Online manipulation

Charting the field

Jongepier, Fleur; Klenk, M.B.O.T.

DOI
10.4324/9781003205425-3

Publication date
2022

Document Version
Final published version

Published in
The Philosophy of Online Manipulation

2 Online manipulation
Charting the field

Fleur Jongepier and Michael Klenk

1 Introduction

When we introduced the main research questions and the contributions of this volume in the previous chapter, we touched upon two broad and fundamental topics. First, what is manipulation? Second, is online manipulation simply "regular" manipulation gone online or a new phenomenon? In this chapter, we tackle both questions and chart the overall terrain of online manipulation, critically considering existing and new possible answers to these questions. Our aim is to provide a conceptual map to the reader and allow them to locate the contributions in this volume on it.[1]

2 Three preliminary questions

In this section, we introduce and discuss three important preliminary questions concerning the study of manipulation. First, what is a good method to do study (online) manipulation and how can we gauge its success? We start with this question because it concerns fairly general points about philosophical methodology that are important to studying online manipulation. It has been pointed out that "manipulation" refers to a number of different phenomena, not all of which overlap in their interesting features (Cave 2007), which puts pressure on the question of how we should go about analysing manipulation, if such a thing can even be done. The subsequent two questions involve asking whether "manipulation" is a *thick* concept (2.2) and whether manipulation is necessarily intentional (2.3).

Though our discussion is critical – mentioning problems and worries where applicable – our aim in this chapter is not to argue for any particular answer to any of these questions. Rather, we want to chart the field and bring to the surface not just which positions are out there but also which challenges or worries one is likely to face when adopting them.

2.1 Method

How should we go about the study of manipulation? More specifically, is conceptual analysis a promising method for the study of manipulation?

DOI: 10.4324/9781003205425-3

Very roughly, conceptual analysis seeks to decompose a concept into its constituent parts.[2] A common and influential interpretation of that method has been to provide an explicit intension that is measured against the intuitive extension (the set of all things to which the concept applies) of a given concept (Queloz 2021, 23). This would lead to specifications of the necessary and sufficient conditions for the correct employment of a concept. We would have mastered a concept at the point where we can say whether the concept applies in any situation, and the criterion for application (e.g. "x is a G") is the concept's intension (Queloz 2021, 25).

Most of the existing philosophical work on manipulation proceeds by conceptual analysis and, therefore, it is worthwhile to enquire about its pedigree (cf. Coons and Weber 2014, 6).[3] The method of cases exemplifies this strand of conceptual analysis, whereby a proposed set of necessary and sufficient conditions is tested by considering (hypothetical) cases to see if the proposed conditions correctly qualify something as manipulation.

There are several reasons to be sceptical about conceptual analysis. Some of these reasons are perfectly general in that they pertain to the viability of conceptual analysis across the board. Conceptual analysis as understood here relies on assumptions about the nature of concepts that come from the classical theory of concepts. According to the classical theory of concepts, a concept like manipulation has a definitional structure that is composed of simpler concepts that express necessary and sufficient conditions for falling under the concept or qualifying as manipulation. The truth of the classical theory of concepts is presumed once we embark on conceptual analysis as interpreted here. But if the classical theory of concepts suffers from problems, then conceptual analysis – as understood here – would also be a method of doubtful pedigree. Existing worries about the classical theory of concepts that carryover to the study of manipulation for instance includes the worry regarding the very existence of conceptual essences that conceptual analysis aims to reveal.

A second challenge about conceptual analysis and studying manipulation more generally comes from experimental philosophy. There are serious questions about the reliability of our intuitions that arguably are the core "data" for conceptual analysis. In particular, there is a question about the legitimacy of claims to universality derived from the conceptual analyses pondered in philosophy. Conceptual analysis is supposed to uncover *the* meaning of a concept by drawing on "our" intuitions as evidence (cf. Climenhaga 2018). But who is the "we" here? Intuitions may differ temporarily and geographically, and the analyses on offer may reflect the highly idiosyncratic intuitions of philosophers from WEIRD – Western, educated, industrialized, rich, and democratic – societies (cf. Henrich, Heine, and Norenzayan 2010) and thus have limited scope. Experimental philosophy, and psychological research on manipulation more specifically, may alleviate some of these worries by systematically eliciting a more diverse set of intuitions (Knobe and Nichols 2008).[4] At the same time, however,

such experimental approaches need to answer questions about the method's validity and reliability, especially if manipulation turns out to be a technical concept that requires some expertise to grasp (cf. Pölzler 2020).[5] For example, it is not clear to what extent we can rely on survey studies that prompt the intuitions of laypeople about manipulation to make inferences about the nature and value of manipulation.

There is also a challenge more specific to the study of manipulation as pointed out by Coons and Weber (2014). They wonder whether the concept of manipulation – quite independently of general worries about concepts – lacks core features that unify all cases of manipulation. Some scholars go as far as suggesting that manipulation lacks core cases because it is "too varied" (Baron 2003, 37) and some thus proclaim the attempt at a conceptual analysis is a "fruitless endeavour" (Kligman and Culver 1992, 175). We do maintain that there are core cases of manipulation (such as the case of Othello discussed in the Introduction), but we remain open as to whether all of them share a set of necessary and sufficient conditions. The concept of manipulation may exhibit what Alston (1967, 220) calls "combinatorial vagueness", which is present in cases where

> [W]e have a variety of conditions, all of which have something to do with the application of the term, yet are not able to make any sharp discriminations between those combinations of conditions which are, and those which are not, sufficient and/or necessary for application.
>
> (cited in Ackerman 1995, 337)

This is a relevant suggestion because there are several conditions often associated with manipulation (which we discuss in more detail here), and yet it is unclear how many or which of them are strictly necessary and sufficient for manipulation (cf. Coons and Weber 2014, 7).[6]

The overall worry here is that a concept like manipulation may simply evade analysis (even if the classical theory of concepts is true), just like the concepts "disability" in law or "species" in science, or indeed concepts like "love" or "consciousness". Concepts that allow for borderline cases may evade successful discovery of necessary conditions. The attempt to boil them down to their highest common factor by conceptual analysis may be the wrong approach to take. There will be counterexamples to almost any interesting intension, as any feature that is not strictly a necessary condition will eventually fall prey to counterexamples. This may leave us, at best, with an analysis that is too thin to be interesting and informative (cf. Queloz 2021, 25).

Arguably, the study of manipulation does not stand or fall with the propensity of the concept "manipulation" to bend to complete analysis in terms of necessary and sufficient conditions. Manipulation, though perhaps vague, varied, and beset with borderline cases, may yet be unified by Wittgensteinian family resemblance, that is, not a set of shared properties but a

resemblance to paradigm cases of manipulation. Borderline cases would be those where the resemblance is not clear or not strong enough (cf. Coons and Weber 2014, 6). Assuming that there are some paradigm cases, and many grey areas, we can still usefully study manipulation.[7] For instance, it would be interesting to say just what the paradigms have in common and how they unify the other cases of manipulation. And even if there are no paradigms at all, there may be a focal core of the concept that we can study. To illustrate, with respect to the complex concept of an "epiphany", Sophie Grace Chappell helpfully describes the notion of a focal-case concept in the following passage, here replaced by the notion of "manipulation" (Chappell 2019, 97):

> There are clear and central cases of [manipulation]. . . . But there are also less clear and less central cases, which we might still want to call [cases of manipulation]; or there again, might not. Nothing much turns on where *exactly* we draw the boundaries around the proper use of the term "[manipulation]". The central territory of the concept is not threatened by minor demarcation disputes about its borders. There are certainly grey areas, and they certainly have their interest. There are equally certainly *non*-grey areas: for instance, the black ones and the white ones. . . . True, there are no non-stipulative necessary and sufficient conditions for something's being an [instance of manipulation]. . . . There are no non-stipulative necessary and sufficient conditions for something's being a mountain, either, and the category of the mountainous typically fades out around its edges into literally small-scale phenomena. That does not stop the geologist from studying mountains, nor the alpinist from climbing them.

The view that manipulation might exhibit some kind of vagueness, admit borderline cases, and lack a clear conceptual core would also have a noteworthy moral implication, for it may well make moral evaluations of manipulation more difficult. If there is no necessary condition common to all cases of manipulation, there cannot be the *same* moral reason against all cases of manipulation because there is no necessary element shared by all manipulative acts. More sceptical approaches about finding any unity, however, may also be positive, as many authors in this volume illustrate, as it may also help our understanding as to why some but not all cases of manipulation are morally problematic. Manipulation may be anything that resembles doing x, y, or z and so we might investigate the moral status of x, y, and z and find differing verdicts (sulking to get your way is always bad, but comforting your friend is ok, though both are, arguably, manipulative). Sometimes, it can be more useful to get a better view on the overall ballpark, as it were, even if the ballpark has a few items that shouldn't be there than having the clearest view on one item in it.

Of course, various methodological approaches will ideally be operative in any concept-heavy debate such as the (online) manipulation debate; and indeed, this volume is itself an illustration of methodological diversity. The central aim and conclusion following from this brief methodological discussion therefore is not that scholars try and work towards methodological consensus and agree on a shared and unified methodology within the online manipulation debate. Rather it is to make implicit methodologies *explicit* so as to learn from their differences and respective strengths and weaknesses and to find ways for various methodological approaches to be complementary and fruitfully run parallel, even if they are methodologically at odds.

2.2 Thickness

We turn now to the question of whether the concept picked out by the word "manipulation" is a thick or moralized concept and of whether and how manipulation depends on intentions (2.3).

Thick normative or thick evaluative concepts have both a significant degree of descriptive content and are normatively loaded. The concept "kindness", for example, may denote descriptive qualities like being self-less, helpful, and caring towards other people. At the same time, characterizing someone as kind typically involves expressing a pro-attitude towards the person or their behaviour and thus an evaluative statement as well. Being kind is being caring towards others, *and that is a good thing*. If "manipulation" is a thick concept, then it also has a significant degree of descriptive content as well as being normatively loaded. It would not only be a particular type of influence (assuming that this is what manipulation is) but it would be a particularly *good* or *bad* – and not normatively *neutral* – type of influence.

Another way to express the thought that manipulation is a thick concept is to say that manipulation may have a normative or evaluative status as a conceptual matter. Ackerman (1995), for instance, notes that several of the features commonly associated with manipulation – such as deceptiveness or using others for one's own benefit – are prima facie immoral. Grasping the concept would thus involve grasping a particular normative or evaluative status. Just like grasping kindness is to understand that being kind is *good*, grasping manipulation may be to understand that it is *bad*. If that were the case, then any analysis of manipulation would have to involve an account of its descriptive content as well as an account of its normative or evaluative content. The analysis need not involve two separate steps, of course. For example, if manipulation is analysed as deceptive influence, then it has both a descriptive component (influence of some sort) and a normative component already built into the concept of deception.[8] Importantly, whether or not manipulation is a thick or moralized concept is largely independent of the question of whether it can sometimes be permissible to manipulate. Manipulation may be, say, morally bad as a conceptual matter. But one

might still maintain from a moral philosophical standpoint that it is merely a pro tanto wrong, which can be outweighed by other factors (e.g. beneficial consequences) (Baron 2003).[9]

Numerous considerations challenge the idea that a moral dimension is part of the concept of manipulation. As a first approximation of that point, consider that the word "manipulation" – and thus presumably the concept expressed by the word – is also appropriately used in ways that are clearly morally neutral. We speak of manipulating inanimate objects like sticks and stones, and we manipulate research subjects in experiments that are cleared by the research ethics board. Thus, there are instances of manipulation that do not appear to carry a specific normative or evaluative judgement with them. To adopt the "thick" reading thus involves explaining why and how manipulation within scientific studies is morally wrong.

The word "manipulation" may express different concepts, and defenders of the thickness of manipulation may claim that there is a distinct thick concept of interpersonal manipulation after all. Accordingly, how we use the word manipulation and the corresponding concepts in cases that do not relate to interpersonal interaction may be beside the point. Still, we can find examples where manipulation of persons is referred to in a morally neutral or even laudatory way. Allan Wood (2014) gives the example of a politician who silences a heckler at a political rally through skillful manipulation, rather than resorting to brute force by calling for security to remove the heckler. We might applaud and praise the politician for this action. It also seems that artists who seek to create a certain effect in us, or politicians who aim for structural reforms, may sometimes do so by means of manipulation and still be applauded for it. Especially the case of the artist may prompt us to consider that manipulation may sometimes be not even pro tanto wrong. Hence, there might be examples of manipulation being appropriately used in a normatively and evaluatively neutral or even positive way. This speaks against the thickness of manipulation, unless we can reasonably maintain that there is at least some pro tanto wrongness associated with manipulation in all of these cases, or if one works out why contrary to appearances this is morally wrong overall and not only in a pro tanto way.[10]

There is also a more general consideration that speaks against the thickness of manipulation. Allan Wood (2014) points out how we use manipulation in the course of moral explanation. For example, when someone enquires what exactly is morally problematic about (aspects of) social media, some may offer as an explanation that social media can be manipulative. Such an explanation would seem entirely reasonable and informative. However, if manipulation is a bad or immoral type of influence as a conceptual matter, then that explanation would lose some of its force. There will be descriptive information conceived in virtue of the descriptive content of the concept of information, but it will not be illuminating in a normative sense because the negative evaluation would be a matter of conceptual course. An additional normative explanation would be superfluous.[11]

A final methodological possibility to consider, which draws on the pragmatic methodology discussed in the previous section, is to grant what might be the main worry with respect to manipulation as a thick concept, namely that sometimes manipulation is harmless or even good and so to allow that manipulation isn't *always* bad or immoral and certainly not *necessarily* so, but then to go on and say: but it is, most of the time. The strategy is thus to agree that, *strictly speaking*, manipulation isn't a thick concept but then to suggest perhaps we should not talk so strictly. This pragmatic approach won't satisfy all philosophers, perhaps, but it might help the online manipulation debate move forward.

2.3 Intentionality

A third preliminary question concerns the relation of manipulation and intentions. Manipulation is almost always portrayed as requiring intentionality on the part of the manipulator. Marcia Baron gives the insightful example of the following apology, which seems strange: "I am so sorry that I manipulated you [treated you manipulatively]. I didn't mean to; I didn't realize I was manipulating you, and I never would have acted as I did had I known" (Baron 2014, 102). The reason this apology is strange, Baron suggests, is because it suggests manipulation can be unintentional. Instead, she and many others claim that manipulators must be capable of having or forming intentions and acting intentionally. Thus, at least on the standard conception of agency, manipulators must be *agents*. We can call this the *general intentionality requirement* for manipulation.

The general intentionality requirement is extensionally plausible because typical cases of manipulation indeed involve agents who perform manipulative actions. Moreover, it looks like manipulation may always be a reason for blame or praise. Since the latter is often thought to be applicable only in cases where we deal with subjects capable of forming intentions, these normative practices related to manipulation support the general intentionality requirement for manipulation.

The general intentionality requirement is particularly interesting in the context of this volume because we will be looking at the relation of manipulation and online technology. If technology, whatever it is, cannot be intentional, then any contribution that technology makes to a manipulative act may seem at best purely instrumental to a real (i.e., human) agent. A manipulative act, perpetuated by an individual or group agent, may turn out to be more effective, more consequential, or, as we dub it in Section 5, "aggravated" in some sense because of the use of technological artefacts. For example, real-time profiling on the web may allow manipulators to wield more powerful influences. We could on such a view allow that technology has a meaningful influence on the agent's choice and behaviour (Klenk 2020), which may change our normative assessment of the situation and the warranted political or legal repercussions (e.g. by partly excusing

manipulators). But given the general intentionality requirement, the contribution of non-intentional technology should not change our assessment of the situation as manipulative.

Apart from broader intentionality requirements, some theorists further advocate a *specific intentionality requirement* for manipulation. Rather than requiring intentionality in the aforementioned, general sense, manipulative action may also require intentions with a particular content (Noggle 1996; Gorin 2014b). These intentions, in turn, could be either topical in the sense that they must be intentions to manipulate or more generally about something else. Perhaps, a manipulator must *intend to x*, where *x* is whatever set of necessary conditions manipulation might have. Several scholars have suggested that there is such a specific intentionality requirement for manipulation. Robert Noggle, for example, argues that manipulators *intend* to have their victims violate some norm that regulates belief, desire, or emotion (1996). Others think this condition is too strong, and that the act of manipulation requires only the fact that people "could have done otherwise", not that they had the explicit intention to violate specific norms. Some theorists, like Kate Manne (2014), go very light on the intentions and instead argue that there can even be something like unwitting manipulation.

The discussion of the intentionality requirements for manipulation nicely leads to a discussion of the conditions of manipulation which we outline in the next section that deals with the demarcating factors of (online) manipulation. Both the general and specific intentionality requirements may be bona fide necessary conditions for manipulation. Still, we do not discuss them as demarcating factors for two reasons. First, as will become apparent, the search for a plausible account of manipulation is often the search for conditions that distinguish manipulation from coercion and persuasion. The intentionality requirement would presumably cut across this discussion. That is, whatever we say about intentionality requirements for manipulation will presumably apply to coercion as well.[12]

3 Manipulation and the search for demarcating factors

In this section, we will introduce and review recent analyses of manipulation. We propose to understand recent work on manipulation as the search for *descriptive demarcating factors* that distinguish manipulative from other types of interpersonal influence.

3.1 The demarcation problem for manipulation

The "demarcation problem" for manipulation is the problem of giving an account of manipulation that demarcates it from neighbouring forms of social influence such as persuasion and coercion (cf. Klenk 2021b). The demarcation problem thus prompts us to say *how* manipulative influence

can be described as a distinct form of social influence, with particular wrong-making features.

Like coercion and persuasion, manipulation is a kind of interpersonal or social *influence* (Coons and Weber 2014, 8). It is widely held to be characteristically distinct from coercion and persuasion in kind or degree (cf. Faden and Beauchamp 1986). But how, precisely?

We should first note the close proximity of manipulation and the typical effects of coercion in terms of autonomy loss and blame-related practices. Wood (2014), for instance, suggests that manipulation is a type of influence on a continuum with coercion, with the latter being more heavy-handed than manipulation. Manipulation "influences choice without quite removing it" (Wood 2014, 26). Similarly, Baron (2003, 42) suggests that manipulation may become so strong so as to be indistinguishable from coercion at some point.[13] Greenspan (2003) suggests that manipulation "seems to have a foot in both the usual categories of intentional interference's in another agent's autonomy, coercion and deception" but is unlike both. Unlike being coerced, being manipulated supposedly never entails being a fully passive victim or instrument. Some autonomy is retained. Likewise, Alm (2015, 256) suggests that manipulatees have "whatever type of control is needed for responsibility". Hence, the manipulated person still *does* something *voluntarily* (Coons and Weber 2014, 8). All the same, manipulation is often, though perhaps not always, seen as antithetical to autonomy (we discuss autonomy violation as demarcating factor later) and some suggest that manipulation implies autonomy loss (Susser, Roessler, and Nissenbaum 2019a, 2019b).[14]

This debate about the autonomy- and blame-related *effects* of manipulation is partly informed by the debate about incompatibilism, and many philosophers (Kane 1996; Pereboom 2001) suggest that an agent in "manipulation cases" is not free, though he or she acts on her own volition (Sripada 2012).[15] The examples of manipulation discussed in that debate are usually much crasser (think of neurological, deterministic interference with people's choices) than the ordinary cases of manipulation that we are concerned with in this volume. The debate is illustrative nonetheless as it suggests how manipulation, understood as detrimental to freedom (of choice), need not undermine volition or autonomy.

Nudging helpfully illustrates the proximity of coercion and manipulation. Nudges influence choices without removing them. Their apparent non-coercive influence is why some consider nudges as morally unproblematic (Thaler & Sunstein, 2008). At the same time, it appears to others that nudges are still a problematic way to influence people, partly because they seem to structure choices in worrisome ways, some of which may lower or hamper autonomy (e.g. Levy 2019). Some types of nudging may thus appear as a paradigmatic way to manipulate people: arguably, they do not remove autonomy, but they may hamper it by structuring choices and thus guiding

people's decisions (cf. Sunstein 2016a). Both manipulation and nudging are typically described as being forms of non-coercive influence (Schmidt and Engelen 2020).

However, one must say more than that, because the negative definition "influence that is not coercive" is not very illuminating. One reason is that this strategy relies on a clear account of the notions of coercion and persuasion to begin with. However, neither coercion nor persuasion is very well understood as a type of influence. There is obviously a tremendous amount of work on coercion, but a lot of it concentrates on characterizing coerced actions and, in particular, their effect on blameworthiness and accountability.[16] Another reason is that we need to find a set of conditions that individually or jointly applies to manipulation but not to coercion or persuasion to solve the demarcation problem for manipulation.

To illustrate the problem, consider the view that, like coercion, manipulation removes, nullifies, or threatens autonomy and, unlike coercion, it operates covertly (e.g. Susser, Roessler, and Nissenbaum 2019b). The suggestion that it operates covertly may, on this view, demarcate manipulation from coercion. But if that is denied, and we will discuss this in the following, then we lose our handle on the distinction between coercion and manipulation.

We can now turn to our search for demarcating factors. We present three families of views that tackle the demarcation problem for manipulation. On what we will call *outcome* views, manipulation requires a particular outcome. On *process* views, manipulation requires a particular process of influence. On *norm* views, manipulation requires the violation of particular norms.

3.2 Outcome views

On outcome views of manipulation, manipulation always, or at least typically, directly, or indirectly, leads to actions or behaviours with particular features. We will discuss just two types of outcome views, according to which manipulation leads to harm or the violation of self-interest or to a loss of autonomy.

3.2.1 Self-interest and harm

Manipulative influences typically go against the interest of the person being manipulated. That is, they lead to outcomes that are directly or indirectly unbeneficial or outright harmful for the person being influenced. Direct outcomes of manipulative influence may include beliefs, emotions, or desires formed by the manipulated person. It is often in one's self-interest to form true beliefs, and to have appropriate emotions, and worthy desires. Manipulation may directly frustrate these. Indirectly, your (false) beliefs or (inappropriate) emotions may lead you to do things that frustrate your self-interest. You may vote for the wrong candidate, buy

the product you do not need at a price that is much too high, or stay at the slot machine for hours on end. The frustration of self-interest is thus often linked to harm, and manipulation may be said to involve harm to the manipulatee. Many paradigmatic cases of manipulation feature frustrations of self-interest on the part of manipulated persons. For example, our introductory case of Othello suffers great harm as a result of the manipulation (he ends up killing his beloved Desdemona). Typical cases of manipulation online, such as voter manipulation or endless doomscrolling, go against self-interest, too.

However, frustration of self-interest and harm is unlikely to be a necessary feature of manipulation. Nudging, at least in some forms, seems to be manipulative. Nudging, at least in some forms, seems to be manipulative, and many such nudges are meant to serve the self-interest of the nudged persons.[17] In many paradigmatic cases it is actually meant to *serve* people's self-interest. Similarly, romantic love is not against self-interest, and yet it is sometimes considered to involve manipulation, especially at the early stages. For example, you may manipulate by presenting yourself better than you actually are or by flattering the other person. When we understand these manipulative manoeuvres as integral parts of romantic relationships and also consider the latter to be unproblematic or even fun, then it becomes problematic to accept the necessity manipulation as always going against self-interest.

Both nudging and romantic love may thus be counterexamples to the view that the frustration of self-interest is a necessary ingredient of manipulation. There is room to argue that these counterexamples are inconclusive. For example, the very act of influence may directly and instantaneously be against self-interest (e.g., there may have been better ways to influence one in a nudging or love relationship) while the situation may be all things considered good for the manipulatee. Moreover, one might have concerns with the method of cases that the aforementioned strategy of showing that manipulation can improve self-interest, relies on.

There are two more general problems with the self-interest proposal, though. First, we are looking for a demarcating factor to distinguish manipulation as a non-coercive type of influence. Coercion typically frustrates self-interest, as least in the minimal sense that a different type of influence may often be better for the person being influenced. So, frustration of self-interest and harm will also be an ingredient of coercion. It does not help us to demarcate manipulation from coercion. At best, such a theory of manipulation would be incomplete.

Second, manipulation is unlikely to be exhaustively characterized by any end state (i.e., the direct or indirect result of the influence) and should at least include features of the process through which manipulation occurs. The reason is that end states like having one's self-interest frustrated may be arrived at in a multitude of ways, and not all of them will be manipulative.[18] This is a more general formulation of the demarcation to coercion. Many

types of influences or events may frustrate people's self-interest. If we want to single out manipulation, we have to find further demarcating factors.

3.2.2 Autonomy

Manipulation as autonomy undermining is an account that shares with the self-interest account a focus on the (direct or indirect) end result of manipulation. It is not always clear whether proposals that link manipulation and autonomy are attempts to spell out its wrong-making features or attempts to give an account of manipulation in descriptive terms (insofar as autonomy can be understood descriptively). As noted earlier, there is surely a close relation between manipulation and autonomy. But how plausible is it that the undermining of autonomy is a necessary feature of manipulative interaction? Paradigmatic examples of manipulation indeed often seem to deprive agents of their autonomy. But, again, that need not imply that the undermining of autonomy is a necessary criterion of manipulation. Manipulation need not interfere with autonomy (Blumenthal-Barby 2012) and may even enhance it (Buss 2005; Gorin 2014b; Klenk and Hancock 2019); for example, when manipulative influence allows you to reach your goals and bring your desires and urges in line with your higher-order volition and desires, as in Harry Frankfurt's classic account.

Notice that autonomy may be lost by means other than manipulation, and so the loss of autonomy is not sufficient for manipulation. The counterexamples to the conceptual link between autonomy and manipulation in the literature suggest that it is not necessary, either. But even if it would be necessary for manipulation, we would need to say more about the nature of manipulation to distinguish it from coercion, in the context of the demarcation problem. As noted earlier, the *outcome* of coercion is *also* less autonomy. So, the autonomy view does not, without further explanation (e.g., distinguishing different types of autonomy), seem sufficient to demarcate manipulation. Of course, whether we can bolster the account by identifying further factors in addition to autonomy loss that, together, demarcate manipulation from coercion remains to be seen.

3.3 Process views

Process views of manipulation interpret manipulation in terms of characteristic processes or modes of influence that lead to a given behaviour or action.

3.3.1 Covert influence

Covert or hidden influence has often been suggested as a defining feature of manipulation in the sense of being a typical (e.g., Baron 2003; Rudinow 1978) or even a necessary condition for manipulation (Susser, Roessler, and

Nissenbaum 2019a; Handelman 2009). This is a plausible proposal because both coercion and rational persuasion take place "out in the open." Persuaders need to get their interlocutors to "see" their reasons for acting, and so do coercers. Manipulators, in contrast, seem to operate undercover. Iago, for example, also deceives Othello and his plan would not succeed if Othello would know what is going on. It seems very plausible that, in order to succeed, manipulation must be hidden in the sense that the intentions of the manipulator, the process of influence, or direct or indirect target outcome remain hidden from the manipulatee.[19] And indeed also in the online manipulation debate, the view that manipulation must be hidden is popular (cf. Susser et al. 2019a).

However, it can be argued that covert or hidden influence is not a necessary condition for manipulation. Again, there are several counterexamples (Gorin 2014b; Krstić and Saville 2019; Barnhill 2014; Klenk 2021b). For example, manipulative guilt trips can be obvious and still be very effective. We can be lucidly aware that we're being manipulated into feeling guilt, even as we feel guilt and act on it (Barnhill 2014, 58).

Counterexamples to the covertness view purport to depict manipulative influences that are not hidden from the manipulatee. With respect to online manipulation, one might wonder whether, say, the manipulation as conducted by Netflix's auto-play or Facebook's newsfeed is going on *non-transparently*. Do internet users in the twenty-first century, post-Cambridge Analytica not know they are being manipulated after all?

Relatedly, it would seem perfectly appropriate to complain about manipulative influence. For example, you may be surveilled and be annoyed by the obviously manipulative attempt of some marketeer to get you to buy a product that you do not need. We've all been irritated by advertisements of things we bought the day before and by targeted ads for camping gear that appear on our screens just after we watched the odd outdoor documentary on Netflix. If manipulation were hidden by definition, our frustration at these ostensibly manipulative influences would betray a conceptual mistake. After all, given that we were aware of the influence, it cannot be classified as a manipulative influence (if the covertness view is true). Clearly, that is not the case.

Again, there is room to resist this conclusion. One can challenge the counterexamples. For instance, it may be argued that what seems like overt manipulative influence that takes place out in the open is actually coercion. Guilt trips, pressuring tactics, and perhaps highly advanced and emotionally salient targeting online may thus fall under coercive and not manipulative influences. This is an attractive answer if we hold on to the view that manipulation and coercion are only gradually distinct. Then there are several ways to refine the thesis. Proponents of the covertness view could, for example, distinguish between covertness being a feature of the influence (i.e., it is actually hidden from the manipulatee) or merely an intended feature of the influence (i.e., the manipulatory intends for the influence to be covert) that

need not be actualized. Netflix and Facebook, the twenty-first century notwithstanding, are still trying to manipulate precisely because they are trying to keep their cunning influence hidden. The task for the defender of the covertness thesis is thus to show exactly what would remain (truly) hidden in manipulative influence. Alternatively, the covertness thesis advocate might want to distinguish between different types of knowing or being aware of being manipulated: one might know, in some "cognitive" sense that one is being manipulated by Facebook (or one's first date), but one still fails to know, in a different sense (whilst being wholly engaged online or enthralled by a date) that one is being manipulated. We will discuss covertness and transparency in some more detail in the following.

3.3.2 Bypassing rationality

Another possible demarcating factor of manipulation is the bypassing of reason (e.g., Noggle 1996; Scanlon 1998). The intuitive idea is that manipulation is a type of influence that does not (adequately) engage the victim's rational capacities (e.g., Sunstein 2016b).[20] It is important to be clear in spelling out what it takes to "bypass reason" and, according to Gorin (2014a), one can understand such accounts in several ways.

One way is to interpret manipulation as actively interfering with rational capacities in the sense that one generates psychological states that are "incompatible with the proper functioning of the person's rational capacities" (Gorin 2014a, 53). Alternatively, one may understand manipulative influence as bypassing rationality in the sense that one impedes the rational capacities of one's victim from functioning, where their functioning can be understood "narrowly" in terms of functioning given the information, beliefs, and preferences available to the agent or "broadly" in terms of functioning given whatever reasons there objectively are (Gorin 2014a, 54–57).

The bypassing-reason view explains well many paradigmatic cases of manipulation. Charming, using olfactory and visual influences, using someone's emotional outbursts, or playing on their jealousy (as in our introductory example of Othello) all seem like paradigmatic cases of manipulation that also seem to bypass the rational capacities of the victim, at least on some interpretations of "bypassing rationality" explicated earlier. For example, charming tactics may impede the proper functioning of your rational capacities by preventing them from picking up the reasons against giving in to your suitor. Many of the phenomena that give rise to a worry about online manipulation such as increasing polarization also seem to drive on emotional and often irrational tendencies of users, for example, a bias in favour of one's in-group.[21]

We can immediately see how the bypassing reason account would help to address the demarcation problem. It is an account that focuses on the

process of influence, rather than the end result. And we noted earlier, how persuasion and coercion require that victims recognize and act on reasons to succeed. Hence, the bypassing-reason criterion is a promising one to resolve the demarcation problem.

However, Gorin (2014a) and several others have documented at length how manipulation can sometimes proceed precisely by exploiting rational facilities (Klenk 2021a; Barnhill 2016). For example, consider a politician, convinced of the rationality of their voters, who finds that voters are very much concerned with saving the environment. The politician proceeds to give good arguments for the protection of the environment, and she is voted into office. The politician herself, however, does not care about the environment herself at all (Gorin 2014b, 91). This seems to be a case of manipulation: she uses voters purely instrumentally. However, it is false in this case that the manipulator aims to make the manipulatees fall short of the ideals that govern their emotions or beliefs, respectively. For example, it is reasonable to accept good arguments for a true conclusion, if anything is.

Moreover, the idea that manipulative acts proceed through some specific pathways – in this case, the process of bypassing rationality – is questionable because "the processing route" or "origin" of an idea or mental state is unlikely to be always unequivocally bad. Certain beliefs or emotions may well have resulted from bypassed rationality (e.g., the result of being madly in love or deeply angry), but that doesn't mean these states are necessarily suspect – quite the contrary (cf. Jongepier 2017). Also, Barnhill (2016) makes a convincing case that the bypassing of rationality cannot convincingly be held to consist in using emotional, non-rational influences because the former are also sometimes bona fide ways to engage with the world. More generally, philosophers have long pointed out that emotional ways of responding to the world are *rational* responses; for example, reacting with a negative emotion towards an injustice.

This doesn't mean that accounts according to which something counts as manipulation in case it (minimally) involves bypassing the rationality of persons are doomed to fail. It's still plausible – if we take the case of propaganda, for instance – that debilitating people's capacity to think clearly and instead to dig their heels in emotional responses such as fear is worrisome. The point, rather, is that bypassing accounts need to explain why and when some bypassed states or emotional ways of responding to the world are bona fide processes and what separates those from the *mala fide* types.

3.4 Norm views

We have reviewed the most promising *outcome-* and *process*-oriented accounts of manipulation and seen their advantages and disadvantages. A different and increasingly influential type of account are *norm-based*

views of manipulation. According to norm-based views, manipulation is associated with behaviour or action that violates norms (Scanlon 1998; Barnhill 2014; Noggle 1996, 2018a; Gorin 2014a, 2014b, 2018; Klenk 2020, 2021b; Sunstein 2016a). There are considerable differences as to how the norm violation that constitutes manipulation is understood. For example, Noggle's influential account of manipulation suggests that manipulation is constituted by the attempt to make someone else (the manipulatee) violate a norm, whereas others like Gorin and Klenk suggest that manipulation is constituted by the manipulator violating a norm of proper influence.

Norm-based accounts are promising and influential in the philosophical literature, but they have not received much uptake in the digital ethics literature yet. The unifying thought behind norm-based accounts of manipulation is that we can explicate the concept of manipulation in terms of epistemic, moral, or practical norms that manipulation violates.

The difference between outcome- and process-oriented views, on the one hand, and norm-based views, on the other hand, is subtle. After all, the fact that an action violates a norm may also be a particular *outcome* of a given interaction, just like some types of *processes* may constitute norm violations. What seems to set norm-based views apart is that the norm violation is constitutive of manipulation, rather than a (common or necessary) side effect.

Norm-based views may seem suspect insofar as they would seem to foreclose the debate about the thickness of manipulation. After all, it would seem that an account of manipulation in terms of a norm-violating social influence would imply that manipulation carries with it a normative or evaluative judgement as a conceptual matter. But that conclusion would be premature. First, insofar as we can give a descriptive account of norms (e.g., in terms of social expectations) we need not conclude that a norm-based account of manipulation implies the thickness of manipulation. Moreover, manipulation may turn out to be morally problematic in all cases without that fact being a constituent part of the concept. As mentioned earlier, these two things should be kept apart. Finally, the question very much depends on the details of the norm-based view under consideration. For example, Noggle's view suggests only that manipulative influence is the attempt to get someone else fall short of certain norms. And while there may be pro tanto norms against attempting such a thing, Noggle does not define manipulation in terms of the attempt to violate that norm. This may be a consequential difference to norm-based views like that of Gorin and Klenk, who analyse manipulation as falling short of certain interactional norms. In either case, however, the thickness of the concept need not be associated with a moral one, as manipulation may also be constituted by a violation of epistemic or practical norms, rather than moral ones.[22]

On Noggle's influential view, manipulation involves a violation of norms that pertain to the outcomes of an interaction, such as the behaviour or

action exhibited by the victim of the manipulative influence. According to Noggle (1996, 44):

> There are certain norms or ideals that govern beliefs, desires, and emotions. Manipulative action is the attempt to get someone's beliefs, desires, or emotions to violate these norms, to fall short of these ideals.

For example, Iago intended for his actions to make Othello believe a falsehood (namely, that Desdemona was cheating on him), and thus he intentionally made Othello violate the norm that legislates believing truths.[23] What norms matter, on Noggle's account? The relevant norms or ideals are the ones that the manipulator envisions for the manipulatee. This retains a parallel with deception (where it matters what the deceiver takes to be the truth, from which he deviates), and it avoids the potential problem of committing to and identifying objective norms that govern belief, desires, or emotions. Most proponents of norm views follow Noggle in classifying manipulation as an "intentionally characterised" action (Noggle 1996) and specify it quite broadly in terms of attempting one's victim to violate some belief, desire, or emotion-related norm (see, for example, Barnhill 2014 and Gorin 2014a). In effect, the breadth of the different norms for emotions, beliefs, and desire that we recognize gives the account tremendous breadth and explanatory power. Thus, a norm-based account avoids the mistake of trying to shoehorn manipulation into the mold of necessary violation of some allegedly more basic outcome or process.

However, the norm-based view has problems with counterexamples, too. For instance, pressuring or charming tactics cannot be explained by the view even though they seem like bona fide cases of manipulation (Noggle 2018b). For example, consider emotional blackmail or related pressuring tactics. It would seem pressuring others provides them with good reasons to act. In light of the threat or the pressure to conform to someone else's demands it may make good sense to believe, desire, or feel just as the manipulator wants. In many cases of pressuring, the pressuring itself creates good practical reasons to yield to the threat. Indeed, the reason-generating nature of pressure is what the perpetrator relies on when they utter their threat. There is thus in that sense no violation of a norm. Indeed, it would seem that the manipulator in these cases relies on the manipulatee to be responsive to the reasons he or she provides in the form of pressure or, more generally, a threat (cf. Klenk 2021a). Insofar as using your emotional power over your significant other, (peer) pressuring your colleague into accepting the undesirable task, or seducing your online date is manipulative, the norm-based view cannot explain it. Since such cases appear to be bona fide cases of manipulation that we should want to explain, that is a problem for norm-based views.

Naturally, these counterexamples may be challenged. Perhaps, the norm-based view and its focus on norm violations could be coupled with additional

conditions to account for these cases, such as a violation of self-interest (cf. Barnhill 2014). However, a deeper concern with the view is that it gives undue attention to the intentions of the manipulator as they concern the manipulatee. Noggle, for instance, suggests that manipulation is constituted by the attempt to make someone else fall short of norms that govern belief, emotion, or desire. Why make the demanding assumption that manipulators aim to have their victims violate a norm, rather than merely assuming that they influence their victims in a way that constitutes or results in a norm-violation?

A variant of the norm-based view that seeks to address this concern is the view that manipulation is negligent influence (Klenk 2020, 2021a). The negligence account is motivated by two problems. First, the aforementioned counterexamples to existing norm-based views of manipulation and the desire to account for these examples as manipulative influence. Second, the observation that these examples can be accounted for on normative terms only at the expense of introducing a proliferation in the type and scope of norms that manipulation violates as a constitutive matter. For example, Noggle's view could account for pressure cases by suggesting that manipulation is constituted by the violation of interactional norms that, amongst other things, imply that pressuring is prohibited. In effect, rather than just considering norms as they supposedly apply to the manipulatee, norm-based accounts would also have to invoke norms as they apply to the manipulator.

The core proposal of the negligence account is to suggest that the latter suffice to satisfactorily account for manipulative influence. Manipulators uniformly seem negligent regarding their chosen means of influence. However they influence their victims, their choice of influence is arguably not best explained by its "reason-revealingness" (to wit, its propensity to reveal reasons to the influenced person) but by its effectiveness in getting people to do what the manipulator wants. This kind of negligence is proposed as the common factor that unifies all cases of manipulation (Klenk 2021a). Marcia Baron suggests a similar line of thought when she writes a manipulator has "the aim of getting the other person do what one wants, together with recklessness in the way that one goes about reaching that goal" (Baron 2014, 103).

The negligence account would amount to a significant shift in thinking about manipulation. Manipulation would not be demarcated from coercion by what it does or adds to it but by what it lacks. Unlike coercion and persuasion, manipulators do not primarily care for reasons (they sometimes might, when it serves their purpose, but it is not an integral part of their endeavour). Gorin (2014b, in this volume) suggests a view along these lines when he analyses manipulation disjunctively as a violation of at least one of four types of norms, amongst them norms that demand being motivated by someone else's reasonable ends. Like the negligence account, Gorin's view also shifts the domain of norms whose violation constitutes manipulation to norms that apply to the manipulator. The open question is how to spell out

those norms in detail and how many different types of norms are violated by manipulation as a constitutive matter.

In any case, the advantage of a negligence-type of account would be that the distinction to coercion could clearly be maintained because coercers *do* care about reasons but manipulators do not (cf. Schelling 1997). After all, coercers rely on their victims being able to appreciate that they are given good reasons (e.g., a threat to life) to comply with what the coercer wants them to do. A lunatic who cares not about reasons can be harmed, but not coerced.

A problem about the negligence account is that it may complicate matters too much when thinking about manipulation and thus be too far removed from ordinary discourse about manipulation (see Coons and Weber 2014 for related discussion). Also, depending on how the negligence relation is spelled out (to wit, the precise sense in which a manipulator fails to acknowledge or care for reasons), there is a question about whether or not norms or duties of care determine domains where manipulation can occur or whether we should better characterize negligent influences in domains without norms of care as benign forms of manipulation (or not as manipulation at all). Finally, and this will connect to the next section, we can ask whether manipulators need to have the capacity to be governed by an absence of negligence or a presence of norms of care to qualify as manipulators in the first place.

4 Intermediary conclusions

We can draw the following intermediary conclusions. First, we should be careful about the intentionality required for manipulation because it may concern the capacity for intention (what we called the general intentionality requirement) or the specific intention to manipulate or do something associated with manipulation (what we called the specific intentionality requirement).

A second major point is that manipulation is a type of influence that is distinct (in kind or degree) from coercion, and manipulated people still do something *voluntarily*. From this observation, we developed the demarcation challenge which is the challenge to define manipulation in contrast to coercion. Coercion notably has normative implications for (moral) responsibility, and it will be important to determine to what extent manipulation exculpates.

Finally, our discussion brings to the surface an important methodological assumption in the philosophical manipulation debate that is transported easily to the digital ethics debate, namely the anti-pluralist assumption that *one* of the accounts of what manipulation is must be right – not a combination of two or more views. The anti-pluralist assumption makes sense. After all, it's strange to think that in some cases what makes it a case of manipulation is that it involves negligence and in others it's because it involves bypassing rationality. Letting go of the anti-pluralist assumption would thus

come at substantial explanatory costs of explaining how manipulation can be so multifaceted and still say it's manipulation across all cases. But it may not be impossible, especially if one were to adopt a "focal case concept" of manipulation. It also depends a great deal on *why* one wants a definition (or better understanding of) manipulation. Is it for getting a better understanding of digital manipulation? Is it for getting a better view on the harms for internet users or the wrongs of digital manipulators? Is it to develop new policy or legal regulations? Depending on the aims, accepting (a degree of) pluralism or conceptual messiness can range from being highly problematic to potentially productive.

The take-home message for this sub-section about theories of manipulation is thus, above anything else, the need to be explicit first of all about one's preferred theory of manipulation, second about one's methodology, and finally about one's aims.

5 Aggravating factors

Having discussed the relevant philosophical terrain and the rich variety of positions to be taken when it comes to defining manipulation and why it's bad or wrong (if it is), it is now time to look at the "techy" side of things. Which technologies can be considered manipulative or used in manipulative ways by corporations? Which aspects of the existing technologies make them effective manipulative tools (if tools they are)? Which technological advancements are especially worrying from a moral point of view?

These questions are the domain of the field of digital ethics, though they are not only questions in the field of digital ethics. The tech side is a vast territory and is, importantly, interdisciplinary territory. The aforementioned questions have also been addressed – often earlier, in fact – by legal scholars, computer scientists and communication scholars, and many others working on (digital) technologies for whom addressing questions about the manipulative and morally problematic nature of these technologies have been inevitable.

When it comes to studying the manipulative and immoral potential of new technologies, there are different approaches one might take. A common approach taken in the wider digital ethics literature is the "ethics of (insert technology)" approach. There are papers covering, for instance, the ethics of recommender systems, the ethics of algorithms, the ethics of automation, self-driving cars, social robots, voice assistances, and so on. The "ethics of" approach is valuable because each new technology or technological implementation will come with its own technical and moral characteristics. Recommender systems and self-driving cars, for instance, are entirely different, each giving rise to different conceptual and moral questions. It's important not to throw everything on one big pile, since doing so feeds into the already all-too-common slogans that "digital technology" as such is manipulating us and undermining our freedom (cf. Harari 2018).

While the "ethics of x" approach is valuable as well as necessary, it's also important that it is not the only approach on offer within the wider discipline of digital ethics. This is because, despite obvious and deep differences between various new technologies, there will also be important similarities in terms of what makes them especially manipulative and/or morally problematic. It's possible to attend to these shared features without having to make sweeping statements about digital technology in general undermining our freedom *tout court*.

These shared features are what we will refer to as "aggravating factors". An aggravating factor is a factor that sometimes or typically either (a) makes manipulation more effective, its effects worse or morally wrong, or (b) makes it harder for individuals to avoid or contest manipulative practices and technologies. In the following, we discuss what we regard as four noteworthy aggravating factors: personalization, opacity, flow, and lack of user control.

5.1 Personalization

Not just our Google searches and the ads we see online but also the health trackers we wear, the TVs we watch, and (future) fridges we use are increasingly personalized, in short, adapted to who we are. The terms "personalized" and "targeted" are often used interchangeably, though a distinction between them can be made. Personalization is typically understood as the way in which (e.g., machine learning) algorithms are designed such that they can deliver something that is in line with the user's preferences, personality, and so on. Targeting can be understood as the active steps, for example, a marketer can take to send specific ads to specific groups. In short, content is personalized (usually to individuals), whereas people are targeted (usually to groups).

In terms of aggravating factors for online manipulation, the main focus is thus on personalization. A first thing to note is that there's nothing wrong about personalization as such, quite the contrary. After all, it's quite nice to enter a record shop and receive personalized advice on the latest Jeff Tweedy or Mavis Staples album you absolutely need to listen to, and it's nice (if sometimes painful) to get tailored love advice from a close friend. Likewise, it can be great to receive personalized recommendations from platforms like Spotify or Netflix, just as it can, in principle, be convenient to be recommended products you might need or like.

However, personalization inside and outside of online contexts also offers opportunities not just for welcome advice but also unwelcome influence. The reason for thinking personalization is a serious aggravating factor when it comes to manipulation is recognizable also outside of discussions about digital influence. The better someone knows us, the greater impact their advices, statements, and warnings have on us because they can tailor their advice to who we are. The existence of the well-researched

phenomenon of gaslighting – a manipulative strategy "aimed at getting another not to take herself seriously as an interlocutor" (Abramson 2014) – illustrates this clearly. Gaslighting can be as manipulative as it is precisely because the gaslighter knows the gaslightee all too well, her vulnerabilities in particular.

Having a lot of knowledge about someone isn't the same as "personalization". However, when such knowledge is put towards certain ends and becomes part of the particular things one says or does to someone, it can become – and in most social contexts, inevitably ends up being – personalized. Answering a person's question about how to get to x by giving them the answer straight is not personalization; telling your friend to get to x via y because you know there's a large flea market going on that they would enjoy (or hate), is. As is apparent, we haven't thereby yet said anything about such personalized advice being problematic or not.

As for online personalization, Susser et al. likewise mention targeting (which they seem to equivocate with personalization) as an exacerbating condition of manipulative technologies, writing that "the more targeted manipulation is the more we ought to worry about it". Or as Alexander Nix said in 2016, when he was still Cambridge Analytica's CEO, by building a psychographic model of "every single adult in the US" and thus by knowing "the personality of the people you're targeting, you can nuance your messaging to resonate more effectively with key audience groups", for instance on "specific issues such as the Second Amendment" (Concordia 2016).

Needless to say, there can also be personalized instances of online manipulation that aren't worrisome and in fact may be welcomed. Various forms of digital healthcare and mental self-care tools can be considered here. There are apps, for instance, that have virtual chat bots that adapt to the often-personal input given to them. There are many things to worry about when it comes to personalized mental health apps, such as privacy, data sale, hacking, undiagnosed conditions, less visits to GPs, and so on. *In principle*, though, online personalization might be desirable and thus not worrisome at all, even if in practice it (almost) always turns out to be.

The phrase of content, ads, or technologies being "adapted to who we are" should of course be taken with a considerable grain of salt. After all, what matters from a commercial or effectiveness perspective is first and foremost the digital profile that is constructed based on online traces a person leaves behind, not who the person really is. That being said, finding ever closer connections to people's "offline selves" – especially given that the online and offline worlds cannot be properly distinguished anymore – is of course also a way of being able to bring personalization to a higher level and influence people more effectively.

Though personalization is a serious aggravating factor when it comes to what makes technologies manipulative, we should also avoid thinking of personalization as something that is necessary to what makes certain online practices or techniques manipulative. It's also important to bear in

mind the impact of impersonal or "sweeping" forms of online manipulation. Again, it's helpful to consider the offline context here. Take propaganda for instance, which is known to have a potentially enormous impact on people's beliefs, values, and actions, but it is not a personalized type of influence, historically it has often been quite the contrary (see Stanley 2015). By steering on feelings of anger or fear, propaganda is typically a broad-scale, sweeping type of influence that intends to resonate with something that large groups of people might fall for. Similarly, online disinformation might manipulate large crowds of people without necessarily doing so in a personalized fashion.

Finally, we need to be aware that it's often also precisely the data mining corporations and political consultant firms who stress the significant impact of personalized influence. In Nix's lecture from which the previous quote was taken, he was outright bragging about the impact of psychographic profiling, mentioning that today "we need not guess" anymore about what solution may or may not work because we now know "exactly which messages are going to appeal to which audiences". This makes good *corporate* sense, but contemporary science tells a much more nuanced story. Scholars keep pointing out that measuring the efficacy of profiling techniques is difficult and that the impact is sometimes said to be questionable (cf. Zarouali et al. 2020). This is not to say personalized online influence is entirely ineffective.

In short, we need to tell a nuanced story: personalization can be a genuine aggravating factor, and thus a serious cause for concern, even if it isn't always necessary to manipulate people online and even if it isn't the "magical marionette technique" that some make it out to be.

5.2 Opacity

Not knowing about someone's manipulative strategies – its being *opaque* or *non-transparent* to someone – generally makes one a lot more prone to being manipulated. Just as with magic: if you see another's trick, the trick won't fool you or not quite in the same way. The experienced online or offline manipulator will therefore generally try to make it the case that you don't see the trick, that you don't realize attempts are being made to steer you in a particular direction.

As mentioned earlier, there is a lot of philosophical discussion about whether or not opacity is a necessary condition for manipulation, and naturally this dispute extends into the domain of online manipulation. Some think it is necessary (Susser et al.) while others don't. It may be worthwhile to adjudicate whether or not it is necessary, but it may equally be more fruitful to agree on the existing common ground: not knowing that one is being manipulated is an aggravating factor to actually being (successfully) manipulated, regardless of whether there might also be ways of being manipulated in broad, digital daylight.

Also, a question that is perhaps worth more attention than it is currently getting is the question of what transparency and opacity in the digital domain mean exactly, given that it is a highly ambiguous concept. Depending on what we take transparency to mean, there's the further question of whether (online) transparency is even a worthwhile ideal to strive towards. Though important work has already been done with respect to both the conceptual and normative questions about transparency, many questions still remain to be answered, indeed formulated (Ananny & Crawford; Sandis & Sellen; Pasquale).

A recurring topic, also in this issue, is what type of communicability or explicitness by a corporation or government institution is sufficient for a type of influence to count as transparent or no longer opaque. Does, for instance, a hard-to-find page on an organization's website suffice as being "transparent" about potentially manipulative techniques such as micro-targeting? And isn't it transparent to us, post-Cambridge Analytica, that social media platforms attempt to manipulate us? These questions cannot be answered in a black-and-white fashion; instead, they require teasing apart the different meanings of transparency and opacity in different contexts.

Though the following is highly incomplete, a rudimentary list of different types of transparency may include the following:

Organizational Transparency: the type of explicit transparency that an organization gives about their digital strategies and means of influence. In this issue, Jared W. Palmer for instance gives the example of the gamifying language platform Duolingo, who made no secret about the fact that it generated profits for its owners by offering the translation services, which were done for free by its language-learning users, to businesses. Duolingo's founder mentioned this explicitly on Duolingo's own forums.

Active Outreach Transparency: this is the type of transparency an organization might give to its subscribers, share- and stakeholders and the broader public about their digital strategies and policies, which takes the form not just of a one-on possibly hard-to-find public message but as part of a continuing project. The messaging app Signal is a possible case in point, which regularly communicates about the technologies they (don't) use and their privacy policies and ethics on their own blog and Twitter, also clarifying how Signal differs from Facebook/WhatsApp and so on.[24]

Factive Transparency: in this type of transparency, an individual knows as a matter of fact (or as a matter of high likelihood) that a service or tool they are using, such as their smart fitness watch or voice assistant, is trying to steer them in certain directions and perhaps selling their data for commercial purposes. A test for factive transparency is simply the positive and explicit answer individuals would give when asked whether they think they are being manipulated by x on platform y through method z.

Engaged Opacity: in this case, an individual has the relevant knowledge just as in Factive Transparency except (1) their knowledge is not available for conscious awareness and (2) they are unaware in this way because they are (kept) engaged in their online behaviour or "in digital flow."

Needless to say, these types of transparency/opacity are hardly exhaustive and for each of these, many sub-types need to be distinguished. But a rudimentary list like this would already be helpful when claims are made about organizations (not) being transparent or something (not) being transparent to individuals. The distinction between factive and engaged transparency, for instance, allows us to recognize that a person might know (as a matter of factive transparency) that Facebook or their smartwatch is trying to steer them in certain ways whilst failing to know (as a matter of engaged opacity) that this is going on (because they're doomscrolling or trying to break personal running records). Making these distinctions also helps us in getting clear on what type of transparency is valuable and what organizations might need to do to "be transparent", as well as bringing out the fact that many corporations do their utmost to prevent people from attaining engaged transparency.

5.3 Flow

Engaged opacity brings out something that ought to be mentioned as a serious aggravating factor in its own right: online flow. Technology is usually, and understandably, designed for comfortable user experience – nothing is as frustrating as websites or gadgets not doing (immediately) what they should be doing. At the same time, being in online flow can prevent one from being aware of relevant knowledge, can hamper one's opportunities to reflect, can bypass one's rationality, and thus prevents one from gearing one's behaviour in directions that better fit one's larger or deeper desires or ideals. This aspect has been well researched for instance by (post)-phenomenologists of technology, who stress that the seamless phenomenological experience of the online world makes that people "forget" that they're not just running in the world but running with a smartwatch, that is, running with a tiny for-profit organization clutched to one's wrist (cf. Keymolen 2018). It is also a topic for philosophers working on how the digital world affects autonomy, authenticity, and weakness of will (e.g., Williams 2018) and numerous authors in this volume).

The topic of online flow – which, given the collapse between the online and offline worlds, usually just amounts to flow in the world – also merits attention because of how it paves the way for thinking about how disrupting flow might counteract existing manipulative forces. Some scholars have for instance begun to examine the potential of introducing "friction" in tech design (Terpstra et al. 2019). If a user's flow is disrupted, this might make it easier for people to stop and think about whether they really want to watch

another video, scroll for another half hour, or insert data about one's menstruation cycle and symptoms in one's health watch.

5.4 Lack of user control

Another aggravating factor is the lack of control of the technologies that attempt to manipulate us. When it comes to being trapped in a filter bubble on YouTube or social media, there is typically little one can do to get out of one's bubble and enter another one. When it comes to recommender systems, again, there is little influence individuals have in changing the values, the settings, the input, and so on, of the technologies they use. In theory, though not in practice, it would be possible for users to select, say, more random news items or getting news from an "anti-bubble", for example, to receive news that is on the opposite end of what your political, social, or moral views are (or in any case what its algorithms believe your views are). Likewise, it is possible in theory, but not in practice, to actively tweak and improve what Spotify or Netflix think you like to listen to or watch, and the same goes for what smart homes recommend to their users. And, finally and most dramatically, it is possible in theory for users to refuse being microtargeted and tracked across the web and to have some control about the extent to which they want to give up on privacy or data traces in return for (free) services or alternatively to have the option to pay for them – but again, not possible in practice. In practice, internet users and owners of smartwatches and smart homes and what not are usually faced with a "take it or leave it" situation. If you want the robot vacuum cleaner, it comes with the corporation knowing not just the size of your rooms but also where your dinner table is and when you're (not) home. One can refuse of course, but in most cases, the service or product fails to work properly or fails to work at all. Lack of control and quasi-coercive circumstances or offers have a distinct way of making people susceptible to manipulation.

One problem about lack of user control is that of accuracy: a lack of user control also obstructs better accuracy of digital profiles. If users had more control about the technologies they engage with, the technologies would be better adapted to "who they really are" and what services or goods they are after (personalization and lack of control are thus importantly connected). But ironically, at the same time, lower accuracy due in part to a lack of user control also makes people *less* susceptible to manipulation. This is because manipulation tends to be more effective the better certain strategies are tailored to individuals' personalities and vulnerabilities. By not being able to change or adapt the digital profile or "digital persona" (Clarke 1994) that is made about us, we might also get out of some of the tech giant's digital clutches.

On the other hand, a lack of user control more often makes one more susceptible to being manipulated, especially if the need for using the technology is high (or quasi-coercive). This can be so because one is repeatedly

being exposed to certain influences even if they do not fit one's digital profile, such as being confronted with political messages that do not necessarily fit one's political views, (perhaps because one has no formulated views as yet) or being constantly confronted with products one has no desire to buy (until one sees them often enough). Without being able to "influence the influence", individuals can slide into certain ways of thinking or behaving. The worst kinds of lack of user control are "dark patterns" (e.g., Gray et al.), such as when users are deliberately refrained from changing, meddling, or refusing certain options or settings (e.g., privacy- or profiling-friendly ones).

Also, it is conceivable – again in theory – for certain services such as social media and the way news is shown to users, to *require* of users to express their preferences, to ask them whether they prefer to be shown news in line with the profile they (the for-profit organization) has constructed of them or whether they prefer an anti-bubble, or alternating filter and anti-bubbles, and so on. Such algorithmic self-governance may help make individuals more robust against manipulation. Commercial corporations are, however, unlikely, depending on their moral compass (see the following) to implement degrees of algorithmic self-governance in their services and products, hence this discussion is mostly a purely idealistic one.

5.5 An organization's moral compass

The list of possible aggravating factors is only a small and non-exhaustive list of factors that can contribute to certain technologies being manipulative. We have here described a few that we believe are particularly acute, but there are many other possible factors that are likely to contribute, such as the *free use* (financially speaking) of technological services which can create the (implicit) thought that being surveilled or manipulated in return is acceptable. Another factor is the *human-likeness* of technologies or their possible anthropomorphic nature which is especially relevant in the context of robots' potential of being manipulative. Yet another is the possible *rogueness* of technologies, that is, when technologies such as self-driving cars or war-drones start doing things on their own account, deviating from human design and plans.

Also, apart from being only a start, the list of aforementioned possible aggravating factors is just that: *possible* aggravating factors. Digital technologies, when they have one or many of the said factors, aren't necessarily manipulative. In fact, most of the factors that can make certain technologies more likely to be manipulative are also the factors that make it that certain technologies can be put to virtuous ends. Care robots that have some human-like aspects (e.g., eyes) and which operate with great flow, and which are designed to be opaque to some degree (given that people in need of care, for example those suffering from dementia or autism benefit from a degree of opacity), are likely to be more effective, for instance. The aggravating

factors, then, are not necessarily sure-fire signs that a certain technology is manipulative, or manipulative in a morally problematic way.

So when should we (not) be worried about opacity, flow, or lack of control? One important guide is the overall *moral compass* of (private or public) organizations (see, e.g., Van de Poel and Royakkers 2011; Vallor 2006; Leonelli 2016). Which values does a corporation or government institution implicitly and explicitly ascribe to? What is their business model and how do the organization's moral values relate to non-moral values such as profit maximization? Which values does it have at heart and which values does it actually carry out? Which risks and problems does it anticipate? How quickly and effectively does it react when such values (autonomy, privacy, human dignity, freedom of speech) are violated? How easy or difficult is it to get non-automated or *human* responses to requests or concerns? Depending on the answer to these questions, the aggravating factors can be worrisome to more or lesser degrees. We should be less worried about high flow and opacity when it comes to a non-profit start-up that builds privacy-friendly apps to improve women's knowledge of their menstruation cycle and moods compared to high flow and opacity when it comes to a corporation like Cambridge Analytica. Which is not to say we have no reason to be concerned even in the first case, as moral compasses of new and rapidly growing tech companies tend to change too.

Needless to say, what an organization's moral compass is, is a notoriously hard question to get an answer to. However, there are some handles to get clues including written statements on the organization's own website, the formulation and design of their Terms and Conditions, whether they have ethicists on board and/or how their ethics committee is chosen and which authority they are assigned, the way they respond to concerns or incidents, whether they engage in ethics washing, and so on.

It is the combination of an analysis of the possible aggravating factors of certain technologies in combination with a sense of an organization's moral compass that designs those technologies or puts them to use that we can get a picture of the level of concern about how likely, and just how impactful, manipulation will be.

6 Conclusion

In this chapter, we have charted the field of the contemporary debate concerning online manipulation. As for the method of studying (online) manipulation, we have discussed the classical conceptual analysis approach and mentioned its problems as well as novel alternative methodological approaches such as the "focal case concept" approach. We also mentioned that, when studying manipulation, one needs to decide and/or be explicit about (1) whether or not one thinks manipulation is a so-called thick or moralistic concept and (2) whether manipulation necessarily involves intentionality and if so, in what sense.

We then moved on to discuss the concept of manipulation and which features might help us distinguish it from coercion and persuasion. To this end, we distinguished outcome-based views (in terms of (3.2.1) self-interest and harm and (3.2.2) autonomy), process-based views (in terms of (3.3.1) covertness or (3.3.2) bypassing rationality), and norm-based views (including the negligence-based view).

In the second half of this chapter we mentioned numerous possible aggravating factors, that is, factors that make manipulation worse or that make it harder for people to get out of a manipulator's clutches. We focused in particular on (5.1) personalization, (5.2) opacity, (5.3) flow, and (5.4) lack of control. Finally, we mentioned that taking into account an organization's *moral compass* – in spite of often being a near-impossible endeavour – is key to knowing whether the said factors are indeed cause for concern.

It should be stressed at this point that "the field" we have chosen to chart has been only a small piece of a larger landscape. As we discussed in the Introduction to this volume, several important and intriguing aspects of (online) manipulation such as its legal, political, and psychological aspects cry out for further study, and they promise much intriguing insight.

Notes

1. We are grateful to Anne Barnhill, Thomas Nys, and Robert Noggle for very helpful written comments on an earlier version of this chapter. The audience at our online workshop series also provided helpful comments and suggestions on an early presentation of the material collected here. Both authors contributed equally to this chapter. Michael Klenk drafted initial versions of Sections 2, 3, and 4, and Fleur Jongepier drafted initial versions of Sections 5 and 6. Michael Klenk's work on this chapter was supported first by a Niels Stensen Fellowship and later by the European Research Council under the Horizon 2020 programme under grant agreement 788321. Fleur Jongepier's work on this chapter was supported by an NWO Veni grant (VI.Veni.191F.056).
2. We simplify the debate about the nature of analysis here for ease of exposition. See Beaney (2021) for further discussion.
3. Though see the discussion by Houk (2018) on alternative approaches.
4. See Feurer and Fischer (2021) and Klenk, Xun Liu, and Hancock (2021) for examples of the nascent experimental work on manipulation.
5. Especially considering the question of whether manipulation has – as a conceptual matter – a normative or evaluative component. See Hopster and Klenk (2020) for further discussion on the limits and benefits of using empirical methods in ethics.
6. The view that several conditions such as deception, autonomy loss, and harm are associated with manipulation is also supported by initial experimental research on non-philosopher's views about manipulation (cf. Klenk, Xun Liu, and Hancock 2021).
7. Of course, if there isn't even a paradigm, as suggested by Baron (2003, 37), then even this approach is put into doubt.
8. Thanks to Anne Barnhill for prompting us to clarify this point.
9. Compare the discussion of the thickness of manipulation in Wood (2014), who like us understands it as a question about the meaning of the concept, versus

the sense in which manipulation is a pro tanto wrong, as discussed by Baron (2014).

10. This almost sounds like a contradiction in terms and is impossible to pull off, but it becomes more feasible if one were to distinguish moral and non-moral forms of laudability. It is sometimes also said of certain populist politicians (clearly not all of them) that they are cunning in such a way that demands our respect, even if their cunningness is used for immoral purposes and so do not demand our *moral* respect.
11. In addition, Coons and Weber (2014) note that we may wonder about whether anything is truly right or wrong independently of some idiosyncratic perspective – as various sceptical challenges in philosophy and beyond demonstrate – but we do not wonder about the reality of manipulation. Proponents of a thick view on manipulation could maintain that a subjective evaluation is part of the concept, but it would be less plausible to suggest that manipulation is stance-independently moralized as a conceptual matter. Based on this sceptical view, there must be some descriptive account of the concept of manipulation independently of moralized considerations.
12. Also note that the process, outcome, and norm-based accounts of manipulation that we discuss in Section 3 may be presented in what we might call deontic or telic fashion. Deontic versions of these accounts portray the demarcating features as the object of an intention. For instance, a deontic covertness thesis would have the manipulator intend to covertly influence her victim. A telic or consequential version would do without intentions and merely require that there is an influence that leads to the manipulatee remaining oblivious about some important feature of the interaction. The distinction between what we call deontic and telic versions of different accounts of manipulation is not always made explicit, nor are decisions for or against a particular view defended. But it seems to be a reasonable and noteworthy distinction to draw. This is especially so given the focus of this volume on interactions mediated by and perhaps with machines that may lack intentionality.
13. Several scholars, like Baron (2003), suggest that manipulation merely limits options, rather than removing them, and that this may be a useful demarcating factor. See also Handelman (2009), who defends the view that manipulation is about presenting some specific choice as best to the agent.
14. See also the debate about incompatibilism, free will, and moral responsibility. The important manipulation cases are supposed to involve a victim performing the manipulator's course of action on its own volition, cf. Sripada (2012). See also Cave (2007) for a discussion of the charge that motive manipulation is morally bad, which seems to be similar to Fischer (2017); Fischer and Illies (2018).
15. The incompatibilism debate is interesting in this context. Incompatibilists argue that the "not fully free" intuition is sensitive to the agent in a manipulation case not being the ultimate source of his or her action. Compatibilists, in contrast, suggest that this intuition is sensitive to the fact that manipulation damages or impairs the agent's cognitive, evaluative, or affective capacities.
16. Garnett (2018) being an illuminating exception. Note also that in fields outside philosophy (e.g., communication studies) persuasion is used to describe tactics commonly associated with manipulation. Note also that it is not entirely clear that an analysis of patient behaviour – such as coerced or manipulated action – allows for inferences about agent behaviour – such as coercive or manipulative action. There may be benefits to dissociating analyses of manipulated from analyses of manipulating action and to offer accounts that are partly independent, for example, Klenk, in this volume.
17. Thanks to Anne Barnhill for prompting us to clarify this point.

18. Thanks to Thomas Nys for prompting us to clarify this point.
19. See also Cohen (2018).
20. A variant of that view might be the one suggested by Blumenfeld (1988), who suggests that manipulation bypasses character, which he understands as an amalgam of reasons, motives, and desires integrated in the manipulatee's character.
21. Some theorists have suggested that manipulation works not only by bypassing reasons but – more specifically – by exploiting vulnerabilities in the subject. Again, this may be correct as a causal statement about manipulation because manipulation may often happen to proceed in these ways. But the claim interpreted as a conceptual claim is more difficult to maintain. The primary problem with this is that vulnerabilities are likely relative to context. For example, the gustatory "bias" to prefer sugary food was great in the environment of our evolutionary development, but in today's world with an oversupply of calorie-rich food it is to our detriment. If some of our dispositions are vulnerabilities given a context, then the account of manipulation as playing on our vulnerability would suggest that we need to appeal to dispositions that are powerful or strong given a context. It is not clear what that would mean, and it is possible it drives on intuitions related to the bypassing reason view or the autonomy view.
22. Thanks to Robert Noggle for helpful feedback on this point.
23. Noggle's account thus makes explicit the specific intentionality requirement that we discussed earlier. Other proponents of norm-based views like Gorin et al. (2017) or Barnhill (2014, 2016), however, do not make the intentionality requirement explicit. In his contribution to our volume, Gorin does make it explicit (cf. Gorin, in this volume).
24. https://signal.org/blog/

7 References

Abramson, Kate. 2014. "Turning up the Lights on Gaslighting." *Philosophical Perspectives* 28 (1): 1–30. doi:10.1111/phpe.12046.

Ackerman, Felicia. 1995. "The Concept of Manipulativeness." *Philosophical Perspectives* 9: 335–40. doi:10.2307/2214225.

Alm, David. 2015. "Responsibility, Manipulation, and Resentment." *Social Theory and Practice* 41 (2): 253–74.

Alston, W. P. 1967. "Vagueness." In *The Encyclopedia of Philosophy*, edited by P. Edwards, 218–21. New York, NY: Collier-Macmillan.

Barnhill, Anne. 2014. "What is Manipulation?" In Coons and Weber 2014, 51–72.

Barnhill, Anne. 2016. "I'd Like to Teach the World to Think: Commercial Advertising and Manipulation." *Journal of Marketing Behavior* 1 (3–4): 307–28. doi:10.1561/107.00000020.

Baron, Marcia. 2003. "Manipulativeness." *Proceedings and Addresses of the American Philosophical Association* 77 (2): 37. doi:10.2307/3219740.

Baron, Marcia. 2014. "The Mens Rea and Moral Status of Manipulation." In Coons and Weber 2014, 98–109.

Beaney, Michael. 2021. "Analysis." In *Stanford Encyclopedia of Philosophy: Summer 2021*, edited by Edward N. Zalta. https://plato.stanford.edu/archives/sum2021/entries/analysis/.

Blumenfeld, David. 1988. "Freedom and Mind Control." *American Philosophical Quarterly* 25 (3): 215–27.

Blumenthal-Barby, J. S. 2012. "Between Reason and Coercion: Ethically Permissible Influence in Health Care and Health Policy Contexts." *Kennedy Institute of Ethics Journal* 22 (4): 345–66.

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.

Cave, Eric M. 2007. "What's Wrong with Motive Manipulation?" *Ethical Theory and Moral Practice* 10 (2): 129–44.

Chappell, Sophie G. 2019. "Introducing Epiphanies." *Zeitschrift Für Ethik Und Moralphilosophie* 2 (1): 95–121. doi:10.1007/s42048-019-00029-4.

Clarke, Roger. 1994. "The Digital Persona and its Application to Data Surveillance." *The Information Society* 10 (2): 77–92. doi:10.1080/01972243.1994.9960160.

Climenhaga, Nevin. 2018. "Intuitions are Used as Evidence in Philosophy." *Mind* 127 (505): 69–104. doi:10.1093/mind/fzw032.

Cohen, Shlomo. 2018. "Manipulation and Deception." *Australasian Journal of Philosophy* 96 (3): 483–97. doi:10.1080/00048402.2017.1386692.

Concordia. 2016. "Cambridge Analytica – The Power of Big Data and Psychographics [video file]." Accessed September 10, 2021. www.youtube.com/watch?app=desktop&v=n8Dd5aVXLCc&feature=youtu.be.

Coons, Christian, and Michael Weber, eds. 2014. *Manipulation: Theory and Practice*. Oxford: Oxford University Press.

Faden, Ruth R., and Tom L. Beauchamp. 1986. *A History and Theory of Informed Consent*. New York, NY: Oxford University Press.

Feurer, Sven, and Alexander Fischer. 2021. *Exploring the Ethical Limits of Manipulation in Marketing: A Discussion Based on Consumer Perceptions*. under review.

Fischer, Alexander. 2017. *Manipulation: Zur Theorie und Ethik einer Form der Beeinflussung*. Berlin: Suhrkamp.

Fischer, Alexander, and Christian Illies. 2018. "Modulated Feelings: The Pleasurable-Ends-Model of Manipulation." *Philosophical Inquiries* 1 (2): 25–44. Accessed August 06, 2020.

Garnett, Michael. 2018. "Coercion: The Wrong and the Bad." *Ethics* 128 (3): 545–73. doi:10.1086/695989.

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

Gorin, Moti. 2014b. "Towards a Theory of Interpersonal Manipulation." In Coons and Weber 2014, 73–97.

Gorin, Moti. 2018. "Paternalistic Manipulation." In *The Routledge Handbook of the Philosophy of Paternalism*, edited by Jason Hanna and Kalle Grill, 236–47. New York, NY: Routledge.

Gorin, Moti, Steven Joffe, Neal Dickert, and Scott Halpern. 2017. "Justifying Clinical Nudges." *The Hastings Center Report* 47 (2): 32–38. doi:10.1002/hast.688.

Greenspan, Patricia. 2003. "The Problem with Manipulation." *American Philosophical Quarterly* 40 (2): 155–64.

Handelman, Sapir. 2009. *Thought Manipulation: The Use and Abuse of Psychological Trickery*. Santa Barbara, CA: Praeger Publishers.

Harari, Yuval N. 2018. "The Myth of Freedom." *The Guardian*, September 14. Accessed August 27, 2021. www.theguardian.com/books/2018/sep/14/yuval-noah-harari-the-new-threat-to-liberal-democracy.

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. "Most People Are Not WEIRD." *Nature* 466 (7302): 29. doi:10.1038/466029a.

Hopster, Jeroen, and Michael Klenk. 2020. "Why Metaethics Needs Empirical Moral Psychology." *Critica* 52 (155). doi:10.22201/iifs.18704905e.2020.1193.

Houk, Timothy. 2018. "The Nature and Morality of Manipulation." PhD thesis, University of California, Davies.

Jongepier, Fleur. 2017. "The Circumstances of Self-Knowledge." PhD thesis, Radboud University Nijmegen.

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

Kane, Robert. 1996. *The Significance of Free Will*. New York, NY: Oxford University Press.

Keymolen, Esther. 2018. "Trust in the Networked Era." *Techné: Research in Philosophy and Technology* 22 (1): 51–75. doi:10.5840/techne201792271.

Klenk, Michael. 2022. "Manipulation, Injustice, and Technology." *In The Philosophy of Online Manipulation*, edited by Jongepier, F. and Klenk, M., pp. 108–132. New York: Routledge.

Klenk, Michael. 2020. "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. 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, 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.

Kligman, M., and C. M. Culver. 1992. "An Analysis of Interpersonal Manipulation." *The Journal of Medicine and Philosophy* 17 (2): 173–97. doi:10.1093/jmp/17.2.173.

Knobe, Joshua, and Shaun Nichols. 2008. *Experimental Philosophy*. New York, NY: Oxford University Press.

Krstić, Vladimir, and Chantelle Saville. 2019. "Deception (Under Uncertainty) as a Kind of Manipulation." *Australasian Journal of Philosophy* 97 (4): 830–35. doi:10.1080/00048402.2019.1604777.

Leonelli, Sabina. 2016. "Locating Ethics in Data Science: Responsibility and Accountability in Global and Distributed Knowledge Production Systems." *Philosophical Transactions of the Royal Society A* 374 (20160122).

Levy, Neil. 2019. "Nudge, Nudge, Wink, Wink: Nudging Is Giving Reasons." *Ergo* 6. doi:10.3998/ergo.12405314.0006.010.

Manne, Kate. 2014. "Non-Machiavellian Manipulation and the Opacity of Motive." In Coons and Weber 2014, 221–46.

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

Noggle, Robert. 2018a. "Manipulation, Salience, and Nudges." *Bioethics* 32 (3): 164–70.

Noggle, Robert. 2018b. "The Ethics of Manipulation." In *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/ethics-manipulation/

Pereboom, Derk. 2001. *Living Without Free Will*. Cambridge: Cambridge University Press.

Pölzler, Thomas. 2020. *Moral Reality and the Empirical Sciences*. New York, NY: Routledge.

Queloz, Matthieu. 2021. *The Practical Origins of Ideas: Genealogy as Conceptual Reverse-engineering*. Oxford: Oxford University Press.

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

Scanlon, Thomas M. 1998. *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

Schelling, Thomas C. 1997. *Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Schmidt, A. T., and B. Engelen. 2020. "The Ethics of Nudging: An Overview." *Philosophy Compass* 15 (4).

Sripada, Chandra S. 2012. "What Makes a Manipulated Agent Unfree?" *Philosophy and Phenomenological Research* 85 (3): 563–93.

Stanley, Jason. 2015. *How Propaganda Works*. Princeton, NJ: Princeton University Press.

Sunstein, Cass R. 2016a. "Fifty Shades of Manipulation." *Journal of Marketing Behavior* 1 (3–4): 214–44. doi:10.1561/107.00000014.

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

Susser, Daniel, Beate Roessler, and Helen Nissenbaum. 2019a. "Online Manipulation: Hidden Influences in A Digital World." *Georgetown Law Technology Review* 4 (1): 1–45.

Susser, Daniel, Beate Roessler, and Helen Nissenbaum. 2019b. "Technology, Autonomy, and Manipulation." *Internet Policy Review* 8 (2): 1–22. doi:10.14763/2019.2.1410.

Terpstra, Arnout, Alexander P. Schouten, Alwin de Rooij, and Ronald E. Leenes. 2019. "Improving Privacy Choice Through Design: How Designing for Reflection Could Support Privacy Self-Management." *FirstMonday* 24 (7): 1–13. doi:10.5210/fm.v24i7.9358.

Vallor, Shannon. 2006. *Technology and the Virtues. A Philosophical Guide to a Future Worth Wanting*. Oxford: Oxford University Press.

Van de Poel, Ibo, and Lambèr Royakkers. 2011. *Ethics, Technology, and Engineering: An Introduction*. Malden, MA: Wiley-Blackwell.

Williams, James. 2018. *Stand out of Our Light: Freedom and Resistance in the Attention Economy*. Cambridge: Cambridge University Press.

Wood, Allen W. 2014. "Coercion, Manipulation, Exploitation." In Coons and Weber 2014, 17–50.

Zarouali, Brahim, Tom Dobber, Guy de Pauw, and Claes de Vreese. 2020. "Using a Personality-Profiling Algorithm to Investigate Political Microtargeting: Assessing the Persuasion Effects of Personality-Tailored Ads on Social Media." *Communication Research*, 1–26. doi:10.1177/0093650220961965.