6 Manipulation, injustice, and technology

Michael Klenk

1 Introduction

Can we be manipulated by technology? Science fiction suggests that the answer is yes. In the 2014 movie, *Ex Machina*, software engineer Caleb falls prey to the empathic android Ava's sly charm. She has a subtle grasp of Caleb's needs and desires and feigns romantic feelings for the engineer. However, as it turns out, she merely uses him as a means to flee from her creator's enclosure. Caleb falls in love with her and helps her escape, and Ava leaves him to die once she is set free.[1]

Leave out the fiction, and we lose Ava's extraordinary and (super-)human intelligence and grasp for emotions. Nevertheless, our daily lives already are filled with interactions with technologies that make reliable predictions about our psychology, possess potent means to influence us, and have 'aims' that potentially conflict with ours. For example, what you see on your social media feed is curated by a recommender system – and intelligent software agent – that adjusts its actions in response to yours. Perhaps you only escape your doomscrolling on Twitter when your fitness wearable – a physical device operated by algorithms – signals you to get a move on. And if the device does not function, your first point of contact with the manufacturer will most likely be – and increasingly so – a customer service conversational AI. These observations warrant an investigation into the manipulative potential of these technologies.

In this chapter, I explain how, precisely, people may end up being manipulated by technology. Rather than focusing on the agent perspective and what it takes to manipulate, I focus on the patient perspective and ask what it takes to be manipulated. I show that being manipulated by technology is possible quite independently of whether or not technology has agency or intentionality. My argument depends on a novel perspective on manipulated behaviour, which I call the explanationist-normative perspective. Accordingly, manipulated behaviour is behaviour explained, in the relevant sense, by an injustice. Because technology can afford or enable injustice we can be manipulated by technology.[2] Thus, the chapter first develops a novel

DOI: 10.4324/9781003205425-7

account of manipulated behaviour and then uses that account to say something about being manipulated by technology.

Section 1 sketches how technology affects us quite independently of its inherent properties, which raises the question of whether we end up manipulated. Section 2 disassociates manipulative and manipulated behaviour, suggesting that the former may cause the latter but that we need a separate account of the latter nonetheless. Next, I introduce and defend the explanationist–normative perspective on manipulated behaviour in Section 3. Finally, Section 4 shows that considerations about epistemic injustice and technology's value-laden affordances imply that some of the effects of technology on us may constitute injustices, quite independently of the agential characteristics of technology.

2 Technology as a cause

Ava's interaction with Caleb is an example of a technology interacting with a human. The outcome is horrible for Caleb. The cause of Caleb's demise, Ava, appears perfectly human-like in the relevant aspects, which is probably why their interaction evokes such a strong reaction (at least it did for me and many other movie-goers!).

Phenomenologically, questions about online manipulation seem more pressing once interactions between humans and technology become overtly indistinguishable from human -human interaction. Outside the uncanny valley, where technology appears very different to humans (cf. Mori, MacDorman, and Kageki 2012), we feel forced to consider how to describe and understand correctly what has happened and how to classify the interaction. Was Caleb manipulated? Or would it be mistaken to understand technology like Ava as capable of manipulation in the first place?

Upon reflection, however, we can see that questions about whether people are being manipulated by technology should arise quite independently of the specific type of technology and its capacities. Ava's specific capacities are not the problem (though they may amplify it, or at least make us consider it with more urgency). Technology of much lower capacities than Ava influences us in already significant ways. Once we lay bare these influences and see how interactions with technology give our mental states and behaviour shape, we should again be prompted to ask how to describe and understand correctly what has happened. To illustrate, consider *social robots, virtual software agents*, and *non-autonomous technology*.

Like Ava, social robots are autonomous and physically instantiated, but they lack Ava's futuristic capabilities. Nonetheless, there should not be a doubt that they can be relevant influences on our psychology and behaviour. When, for instance, social robots are proposed to take over important roles in education (Belpaeme et al. 2018), we may worry about them spurring on learners in a problematic way. Granted, there seems to be little

hard evidence about the impact of autonomous and physically instantiated technology, yet we demand that their influence on us be measured by scientific experiments. For example, there are only weak indicators about the effects of social robots in elderly care on well-being (Broekens, Heerink, and Rosendal 2009) and the lowering of depression (Chen, Jones, and Moyle 2018) and that of sex robots on well-being (Döring, Mohseni, and Walter 2020). However, for our purposes, we need not be as strict with the concept of a cause. A new colleague's behaviour and influence on you may worry you even if we cannot scientifically establish his influence on your psychological well-being or some other factor. In the same vein, we can ask what happened when someone who feels grateful to a care robot or in love with a sex robot.

Virtual software agents are, like Ava, autonomous, but they lack a physical instantiation, and yet again, they have an effect on us that we might have reason to classify as manipulation. Consider that people's online consumption, be it social media, videos, music, or other goods, is in large parts orchestrated by recommender systems. Like Ava, these systems are instances of a technology that is intelligent and autonomous in that it can perceive its environment and take actions that maximise its chance of achieving its goals (Aggarwal 2016). Your interaction with such a system can be understood as an interaction between you and an intelligent software agent (Burr, Cristianini, and Ladyman 2018). For example, it may push an anti-vaxxer video into your feed rather than any of the other billion possibilities (Alfano et al. 2020). Virtual software agents still impact out mental states and, ultimately, our behaviour. They can, to a considerable degree, 'read our minds,' that is, make reliable inferences about our beliefs and dispositions based on information gathered about the human user in interaction (Burr and Cristianini 2019). They have been shown to have measurable influences on our affective states as in the well-known emotional contagion study where Facebook users' affective states were influenced by the recommender system (Kramer, Guillory, and Hancock 2014).[3] When you are led down a rabbit hole of, say, more and more far-right videos on YouTube, and you come to believe, say, that the Democrats stole the 2020 US election, then what has happened to you? Were you manipulated? Or would it be a mistake to understand a technology like a recommender system as capable of doing that? These questions are rightly prompted by the nature of the effects that technology has on us. Caleb dies, and Internet users may end up more likely to believe a conspiracy theory. These are bad things, and they cry out for an explanation. But, again, insofar as the effects of technology on us prompt questions about manipulation, we should not restrict ourselves to artificially intelligent technology.

Non-autonomous technology influences us in relevant ways, too. User-friendly design concerns just the exterior features of a technology without requiring that it be autonomous or physically instantiated. However, it has been shown to have distinct effects on thought and cognitive integration (see

Schwengerer, this volume). Moreover, technology is physically instantiated but not autonomous, can also have dramatic effects on us. A prominent example from the philosophy of technology concerns the socio-technical effects of technology. Winner describes underpasses that were intentionally built low and, due to a combination of various technical and sociological factors, prevent certain classes people from reaching desirable areas for recreation. Winner takes this to show that artefacts have politics (cf. Winner 1980).[4] But we need not go as far and ascribe powers to the technology itself. We can focus on the effects it has on people. What has happened to those people who were prevented by the underpasses to go to the beach? Were they manipulated? Or would it be a mistake to describe the influence on them in that way?

With all these different types of technology, it is entirely plausible that we have to attend closely to the capacities of the respective technology to understand what *it* did. It may be more plausible to describe Ava as *being manipulative* than an underpass in a city. However, when we are interested in Caleb's plight, or anyone else who is being influenced by technology, we must focus on what has happened to them.

Therefore, the takeaway from this section is that the specific properties should not matter for the general question of whether technology manipulated us. Technology is interesting for its potential to manipulate us. Some manifestations – notably artificially intelligent technologies with a physical manifestation – may have particularly significant or powerful effects (see Jongepier and Klenk, this volume). But any type of technology can affect us. And that effect may prompt the question of whether we must describe it as a manipulative influence and us, in turn, as being manipulated by the technology.

Next, I put technology aside and ask what manipulated behaviour is, before showing that technology can be a cause of manipulated behaviour in Section 4.

3 Manipulated and manipulative behaviour

One puzzle with manipulated behaviour is that it is not overtly different from non-manipulated behaviour.[5] Their environment, including other agents, constantly influences agents. However, whether the actions they perform or the mental states they adopt as a result are manipulated or not is not evident from the overt mental state or action. For example, falling in love, believing that the election was rigged, buying a new flat-screen TV, getting angry, or starting to cry can be non-manipulated mental states and actions as well as manipulated mental states and actions. Their difference is not readily discernible under an overt description.[6]

Moreover, 'ion' terms like manipulation are ambiguous between process and result. As Hacking (1999) suggests, each of these terms negotiates the difference between both in its own way, and manipulation allows for

a distinction between the active process of manipulating and the passive, receptive upshot of being manipulated.

The existing literature on the nature and value of manipulation is predominantly focused on the former. But although manipulative and manipulated behaviour is related, they are different phenomena. This disassociation is crucial because we cannot rely on existing accounts of manipulative behaviour to say what manipulated behaviour is. After I manipulate you, the behaviour you exhibit will not necessarily overtly differ from non-manipulated behaviour.

Nonetheless, there is a bridge between accounts of manipulative behaviour and manipulated behaviour. On my preferred analysis of manipulative behaviour, manipulation is a kind of negligence in revealing reasons to others (Klenk 2021a, 2021b). A manipulator is negligent in the sense that they ultimately choose their means of influence because it is effective in getting the manipulatee to believe, feel, or desire in a certain way and not because it reveals reasons to the manipulatee. Similar to other norm-based accounts of manipulation (Noggle 1996; Gorin 2014; Barnhill 2014), the negligence account of manipulation suggests that manipulative influence violates a norm. However, unlike previous views, it suggests that the violated norm is best understood as a lack of care to reveal reasons to the manipulatee rather than an active perpetration or ill will on the part of the manipulator. In this sense, manipulative influence is more like bullshit (in the technical sense, introduced by Harry Frankfurt, as not caring for the truth) than lying (intending to communicate a falsity). I will suggest in the next section that manipulative behaviour, thus understood, may often (though not always) be behind manipulated behaviour. Nevertheless, you can already see that whatever we may say about this view of manipulative behaviour, it illustrates a lot about the manipulator and next to nothing about the manipulatee.

Therefore, we need an account of manipulated behaviour – and if that can be shown to connect to and extend existing work on manipulative behaviour, then all the better for it.

4 An explanationist-normative perspective on manipulated behaviour

Some causes of our mental states, and ultimately our behaviour, are injustices. For example, a violation of your right to be treated with dignity – a violation often but – notably – not exclusively perpetrated by manipulators – may cause you to believe falsehoods and do things you did not want. Caleb, for instance, was played with and used as a mere means to Ava's nefarious ends.[7] Similarly, Othello, a prime exhibit of manipulated behaviour (whom we will discuss more later), was lied to and thus got his entitlement to truth frustrated.

In these cases, an injustice explains how the behaviour came about. Thus, injustices are at least correlated with seemingly manipulated mental states

and behaviours. I expand on that correlation and argue that manipulated behaviour is constituted by an injustice that explains the behaviour in a to-be-specified sense. I will call this an explanationist-normative account of manipulative behaviour.

> **Explanationist-normative account of manipulated behaviour:** Some behaviour x is manipulated behaviour if and only if and because x is explained in the relevant sense by an injustice.

My defence of the account can be summarised as follows. I illustrate the account with Othello's paradigmatic case of manipulated behaviour (Section 3.1). Real-world cases of deep oppression provide another example. Deep oppression seems morally problematic, but it has been very hard to account for that. Enoch (2020) argued that it could be accounted for in terms of injustice. If an injustice explains problematic adaptive preferences, then there is prima facie reason to think that manipulated behaviour is explained by injustice, or so I will argue (section 3.2). This can be shown to explain common concerns with adjacent accounts of manipulated behaviour (3.3). Moreover, it would offer a unified account of manipulated behaviour, which is important for independent reasons (3.4).

The argument is thus preliminary in many ways. Most importantly, it is abductive and thus leaves open that a yet deeper unifying explanation of manipulated behaviour can be found. Unless we find such a factor, however, the explanationist-normative account should stand as a serious contender.

4.1 Manipulated behaviour and injustice

Shakespeare's Othello illustrates the constitutive link between injustice as an explanation and manipulated behaviour.

Othello falls for the red herrings planted by his confidante Iago and comes to falsely believe that his wife Desdemona is cheating on him. He is so enraged by her supposed betrayal that he ends up killing her. Iago's scheme succeeded beautifully. Naturally, Othello was manipulated by Iago. In classifying Othello's behaviour as manipulated, we inadvertently suggest that something demarcates his relevant mental states, his belief that Desdemona cheated on him, his infuriation, and the desire to punish her, from your typical non-manipulated mental states.

Manipulat*ive* behaviour can constitute one of the injustices that explain in the right way some manipula*ted* behaviour. The injustice that played a role in Othello's behaviour was Iago's manipulative influence on him. All Iago cared about was his plan to succeed, and thus, his influence on Othello was reckless and negligent. He did not care the least whether Othello saw these reasons except that they made Othello behave as desired. Thus, though Iago was scheming, clever, and highly deliberate in his behaviour towards Othello, he was utterly negligent regarding Othello's reasons. This

description fits the view that manipulative influence is negligence regarding the grounds on which one chooses one's behaviour (Klenk 2020a, 2021a,b). Insofar as norms legislate attention toward revealing reasons to others in interaction, we have here a violation of these norms and thus an injustice. The very violation of Othello's right leads him to have a false belief and unwarranted anger about Desdemona. So, his belief (a particular mental state) was manipulated. Since that belief figured crucially in Othello's subsequent killing of Desdemona, Othello's behaviour was manipulated. The explanation illustrates how the thesis that being manipulated tracks injustice gives us the correct analysis of a pertinent case.

Manipulated mental states are not per se faulty. Like other victims of manipulation, Othello is troubled by their plight. Often, this will be the case because victims of manipulation end up with faulty mental states and even more so if their manipulation engenders horrible behaviour like in Othello's case.

Nevertheless, it is perfectly conceivable that one rightly laments an accurate and non-faulty but manipulated mental state. For example, suppose that Desdemona has, in fact, cheated on Othello. His belief that she cheated on him would be true and not faulty in the propositional sense. Nonetheless, Othello might rightly complain that Iago's scheming and ill will towards him make his resulting belief a manipulated one. 'I have come to a true belief, what I arrived at it in bad ways' he might say. This suggests that it is not the substantive content of a mental state that makes it manipulated or not but how the mental state came about. Thus, we must look to its genesis to understand why it counts as manipulated.

Manipulated mental states are explained in a certain way because something about their genesis is amiss. Importantly, what is amiss is measured in inherently normative terms. The violation of a right or an entitlement – an injustice – plays an appropriate role in the genesis of manipulated behaviour. A normative explanation thus demarcates manipulated from non-manipulated mental states.

Importantly, it does not seem important per se *where* the relevant mental states came from or *who* caused them. For example, when Othello would complain about being manipulated, he would perhaps be saying something about Iago's personality, intention, or capacities (the source of his mental state). However, certainly, he would be saying something about Iago's influence on him and the mental states that it engendered. Thus, it is not important who or what Iago is, but how he influenced Othello.

To illustrate, imagine a rewrite of Shakespeare's Othello, where Iago turns out to be a cyborg just like Ava. Whatever is wrong with Othello's mental states (e.g., they were manipulated) would seem to be the same, irrespective of whether he is dealing with the original Iago or his futuristic counterpart.

This points to the *independence* of the manipulator's capacities from facts about whether or not someone was being manipulated. The independence claim already points to a connection with the larger concern of this chapter,

namely to explore how we can be manipulated by technology in virtue of its influence on us and quite irrespective of its capacities.[8] I will return to this point in Section 4.

For now, it just matters that Othello's mental states seem to differ in some non-substantive sense from his other mental states and the – counterfactual – mental states he would have had, had Iago not manipulated him.

The relevant causal-normative explanation of a given behaviour is both sufficient and necessary for some behaviour to count as manipulated. Take *necessity* first. Once we remove the injustice from the explanation of Othello's mental state, the manipulated behaviour disappears. Suppose Iago had been honest and not manipulative towards Othello. Nonetheless, Othello ends up with the faulty belief that Desdemona cheated on him because of a bad dream or the onset of insanity. Othello's behaviour would seem tragic and wrong but not manipulated. So, without the injustice, we do not seem to have a case of manipulated behaviour, which suggests the necessity of injustice for manipulated behaviour.

To deny the necessity claim, one would have to find manipulated behaviour that did not involve an injustice, however small. Critics may suggest that my proposed account rules out – illegitimately – the possibility of manipulated behaviour with good causal histories.[9] Some examples that may push us against the necessity of injustice for manipulated behaviour may be manipulating a consenting subject in an experiment, sulking to get your partner's attention, or flirtatious behaviour to get someone to desire you in the first place.[10]

However, injustices need not be egregious and fulminant to play their relevant explanatory role in manipulated behaviour. Injustices can be minor, even trivial, perhaps. Many instances of manipulated behaviour can be all things considered permissible, notwithstanding that they remain a prima facie problem. For example, if flirting, paternalistic advice, and treating subjects in experiments results in manipula*ted* behaviour, this is because it involves manipulat*ive* behaviour, which constitutes an injustice.[11] And so being seduced, nudged, or experimented on might come with injustices and are instances of manipulated behaviour. But at no significant cost, so it might be quite plausible to say that, at the end of it, it is to be welcomed. Thus, if there is a case we are prepared to classify as a case of manipulated behaviour, then there will be some injustice – however minor – to be found.[12]

The sufficiency claim is supported by the example of Othello and other run-off-the-mill cases of manipulation. To deny this claim, one would have to show that there are cases where an injustice explains behaviour without that behaviour counting as manipulated.

However, not any kind of explanation will do. Two examples illustrate the sense of appropriate or relevant explanation that I am after. Suppose that illegitimately withheld gratitude is an injustice. Consider first someone who was denied gratitude where gratitude is due. That is an injustice. That person may become acutely aware of a desire to be thanked. The person's

desire to be thanked is undoubtedly caused by an injustice here in some sense. Is he manipulated? No, because the injustice does not explain his behaviour in the *appropriate* sense. The desire to be thanked where gratitude is due will probably always have been there – it was only brought to the person's attention due to the injustice. The injustice is not the root cause of the desire, if you will.

Enoch (2020) discusses a related example to illustrate the appropriate explanation. Someone is taken hostage and kept in a cellar. Her being taken hostage is an injustice. Now, there is almost nothing in that cellar except for a piano, and to pass the time, she starts playing. Eventually, she develops a passion for piano play. Her passion is in some sense caused and explained by her being held captive in a cellar, which is an injustice. Despite that, her desire for piano play is not manipulated.

Neither case is a counterexample to the sufficiency thesis because the injustice does not in the relevant sense explain the resulting behaviour. The injustice is not required, counterfactually, for the desire to arise. It has already been there (as in our first example), or it would have been there in similar circumstances minus the injustice. But for proper cases of manipulated behaviour, the injustice seems to be an essential factor in explaining how the relevant mental state was formed. Thoroughly assessing this claim would require a discussion wider than I can offer here about the conditions for a relevant explanation.[13] However, it seems plausible that behaviours that have an injustice as some (distant) part of their causal chain are not relevantly explained by the injustice. This observation seems to be sufficient to rule out the most pertinent counterexamples to the sufficiency claim.

This section illustrated the constitutive link between injustice as an appropriate explanation of behaviour and manipulated behaviour. So far, the case for that account has been illustrated almost exclusively by a discussion of Shakespearean fiction. However, very real behaviour in our world is no less influenced by injustices and thus no less manipulated. The next section will explore the fruitfulness of the explanationist-normative perspective applied to complex cases in the real world.

4.2 Advantage I: explanatory power

Adaptive preferences as a class of mental states are seemingly morally pernicious, yet it is puzzling to explain their perniciousness. Suppose a recent analysis of the problem due to Enoch (2020) is correct. In that case, Enoch's analysis supports the explanatory-normative account of manipulated behaviour defended earlier while the latter simultaneously extends Enoch's analysis.

Roughly, someone's adaptive preference for x is a preference that person adopted upon realising that y was not among her set of feasible options to the extent that she would now prefer x even if y would become feasible (Bruckner 2009; Enoch 2020). Thus, for example, desiring to have drinks

with your friends via videoconferencing rather than meeting them in person may be an adaptive preference in light of the restrictions on your feasible options surrounding the COVID-19 pandemic. This example is similar to La Fontaine's fox, which realises that it cannot reach the grapes that it so desires, and rather than admitting defeat, resolves that they look sour and that it did not want them in the first place.

Deep oppression cases illustrate this well. Martha Nussbaum gives several compelling examples of women in oppressive personal and socio-economic contexts in India. In their cases, it seems evident that the standards by which they measure their well-being or internal state are distorted and whose resulting preferences appear problematically adaptive (Nussbaum 2001, 112–13). Mitchell (2018) recaps these cases as follows:

> Vasanti, after years in an abusive marriage, thought her abuse to be a normal part of a woman's life, something to be expected once she left her family home to live with her husband. Jayamma, despite being paid less than men for more demanding factory work, accepted that this was how things were, and, knowing change was not possible, did not even waste energy lamenting her situation. And severely malnourished women in Andhra Pradesh, prior to the efforts of a government consciousness-raising program, didn't consider themselves to be malnourished, or their conditions to be unhealthy.
>
> (discussed in Mitchell)

Vasanti, Jyamma, and the other women seem manipulated, and their adaptive preferences are morally worrisome. Nevertheless, despite being harmed by the oppressive practice that they adapted to, they also appear to be strong advocates of the practice. Thus, several well-known attempts to spell out the moral problem in terms of an autonomy deficit for adaptive preferences seem to run into problems.[14] Some analyses may succeed in the future. However, Enoch (2020) makes a compelling case that this is unlikely. The problem is not how the preferences of deeply oppressed persons relate to their other preferences or whether they are preferences for things that are morally good or bad (though that may be *another* problem, cf. Nussbaum 2001).

Instead, Enoch (2020) suggests that their preferences are deeply oppressed and thus morally problematic because they were caused by injustice. Accordingly, Vasanti's preference turns out to be non-autonomous in problematic ways because her adaptive preference for a certain kind of marriage is explained by the injustice of living in such an arrangement for years. The latter is an injustice in many ways, not least because it violates Vasanti's right not to be harmed.

My analysis of manipulated behaviour draws heavily on and is indebted to Enoch's analysis of adaptive preferences. Irreducible normativity is the most crucial element that the explanationist-normative account of manipulated

behaviour takes from Enoch (2020). We cannot understand what manipulated behaviour is without a moral perspective on what counts as an unjust influence. Deep oppression cases were helpful to illustrate this point because they feature agents whose moral problem seemingly evades us when we myopically focus on their autonomy. However, we can always conjure up cases that meet whatever criterion of autonomy we can think of and nonetheless seem problematic. Thus, we need to normatively evaluate a preference' genesis to explain *why* adaptive preferences are problematic if they are autonomous and not preferences for bad things (in that respect, the deep oppression cases discussed here are unfitting examples).

However, I also think that Enoch is analysing manipulated behaviour rather than merely problematic adaptive preferences. It is not entirely clear whether Enoch suggests that problematic adaptive preferences in cases of deep oppression are problematic in virtue of being non-autonomous or problematic in virtue of how their non-autonomy is explained, namely in terms of injustice. The latter would, on Enoch's analysis, entail the former. I suggest that the latter explains manipulated behaviour and not just non-autonomy.

First, there is no principled reason to believe that, in some sense, problematic preferences are to be explained differently than problematic mental states in general. On the contrary, corrupted preferences, desires, beliefs, and emotions are precisely the ingredients of manipulated behaviour.

Second, the set of manipulated behaviours intersects only the set of non-autonomous preferences. Some non-autonomous preferences do not amount to manipulated mental states, as other forms of influence like coercion also engender non-autonomy. And insofar as being manipulated does not require non- or less-than-fully autonomous preferences, there are some fully autonomous but nonetheless manipulated behaviours (compare Buss 2005). The latter, of course, is controversial and goes beyond anything I can hope to discuss in sufficient detail here (though see Klenk and Hancock 2019). But suppose it is true that there is no conceptual connection between being manipulated and being less-than-fully autonomous. Then we can still explain the problem in cases of seemingly problematic adaptive preferences like deep oppression in terms of being manipulated, and we need not find a further explanatory connection between non-autonomy and being manipulated.

Therefore, the explanationist-normative account of manipulated behaviour can explain what is wrong with deep oppression while also explaining how seemingly problematic influences that do not impact autonomy are problematic. The account thus explains well a set of highly relevant real-world cases.

Next, we will see how the account explains a common concern behind adjacent but competing accounts of manipulated behaviour, thereby extending its support.

4.3 Advantage II: explains common concerns

A central motivation of the explanationist account of manipulated behaviour is that adjacent but competing attempts to explain manipulated behaviour often fail. They fail for being too narrow (they do not explain all cases of manipulated behaviour) because they require conditions that manipulated behaviour does not need. Therefore, the explanationist account of manipulated behaviour should be preferred over these other accounts.

First, manipulated behaviour has sometimes been associated with a particular *process* that brought it about. Specifically, several authors have emphasised the connection between affective formation of mental states and manipulation, suggesting that being manipulated has something to do with having mental states formed through such processes (cf. Fischer in this volume; Wildman, Rietdijk, and Archer in this volume). However, it is not the process that is at fault but the injustice behind it. The association between emotion, affect, and being manipulated is indeed often there, but it is merely a spurious connection, and it cannot explain all cases of being manipulated. Manipulated mental states can be formed on a purely cognitive and rational basis such as Othello's belief that Desdemona cheated on him. Moreover, the epistemic or moral warrant of a form of mental state genesis, or the rationality of a type of influence, does not depend on the type of information per se but on the contextual factors at hand.

This claim can be briefly illustrated with the debate around System 1 and System 2 processing (cf. Kahneman 2012). The former is associated with 'non-rational' mental processes such as heuristic decision-making. In contrast, the latter is associated with 'rational' mental processes such as reflection and conscious deliberation. But that does not settle questions about the normative rationality irrationality of System 1 versus System 2 processing. For example, fast affective heuristics are rational when decisions must be made quickly in familiar environments (Gigerenzer 2008). The type of information or the manner of its processing per se is epistemically and morally neutral. Therefore, we cannot identify manipulated behaviour with the type of informational source nor the processes that lead to the behaviour in question.

Second, it is implausible that mental states can be distinguished into manipulated and non-manipulated based on their relation to the *agent's plans*, *aims*, *or* (*self-*)*interest*.[15] Several scholars have championed this proposal (e.g., Barnhill 2014; Rudinow 1978), and it is evident that being manipulated is often not good for you. But clearly, our manipulated mental states are sometimes conducive to our objectively warranted plans or aims. For example, a little nudging may help me avoid the temptation to book a transatlantic flight as soon as travel restrictions abate. Because at least some nudges lead to manipulated behaviour, this is a counterexample to the proposal that manipulated mental states and actions are at odds with

our objectively warranted plans or aims (cf. Sunstein 2016; Klenk 2020a). Neither do our subjectively held plans or aims need to be at odds with manipulated mental states or actions. Being nudged to avoid booking the transatlantic flight seems like manipulated behaviour, and yet it may serve my aim to live carbon-neutral, whether or not this is an appropriate aim to have. Therefore, manipulated behaviour need not be at odds with our aims, plans, or (self-)interest (cf. Gorin 2014). However, what is required is that we judge the genesis of the relevant mental state to contain an injustice. And that seems to be the case. Take nudging as an example. At least some nudgers are manipulative in the sense that they are negligent about revealing reasons to their interlocutor, which we identified with an injustice earlier. This can account for the resulting manipulated behaviour, quite independently of whether the manipulatee's plans, aims, or self-interest were frustrated.

The most promising lead is that manipulation is a kind of interference and, consequently, manipulated mental states are those that were meddled with or interfered in in problematic ways. This is the popular image of the manipulator as the puppet master and the manipulated person as a puppet on a string, as a prop in someone else's play.

However, as alluded to in the previous section, the link between manipulated mental states and behaviour and autonomy is not conceptual. How manipulation impacts autonomy would have to be explained in a more substantive sense (cf. Klenk and Hancock 2019). So, there is a more general lesson here. Any account of manipulation that wants to understand manipulation as a kind of interference that diminishes autonomy must account for how 'normal' or non-interfered processing goes. And I reckon it will be incredibly tough to say how people who are always influenced by their past and present and who at various points are prone to endorse those influences as in deep oppression cases reflectively are functioning in a 'normal' or non-interfered way.

What seems problematic for manipulated people is that their rights are violated, perhaps because they have been influenced negligently. This can be given a distinct Kantian flavour, in that victims of manipulation were not treated with due respect, which clarifies what right is violated (see Jongepier and Wieland, in this volume). The image of being a puppet on a string is misleading if it suggests that we necessarily are less than fully autonomous when we are being manipulated. However, it is apt to evoke a sense of disrespect and violation of one's rights – after all, we are not puppets on a string and should not be treated that way. In this sense, the explanationist-normative account of manipulated behaviour illustrates very well the core concern with manipulated behaviour and ties in nicely with accounts of manipulative behaviour.

4.4 Advantage III: avoids error theory

Finally, the explanationist-normative account of manipulated behaviour is supported by a *reductio* argument.

We should accept the thesis that being manipulated tracks injustice to avoid an error theory about judgements about manipulation (i.e., a theory that explains how and why people are frequently mistaken in their judgements about manipulation). So, suppose that it is false that being manipulated tracks injustice. Then, manipulation is *not* related to injustices that play an appropriate causal role in the formation of a mental state or the generation of an action, and we should *not* expect normative judgements to track causal histories of mental states affected by injustices (in the appropriate way). However, there is widespread disagreement about what makes some behaviour an instance of manipulated behaviour.[16] I call this phenomenon:

Classification Variety: There is widespread disagreement about the conditions for manipulation.

Classification Variety is supported by two sources, preliminary empirical studies and the discussion in the philosophy of manipulation. The 'charting the field' chapter for this volume has shown that there is considerable disagreement about the nature of manipulation (Jongepier and Klenk, chapter 2 in this volume). Normative concepts are controversially discussed more generally. Further defence for this claim that professional philosophers disagree about the nature of manipulation may be produced at will.

A novel data point is that laypeople seem to disagree about the nature of manipulation, too. In an unpublished experiment, Klenk, Xun Liu, and Hancock (2021) asked participants to evaluate short vignettes that described paradigm cases of manipulation (e.g., Shakespeare's Othello) on four dimensions concerning the effect on the manipulatee: they were 'deceived,' 'harmed,' 'played,' and 'unconsciously influenced.' The four answer options were pre-experimentally selected based on the philosophical discussion about necessary and sufficient conditions for manipulation. The results showed that while subjects considered the vignettes as examples of manipulation, they disagreed significantly about the underlying condition. Just like the professional philosophers, laypeople identify several distinct causes as the underlying condition of manipulation. This supports Classification Variety. Now, we must take an important mental note. All the relevant examples plausibly include a causally relevant injustice (depending on the right theory of justice at the end of the day, of course). If we can interpret these varying judgements as tracking injustice instead, then we might explain away Classification Variety. But we are getting ahead of ourselves. First, I need to show that we should try to explain away Classificatory Variety.

Suppose also that there is a unified condition for manipulated behaviour (though it is not injustice!). It might be that manipulated behaviour depends on undermined autonomy. Or on deception. Or on emotional influence. But only one. These assumptions (the rejection of my thesis and that there is a unified condition for manipulated behaviour) coupled with Classification Variety would imply that a sizeable portion of the beliefs about the

conditions for manipulated behaviour – advanced by professional philosophers and laypeople alike – are false. That is because there is one underlying condition of manipulation, while people apparently hold widely differing beliefs about what that condition happens to be. So, Variety implies what I will call

> **Classification Error**: Many beliefs about the conditions for manipulation are false.

I can now show that we should not accept Classification Error and thus reject any assumption that commits us to it. Classification Error is unpalatable, as evolutionary considerations will show. Humans developed a reasonably elaborate capacity to detect cheaters (and, alas, to cheat ourselves). This does not mean that people are good at detecting. But it suggests that we usually know that we are being cheated *when we see it*. Being deceived and being manipulated are some of how we can be cheated. We should expect that social animals like us are good at recognising deception and manipulation, at least when they occur in environments similar to our environment of evolutionary adaptation.[17] When people agree that a given case exemplifies manipulated behaviour, we have good prima facie reason to think that the case indeed does exemplify manipulated behaviour. But given the assumption that it is *false* that manipulated behaviour is caused by injustice, we lack a unifying explanation of these judgements. Absent an explanation, we have to assume that most of these beliefs are false.[18] This is not what we should expect given our evolutionary history.

Considering an objection to this line of thought will further strengthen it. Evolution, the objection goes, did not select for a correct appreciation of the underlying condition for manipulation but the mere 'blind' application of the concept. For instance, classifying behaviour as being manipulated may serve a function, and adaptive pressure may have applied to the utilisation of that function, not correctly identifying the conditions for manipulation.

However, correct classification absent an understanding for the underlying reasons for why something is an instance of manipulated behaviour is insufficient for two reasons. First, it would be an open question just why people have competence in applying the term without some kind of insight. Positing insight would answer this question. Second, even setting that worry aside, there is a substantive problem because different ascriptions of the underlying conditions behind manipulation are plausibly functionally differentiated. That is, different conclusions follow from calling something caused by autonomy-undermining or from deception. Thus, even if evolutionary pressure applied to whatever functional implications (the concept of) manipulation may have, they plausibly indirectly put pressure on the correct recognition of the conditions for manipulation.

Therefore, if it is false that being manipulated tracks injustice, we get the problematic implication that people do not understand the conditions

that ground manipulated behaviour and make many mistakes in applying it. Because this implication is problematic, I conclude that we should not reject the thesis that being manipulated tracks injustice. In other words, the explanationist-normative account of manipulated behaviour should be accepted.[19]

5 Technology's manipulative potential

So far, we have established that interacting with technology puts as at risk of being manipulated by it. For example, it makes sense to classify Caleb as being manipulated by Ava. More generally, if manipulated behaviour is behaviour explained, in the relevant sense, by an injustice then we can be manipulated by technology, quite independently of whether it possesses agential features such as intentionality. That is because agential features are not required for an injustice to explain a mental state and, ultimately, behaviour. So, whether or not we would be correct in ascribing mental states and intentions to Ava does not matter for the question of whether Caleb has been manipulated.

Are there any more general ways in which technology may contribute to an injustice? I will first discuss a general non-agential injustice and then elaborate on technology's causal effects in support of this claim.

Epistemic injustice gives us reason to think that agential features are not required for injustice to contribute to a mental state and behaviour.[20] Fricker (2011) introduces the notion of 'epistemic injustice,' which arises when somebody is wronged in their capacity as a knower. The stock example of epistemic injustice, of the hermeneutical kind, is that of a person or social group that is unfairly deprived of knowledge because of their lack of access to education or other epistemic resources. Fricker discusses two kinds of epistemic injustice in greater detail. First, testimonial injustice occurs when somebody is given less credibility than due to prejudice about the social group to which the speaker belongs. Second, Fricker describes hermeneutical injustice, which occurs when members of a social group fail, because of a linguistic gap in collective understanding, to make sense of certain distinct experiences (e.g., sexual harassment). The idea is that women, for example, were socially powerless in the 1970s and, partly because of that, could not communicate their experiences adequately (cf. Keane 2016). When people are subject to hermeneutical injustice, no direct agent (nor a group agent) perpetuates the injustice, even though at some point agents may have been involved in contributing to the injustice. But whatever 'original' agential contribution there is, it is most likely not required to explain the effects of the injustice today. Whether or not this or that agent was involved in creating systematically oppressive circumstances may matter for questions about responsibility but not for the question of whether your or my behaviour today is explained by the injustice in the appropriate way. Therefore, agential contribution is not required for injustice to appropriately explain a mental state.[21]

Technology can contribute to injustice and thus make us manipulated because of technology's value-ladenness. The idea that technology is more than a 'mere tool' is deeply ingrained in the philosophy of technology. If technology were a mere tool, then any of its effects would have to be attributed to – roughly – the designer of the tool or its user. The NRA makes use of that idea when they claim that 'Guns don't kill people, people do.' But surely guns contribute in some sense to extraordinarily high murder rates in the United States compared to other countries (cf. Grinshteyn and Hemenway 2016), though we need not understand their contribution in an agential sense. From that perspective, technology does not seem to be morally neutral. One way to make sense of sense of technology's value-ladenness without ascribing agential features to it is in terms of affordances. Technology has affordances, which are relational properties that depend on the material properties of the technology as well as contextual factors such as biological, psychological, and social factors concerning the user of the technology (Klenk 2020b). Affordances make certain mental states and behaviours more likely and others less likely. It makes sense to speak of a chair 'inviting' us to sit on it. The affordance perspective on technology helps us interpret this claim without retorting to an implausible ascription of agency or intentionality to technological artefacts. For example, the fact that a gun affords killing indicates that handling a gun will make deadly outcomes in some scenarios, like a heated argument, more likely. Similarly, social robots are suspected of lowering depression and increasing well-being, and virtual software agents have been shown to afford more and more extreme viewing behaviour on YouTube. Even non-autonomous technology like user-friendly websites or low-built overpasses affords some mental states but not others, such as trust in the case of user-friendly websites and not going to the beach in the case of overpasses. And Ava's incredible artificial intelligence made it likely that Caleb fell in love with her without seeing her nefarious scheme. We can also evaluate the affordances of a given technology in moral terms (cf. Klenk 2020b). So, we can also see how technology is not value-neutral from the affordance perspective.

Most importantly, it is now straightforward to see that the affordances of technology can constitute injustices that explain, in relevant ways, our mental states and behaviour. For example, all of a city's citizens are entitled to frequent the city's public beach. Low-hanging underpasses that prevent some citizens from going to the beach violate their entitlement and thus constitute an injustice. It would follow that citizens in that situation are being manipulated by the architectural features of the city. Similarly, we are entitled to truth (suppose). Virtual software agents in recommender systems make it more likely that we believe falsehoods. Thus, they contribute to a violation of our entitlement. That injustice may explain why some end up believing that the 2020 US election was rigged. They are manipulated, according to the explanations-normative account of manipulated behaviour. Caleb, finally, has a right to be shielded from seduction. Ava violated that right, and that injustice explains Caleb's behaviour. Therefore, Caleb was

manipulated by Ava, even if Ava lacks the capacities for genuine manipulative behaviour as intentionality. These observations about concrete cases of technological manipulation depend on identifying a relevant injustice and explaining the relevant mental states and behaviours. However, they should be sufficient to show how technology has manipulative potential quite independently of its agential characteristics.

My argument for the manipulative potential of technology suggests that a prominent and competing type of argument in the ethics of technology is beside the point. I call this type of argument a *condicio sine qua non argument*. Proponents of such arguments describe conditions for manipulative behaviour and then suggest that technology currently or in principle lacks the conditions for manipulative behaviour (compare the contributions by Pepp et al., Gorin, and Nyholm, in this volume).

For example, it may be claimed that manipulativeness requires intentionality and that technology lacks intentions. Therefore, one might conclude, technology cannot manipulate us. However, arguments along these lines miss the possibility, demonstrated earlier, that manipulating (the agent side) can come apart from being manipulated (the patient side). Even if technology cannot be manipulative – for example, because there is no sense in which it can be negligent[22] – it may contribute to injustices that result in manipulated behaviour on our part. Of course, this is because being manipulated does not require that one interacts with a manipulator with intentions, even if the latter will, in many cases of human-to-human manipulation, be the cause of manipulated behaviour.[23]

Thus, technology may relevantly contribute to an injustice that plays an appropriate role in explaining our behaviour. Therefore, there is potential for us to be manipulated by technology.

Usually, we would be wont to ask about the perpetrator and the person culpable of manipulative action. This raises an important question. If there can be manipulated people without manipulators we face a responsibility gap. Some questions about passive responsibility may not be satisfactorily answered. But note two points. First, the explanationist-normative account of manipulated behaviour does not replace the need to ask questions about passive responsibility about the inventors and deployers of technology. Clearly, facts about whether or not soldiers are assessable in terms of responsibility do not absolve their higher-ups from such questions. Second, questions about passive responsibility arguably should not focus on appropriate ethics of technology in the first place (Klenk and Sand 2020). We can still ask questions about forward-looking responsibilities to prevent manipulated behaviour, which is indeed what we should focus on.

6 Conclusion

Interacting with increasingly autonomous technology raises all sorts of problems, as the burgeoning debate, especially in AI ethics, demonstrates. Is one of the problems that we can be manipulated by technology?

This chapter explored a novel approach to that question – by focusing on the patient rather than the agent side of manipulation – and suggested that the answer is affirmative. Manipulated behaviour is behaviour that is explained, in the relevant sense, by an injustice. Agential features like intentionality are not required for injustice, as the case of epistemic injustice demonstrates. Technology can contribute to said injustices in virtue of its affordances. Therefore, we can be manipulated by technology, even if it lacks agential features such as intentionality and thus does not meet the conditions for being manipulative.

That leaves the practically most relevant question of whether we are, in fact, being manipulated by technology. My chapter suggests concrete ways forward with this question. We must assess whether the influence of technology on us constitutes injustices. That involves a question about the proper explanation of our mental states and behaviour and a normative account of what injustices are. Thus, two broad research challenges arise. First, we need much more empirical work to substantiate the concrete ways in which particular instances of technology influence our mental states and behaviour. Second, those influences need to be assessed in light of an appropriate theory of justice to see whether they violate our rights and entitlements. Given the manipulative potential of technology, it is our forward-looking responsibility to ensure that it does not materialise, and we are spared Caleb's plight.

Notes

1. Many thanks to Fleur Jongepier, Michael Madrey, Nathan Wildman, Sven Nyholm, and Jan Willem Wieland for written comments on an earlier version of this chapter. Also, I thank Steffen Steinert and the audiences at a TU Dresden workshop and the online symposium series we organised for this volume for very helpful discussion. My work on this volume was made possible by a Niels Stensen Fellowship. I gratefully acknowledge generous support by the European Research Council under the Horizon 2020 programme under grant agreement 788321.
2. I will often suggest for illustrative purposes that manipulated behaviour – or a manipulated action – is based on a manipulated mental state. Whether that claim about the relation of mental states, action, and behaviour is plausible depends in part on wider issues than I can discuss here. Readers who see a problem in that simple sketch may just focus on my core claim about the conditions of manipulated mental states.
3. This case is prominently Wildman, Rietdijk, and Archer, in this volume.
4. Also consider a point I do not address here, namely that some kinds of interactions may be made possible in the first place by new technology, like augmented many-to-many interaction. See Cappuccio et al. (2021), in this volume.
5. Note that I use the term 'manipulated behaviour' to refer to manipulated mental states and manipulated actions.
6. Note that I may be the first to make this distinction explicit, but I am not the only one who defends it. (Wilkinson 2013) notes in his discussion of a general account of manipulation that it may be premature to assume that manipulative action leads to manipulated action. His point is that social science is difficult.

But it obviously depends on the thought that manipulative actions do not imply manipulated actions. The converse may also hold true. At least, that is what I will assume in what follows.

7. Relatedly, a violation of your right to bodily integrity may cause you to feel threatened and cave in to illegitimate demands. Or a frustration of your civic entitlement to be informed by media in a factual manner about politics may cause you to believe falsehoods, to desire irrational things, and to vote for the wrong party.
8. Most relevantly, as discussed earlier, is probably the capacity for intentionality. See the overview by Jongepier and Klenk, in this volume.
9. Thanks to Fleur Jongepier for pressing me to address this point, and to Jan Willem Wieland for putting this point to me in that way.
10. Perhaps my proposed analysis of manipulation would seem to require a revision of our concept manipulation. Compare Pepp et al., in this volume.
11. Thanks to Fleur Jongepier for helpful feedback on this point.
12. One class of counterexamples are cases of manipulation in the context of a game. Nathan Wildman suggested a case along the following lines. Suppose that Iago and Othello are playing chess, and Iago manipulates Othello by making a series of moves in order to get him to think that he's going to attack queenside, when in fact he's going to go kingside. As a result of Iago's manipulation, Othello builds up his defences in the wrong spot and ends up eventually losing. That strikes some as a case of manipulation, but one that tracks no injustice: Iago manipulated Othello, but he did nothing wrong! I would maintain that we do not have a case of manipulation here because Iago stuck entirely to the rules of the game. So, even though he presumably did not care for whether Othello recognised his reasons for acting, there is no norm within the game that would demand such care. Perhaps Othello was fooled then or duped but not manipulated.
13. Thanks to W. Jared Parmer for pressing me on the distinction between being caused and being explained by an injustice.
14. To illustrate, consider that the women's preferences are not irrational or non-autonomous insofar as they internalised the practice to an extent that they desire what they want to want (on an 'internal' conception of autonomy, cf. Frankfurt 1971) or reflectively endorse the desire, as part of a self-affirming practical identity (cf. Bruckner 2009; Christman 2014). Insofar as the oppression is sufficiently thorough, it is likely that their seemingly problematic preferences are in harmony with their other preferences, thus denying the claim that the problem is formal (cf. Bovens 1992).
15. More generally, we could also call these objective list theories of manipulated behaviour, because they propose lists of goods (e.g., alignment with one's aims or plans) that manipulated behaviour arguably lacks. The short rebuttal of these proposals is that for any entry on a list we can imagine a behaviour that possess that item but still counts as problematic or a behaviour that lacks the item but does not count as problematic.
16. Another crucial clarification concerns the claim about injustice. On its face, my thesis is ambiguous between people judging that an injustice played an appropriate role in some behaviour (the mentalist interpretation) and the fact that an injustice played an appropriate role in some behaviour (the causalist interpretation). Making the distinction clear is important because my thesis drives on an argument about people's judgements being on track.
17. Which is precisely why I might be especially worried about technological manipulation as it supersedes our adaptations. Fleur Jongepier and I discuss this as an aggravating factor, in chapter 2 of this volume. The current point, however, is

not that we should be good at detecting when a machine manipulates us but at identifying the criteria for manipulation.

18. This is a bit quick: it would be reasonable to assume that at least one already identified candidate condition is correct (e.g., deception). Then beliefs whose content portrays manipulated behaviour as depending on other factors are false.
19. The assumption that there is a unified or single condition for manipulation may be controversial, and my argument depends on it. But there is good reason to accept it. But suppose you deny that there is but one condition for manipulation and insist that there are multiple, disjunctive, and individually sufficient conditions for manipulated behaviour. If that is true, then we can explain Variety without accepting Error. People may simply classify correctly several conditions for manipulation and the allegedly absurd consequence Error would not follow from the rejection of the thesis that being manipulated tracks injustice. However, turning to pluralism about the conditions for manipulation is ultimately unconvincing. First, we are still in the dark about the necessary conditions for manipulation. We now assume that there are many sufficient ones. But there is no apparent structure to the many that emerge from people's classifications. But which ones, precisely? All of the ones that we have discussed so far? Or only some? Our understanding of manipulation has not been illuminated. But even if we grant the assumption that there are multiple conditions for manipulation, there is a deeper problem. The explanation in terms of pluralism does not jibe well with the aim of explanatory parsimony. A simpler theory is more likely to be correct. There are constraints about applying the criterion of parsimony in the normative case, see (Sober 2015), but they do not change that a simple explanation is to be preferred to a potentially complicated explanation. Therefore, there is good reason to accept the view that being manipulated tracks injustice.
20. Thanks to Steffen Steinert for suggesting epistemic injustice in discussion as a point in favour of the explanationist-normative account of manipulated behaviour.
21. See Liao and Huebner (2021) who present a fuller account of how technology can be a relevant cause in the injustices that we suffer. Unfortunately, I could not engage with their account more fully in this chapter. Thanks to Sven Nyholm for the pointer.
22. Note that I previously argued that the fact that technology cannot care for our reasons supports an a priori argument about their manipulativeness, (Klenk 2020a). I am now not sure anymore whether the impossibility of technology to have agential features would make it a priori manipulative or just altogether remove it from the category of things that can or cannot be manipulative.
23. Note that Sharkey and Sharkey (2020) have recently suggested an argument along similar lines in the case of deception. Thanks to Sven Nyholm for the pointer.

7 References

Aggarwal, Charu C. 2016. *Recommender Systems: The Textbook*. Cham: Springer.

Alfano, Mark, A. E. Fard, J. A. Carter, P. Clutton, and C. Klein. 2020. "Technologically Scaffolded Atypical Cognition: The Case of YouTube's Recommender System." *Synthese*, 1–24.

Barnhill, Anne. 2014. "What is Manipulation?" In *Manipulation: Theory and Practice*, edited by Christian Coons and Michael Weber, 51–72. Oxford: Oxford University Press.

Belpaeme, Tony, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. 2018. "Social Robots for Education: A Review." *Sciences Robotics* 3 (21). doi:10.1126/scirobotics.aat5954.

Bovens, Luc. 1992. "Sour Grapes and Character Planning." *Journal of Philosophy* 89 (2): 57. doi:10.2307/2027152.

Broekens, J., M. Heerink, and H. Rosendal. 2009. "Assistive Social Robots in Elderly Care: A Review." *Gerontechnology* 8 (2). doi:10.4017/gt.2009.08.02.002.00.

Bruckner, Donald W. 2009. "In Defense of Adaptive Preferences." *Philosophical Studies* 142 (3): 307–24. doi:10.1007/s11098-007-9188-7.

Burr, Christopher, and Nello Cristianini. 2019. "Can Machines Read our Minds?" *Minds and Machines* 29 (3): 461–94. doi:10.1007/s11023-019-09497-4.

Burr, Christopher, Nello Cristianini, and James Ladyman. 2018. "An Analysis of the Interaction Between Intelligent Software Agents and Human Users." *Minds and Machines* 28 (4): 735–74. doi:10.1007/s11023-018-9479-0.

Buss, Sarah. 2005. "Valuing Autonomy and Respecting Persons: Manipulation, Seduction, and the Basis of Moral Constraints." *Ethics* 115 (2): 195–235. doi:10.1086/426304.

Cappuccio, M. L., E. B. Sandoval, O. Mubin, and M. Velonaki. 2021. "Robotics Aids for Character Building: More than Just Another Enabling Condition." *International Journal of Social Robotics* 13: 1–5.

Chen, Shu-Chuan, Cindy Jones, and Wendy Moyle. 2018. "Social Robots for Depression in Older Adults: A Systematic Review." *Journal of Nursing Scholarship* 50 (6): 612–22. doi:10.1111/jnu.12423.

Christman, John. 2014. "Coping or Oppression." In *Autonomy, Oppression, and Gender*, edited by Andrea Veltman, 201–26. Oxford: Oxford University Press.

Döring, Nicola, M. R. Mohseni, and Roberto Walter. 2020. "Design, Use, and Effects of Sex Dolls and Sex Robots: Scoping Review." *Journal of Medical Internet Research* 22 (7): e18551. doi:10.2196/18551.

Enoch, David. 2020. "False Consciousness for Liberals, Part I: Consent, Autonomy, and Adaptive Preferences." *Philosophical Review* 129 (2): 159–210. doi:10.1215/00318108-8012836.

Fischer, Alexander. 2022. "Manipulation and the Affective Realm of Social Media." In *The Philosophy of Online Manipulation*, edited by Jongepier, F. and Klenk, M., 327–352. New York: Routledge.

Frankfurt, Harry G. 1971. "Freedom of the Will and the Concept of a Person." *Journal of Philosophy* 68 (1): 5. doi:10.2307/2024717.

Fricker, Miranda. 2011. *Epistemic Injustice: Power and the Ethics of Knowing*. Repr. Oxford: Oxford University Press.

Gigerenzer, Gerd. 2008. "Why Heuristics Work." *Perspectives on Psychological Science* 3 (1): 20–29. doi:10.1111/j.1745-6916.2008.00058.x.

Gorin, Moti. 2022. "Gamification, Manipulation, and Domination." In *The Philosophy of Online Manipulation*, edited by Jongepier, F. and Klenk, M., 199–215. New York: Routledge.

Gorin, Moti. 2014. "Do Manipulators Always Threaten Rationality?" *American Philosophical Quarterly* 51 (1): 51–61. Accessed June 04, 2019.

Grinshteyn, Erin, and David Hemenway. 2016. "Violent Death Rates: The US Compared with Other High-Income OECD Countries, 2010." *The American Journal of Medicine* 129 (3): 266–73. doi:10.1016/j.amjmed.2015.10.025.

Hacking, Ian. 1999. *The Social Construction of What?* Cambridge, MA: Harvard University Press.

Jongepier, Fleur, and Michael Klenk. in 2022 a. "Online Manipulation: Charting the field." In *The Philosophy of Online Manipulation*, edited by Jongepier, F. and Klenk, M., 14–48. New York: Routledge.

Jongepier, Fleur, and Michael Klenk, eds. in 2022. *The Philosophy of Online Manipulation*. New York, NY: Routledge.

Jongepier, Fleur, and J. W. Wieland. 2022. "Microtargeting People as a Mere Means." In *The Philosophy of Online Manipulation*, edited by Jongepier, F. and Klenk, M., 156–179. New York: Routledge.

Kahneman, Daniel. 2012. *Thinking, Fast and Slow*. London: Penguin.

Keane, Webb. 2016. *Ethical Life: Its Natural and Social Histories*. Princeton, NJ: Princeton University Press.

Klenk, Michael. 2020a. "Digital Well-Being and Manipulation Online." In *Ethics of Digital Well-Being: A Multidisciplinary Perspective*, edited by Christopher Burr and Luciano Floridi. Cham: Springer. Accessed November 17, 2019. 81–100. doi: 10.1007/978-3-030-50585-1_4.

Klenk, Michael. 2020b. "How Do Technological Artefacts Embody Moral Values?" *Philosophy & Technology*, 1–20. doi:10.1007/s13347-020-00401-y.

Klenk, Michael. 2021a. "Interpersonal Manipulation." *SSRN Electronic Journal*. doi:10.2139/ssrn.3859178.

Klenk, Michael. 2021b. "Manipulation (Online): Sometimes Hidden, Always Careless." *Review of Social Economy*. 80: 1, 85–105. doi:10.1080/00346764.2021.1894350.

Klenk, Michael, and Jeff Hancock. 2019. "Autonomy and Online Manipulation." *Internet Policy Review*. Accessed February 28, 2020. https://policyreview.info/articles/news/autonomy-and-online-manipulation/1431.

Klenk, Michael, and Martin Sand. 2020. "Prometheus' Legacy: Responsibility and Technology." In *Welche Technik?* edited by Birgit Recki, 23–40. Dresden: Text & Dialog.

Klenk, Michael, Sunny Xun Liu, and Jeff Hancock. 2021. *Pulling the Rug from under the Tech-lash: Online Influences are Perceived to be More Manipulative than Similar Offline Influences*. Under review.

Kramer, A. D. I., J. E. Guillory, and Jeff Hancock. 2014. "Experimental Evidence of Massive-scale Emotional Contagion Through Social Networks." *Proceedings of the National Academy of Sciences* 111: 8788–90.

Liao, Shen-yi, and Bryce Huebner. 2021. "Oppressive Things." *Philosophy and Phenomenological Research* 103 (1): 92–113. doi:10.1111/phpr.12701.

Mitchell, Polly. 2018. "Adaptive Preferences, Adapted Preferences." *Mind* 127 (508): 1003–25. doi:10.1093/mind/fzy020.

Mori, Masahiro, Karl MacDorman, and Norri Kageki. 2012. "The Uncanny Valley." *IEEE Robotics & Automation Magazine* 19 (2): 98–100. doi:10.1109/MRA.2012.2192811.

Noggle, Robert. 1996. "Manipulative Actions: A Conceptual and Moral Analysis." *American Philosophical Quarterly* 33 (1): 43–55.

Nussbaum, Martha C. 2001. *Women and Human Development: The Capabilities Approach*. Cambridge: Cambridge University Press.

Nyholm, Sven. 2022. "Technological Manipulation and Threats to Meaning in Life." In *The Philosophy of Online Manipulation*, edited by Jongepier, F. and Klenk, M., 235–252. New York: Routledge.

Pepp, Jessica, Rachel Sterken, Matthew McKeever, and Eliot Michaelson. 2022. "Manipulative Machines." In *The Philosophy of Online Manipulation*, edited by Jongepier, F. and Klenk, M., 91–107. New York: Routledge.

Rudinow, Joel. 1978. "Manipulation." *Ethics* 88 (4): 338–47. doi:10.1086/292086.

Sharkey, Amanda, and Noel Sharkey. 2020. "We Need to Talk about Deception in Social Robotics!" *Ethics and Information Technology*. doi:10.1007/s10676-020-09573-9.

Sober, Elliott. 2015. *Ockham's Razors: A User's Manual*. Cambridge: Cambridge University Press.

Sunstein, Cass R. 2016. *The Ethics of Influence: Government in the Age of Behavioral Science*. Cambridge: Cambridge University Press.

Wildman, Nathan, Natascha Rietdijk, and Alfred Archer. 2022. "Online Affective Manipulation." In *The Philosophy of Online Manipulation*, edited by Jongepier, F. and Klenk, M., 311–326. New York: Routledge.

Wilkinson, T. M. 2013. "Nudging and Manipulation." *Political Studies* 61 (2): 341–55. doi:10.1111/j.1467-9248.2012.00974.x.

Winner, Langdon. 1980. "Do Artifacts Have Politics?" *Daedalus* 109 (1): 121–36.