. 2020. "Mass Surveillance: A Private Affair?" *Moral Philosophy and Politics* 7 (1): 9-27.
Margalit, Avishai, and Joseph Raz. 1990. "National Self-Determination." *The Journal of Philosophy* 87 (9): 439-461.
Matz, S. C., M. Kosinski, G. Nave, and D. J. Stillwell. 2017. "Psychological targeting as an effective approach to digital mass persuasion." *PNAS* 114 (48): 12714-12719.
Moore, Martin. 2019. "Protecting democratic legitimacy in a digital age." *POLITICAL QUARTERLY* 90: 92-106.
Nissenbaum, Helen. 2004. "Privacy as contextual integrity." *Washington Law Review* 79 (30): 101-139.
Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.
Paul, Kari. 2021. "A few rightwing 'super-spreaders' fueled bulk of election falsehoods, study says." *The*

Guardian, Mar 5, 2021. https://www.theguardian.com/us-news/2021/mar/05/election-misinformation-trump-rightwing-super-spreader-study.
Pridmore, Jason. 2012. "Consumer surveillance: Context, perspectives and concerns in the personal information economy." In *Routledge Handbook of Surveillance Studies*, edited by Kirstie Ball, Kevin D. Haggerty and David Lyon, 321-329. Routledge.
Quinn, Warren S. 1989. "Actions, Intentions, and Consequences: The Doctrine of Doing and Allowing." *Philosophical Review* 98 (3): 287-312.
Rawls, John. 1999. *A theory of justice*. Rev. ed. Oxford: Oxford University Press.
Richards, Neil M. 2013. "The Dangers of Surveillance Symposium: Privacy and Technology." *Harvard Law Review* 126 (7): 1934-1965.
Romm, Tony, Cat Zakrzewksi, and Rachel Lerman. 2020. "House investigation faults Amazon, Apple, Facebook and Google for engaging in anti-competitive monopoly tactics." *Washington Post*, Oct 7, 2020. https://www.washingtonpost.com/technology/2020/10/06/amazon-apple-facebook-google-congress/.
Selbst, Andrew D., and Solon Barocas. 2018. "The intuitive appeal of explainable machines." *Fordham Law Review* 87: 1085-1139.
Solove, Daniel J. 2004. *The Digital Person: Technology and Privacy in the Information Age*. New York University Press.
Southwood, Nicholas, and Geoff Brennan. 2007. "Feasibility in Action and Attitude." In *Homage à Wlodek: Philosophical Papers Dedicated to Wlodek Rabinowicz*, edited by T. Rønnow-Rasmussen, B. Petersson, J. Josefsson and D. Egonsson. www.fil.lu.se/homageawlodeck.
Srnicek, Nick. 2017. *Platform capitalism.Theory redux*. Cambridge, UK ; Malden, MA: Polity Press.
Susser, Daniel, and Vincent Grimaldi. 2021. "Measuring Automated Influence: Between Empirical Evidence and Ethical Values." Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA. https://doi.org/10.1145/3461702.3462532.
Susser, Daniel, Beate Roessler, and Helen Nissenbaum. 2018. "Online manipulation: Hidden influences in a digital world." *Georgetown Law Technology Review*.
Susskind, Jamie. 2018. *Future politics: living together in a world transformed by tech*. First edition. ed. Oxford: Oxford University Press.
Suzor, Nicolas P. 2019. *Lawless : the secret rules that govern our digital lives*. New York: Cambridge University Press.
Taylor, Charles. 1995. "Irreducibly Social Goods." In *Philosophical arguments*, 127-145. London: Harvard University Press.
Taylor, Linnet, Luciano Floridi, and Bart van der Sloot, eds. 2017. *Group Privacy: New Challenges of Data Technologies*. Edited by Luciano Floridi and Mariarosaria Taddeo, *Philosophical Studies Series*: Springer.
Törnberg, Petter. 2018. "Echo chambers and viral misinformation: Modeling fake news as complex contagion." *PLOS ONE* 13 (9): e0203958.
Tucker, Catherine E. 2014. "Social Networks, Personalized Advertising, and Privacy Controls." *Journal of Marketing Research* 51 (5): 546-562.
Tufekci, Zeynep. 2018. "It's the (democracy-poisoning) golden age of free speech." *Wired*, Jan 16, 2018. https://www.wired.com/story/free-speech-issue-tech-turmoil-new-censorship/?src=longreads.
Turow, Joseph, and Nora Draper. 2012. "Advertising's new surveillance ecosystem." In *Routledge Handbook of Surveillance Studies*, edited by Kirstie Ball, Kevin D. Haggerty and David Lyon, 133-140. Routledge.
Vaidhyanathan, Siva. 2018. *Antisocial Media: How Facebook Disconnects Us and Undermines Democracy*. New York: OUP.
Véliz, Carissa. 2020. *Privacy is power : why and how you should take back control of your data*. London: Transworld Digital,.
Viljoen, Salome. 2020. "Democratic Data: A Relational Theory For Data Governance." *Available at SSRN 3727562*.
Viljoen, Salomé, Jake Goldenfein, and Lee McGuigan. 2020. "Design Choices: Mechanism Design and Platform Capitalism." *Unpublished MS*.
Vold, Karina, and Jessica Whittlestone. 2019. *Privacy, Autonomy, and Personalised Targeting: rethinking how personal data is used.* IE University's Centre of Governance of Change.
Vosoughi, Soroush, Deb Roy, and Sinan Aral. 2018. "The spread of true and false news online." *Science* 359 (6380): 1146-1151.
Wachter, Sandra. 2020. "Affinity Profiling and Discrimination by Association in Online Behavioural

Advertising " *Berkeley Technology Law Journal* 35 (2).
Waldron, Jeremy. 1987. "Can communal goods be human rights?" *European Journal of Sociology / Archives Européennes de Sociologie / Europäisches Archiv für Soziologie* 28 (2): 296-322.
Wertheimer, Alan. 1987. *Coercion*. Princeton: Princeton University Press.
Yeung, Karen. 2017. "'Hypernudge': Big Data as regulation by design." *Information, Communication & Society* 20 (1): 118-136.
Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism*. New York: Public Affairs.
---. 2020. "You Are Now Remotely Controlled." *New York Times*, 24 January, 2020. https://www.nytimes.com/2020/01/24/opinion/sunday/surveillance-capitalism.html.