5 Manipulative machines

Jessica Pepp, Rachel Sterken, Matthew McKeever, and Eliot Michaelson[1]

1 Introduction

One theoretical approach to mitigating existential risk to human beings from artificial "superintelligence" is Oracle AI. The idea of Oracle AI is to design an AI that cannot act except to answer questions. The Oracle can thus be used by humans to achieve their goals but cannot affect the outside environment to pursue its own, potentially dangerous, goals. A critical conceptual problem with this idea is that an Oracle AI would still have a channel by which to influence the world, namely its answers to human questions. In particular, it could *manipulate* the humans with which it interacts into "setting it free" such that it could influence the world in more direct ways (Armstrong, Sandberg, and Bostrom 2012; Armstrong and O'Rourke 2018; Chalmers 2010).[2]

It is a matter of controversy how great the threat from superintelligence is and whether Oracle AI is a good approach to risk mitigation. We will not be entering into those fascinating discussions here. Rather, we will take the claim that the potential manipulation of humans by AIs is part of this threat as a useful jumping-off point for a philosophical study of the concept of manipulation in the context of human–machine interactions.

It might seem obvious that a superintelligence like Oracle AI could manipulate us, and theorists (like those cited above) who study these potential beings do not hesitate to describe worries about their behavior in these terms. A superintelligence, after all, even if it is a machine, is by definition (way) more intelligent than a human being. So, if human beings can manipulate each other, there might seem no reason to think a superintelligence could not manipulate human beings. However, human–machine interactions present a challenge for the analysis of the concept of manipulation. For whether machines can manipulate us depends on whether one entity's manipulating another is simply a matter of the first entity having a certain kind of effect or influence on the second or whether it also requires a certain kind of mental state on the part of the manipulator. A superintelligence surely could influence us in ways that might seem manipulative. But if manipulation requires the manipulator to have certain kinds of thoughts,

DOI: 10.4324/9781003205425-6

desires, beliefs, or intentions, then the notion of a superintelligence (as it is usually defined) leaves open whether a superintelligence could manipulate. This is because it leaves open whether a superintelligence would have such "intentional" states (Bostrom 2014, 22, n 2).[3]

In light of this challenge, we will explore three ways to make sense of the reasonable-sounding claim that manipulation by machines is possible, and in the extreme case could even pose an important kind of threat to humankind, that might go as far as to be an existential threat (though our argument doesn't turn on such dire possibilities). The first is to argue that manipulation by machines is included under the (or, at any rate, *a*) current concept of manipulation.

The second is to allow that machine manipulation is not included under current concepts of manipulation, but to argue that there are nonetheless good reasons to group it together with human manipulation, or to treat it in parallel with human manipulation. On this view, machines might, *speaking loosely*, be referred to as manipulators, without any commitment to an analysis of manipulation according to which they are, strictly speaking, capable of manipulation. We might describe machines in this way just as we describe a crime scene as "suggesting" or "showing" or "casting doubt" on whether a particular crime occurred, even though we don't attribute sentience to the scene – presumably because certain features of the crime scene might have the same probative value as intentional testimony by witnesses. So it's helpful to classify the crime scene *as if* it were a witness. This second approach could be taken by a theorist who thinks that the correct conceptual analysis of manipulation simply excludes machines from being manipulators or by a theorist who does not care about the analysis of our concept of manipulation. The latter type of theorist may be more interested in understanding why it makes sense to speak of machine manipulation – as those who study existential threat from superintelligent AI readily do – than in arriving at an account of the concept.

The third approach we will consider is what Haslanger (2000, 2012) calls an "ameliorative project" concerning the concept of manipulation. The ameliorative approach starts by asking what legitimate purposes there are to having a concept of manipulation and then seeks the concept that best serves these purposes. On this approach one can make sense of machine manipulation by arguing that a concept of manipulation which includes certain activities of machines best serves the legitimate purposes of a concept of manipulation. For example, one purpose of a concept of manipulation may be to allow us to identify, call out, and mitigate certain concerning effects or influences, and if our current concept(s) of manipulation exclude significant activities which cause such effects or influences, then an ameliorated concept of manipulation which includes such activities would better serve our purposes.

Call the first approach, that of fitting machine manipulation under our current concept, *conservative conceptual analysis*. The challenge for this

approach is that many extant analyses of the concept of manipulation require either certain sorts of intentions or states of mind on the part of the manipulator or norm violations by the manipulator. But whether machines, even superintelligent ones, can have intentional states is a famously fraught question and still generally viewed as unsettled. (The literature is gigantic, but thankfully we can mostly ignore it: Turing 1950 and Searle 1980 are seminal; Cole 2020 provides an overview of the state of the art on Searle's Chinese room argument.) Similarly, it is unclear whether machines can be subject to norms of any kind.

We will set out one way of pursuing the first approach so as to deal with this challenge in Section 2. There, we lay the groundwork for an analysis of the concept of manipulation that involves neither the manipulator's intentional states nor the various norms she may be subject to. Our approach here is based on extant analyses of manipulation that focus on the manipulative influence rather than on the manipulator's state of mind; thus, we call it the *influence-centric approach*.

Next, in Section 3, we will explore the view that, whether or not it is strictly speaking, the case that machines can manipulate us, it is useful to use the word "manipulate" and its cognates to describe certain kinds of influence that machines have on us. This approach is compatible with an error-theoretic stance according to which the concept of manipulation does not extend to the phenomena that are candidates for being instances of manipulation by machines, perhaps because machines cannot have beliefs, desires, and intentions. But it is also compatible with a broader skepticism about conceptual analysis and the stance that there is no interesting answer to the question of whether that which one might be inclined to call "machine manipulation" is really manipulation.

Finally, in Section 4, we will recast our proposal from Section 2 as an ameliorative analysis of the concept of manipulation. Whether or not this proposal captures the current concept of manipulation, a concept along these lines might be most helpful in serving the legitimate purposes of a concept of manipulation. Although we will not offer a full defence of the ameliorative approach over the others we canvas, we will make a preliminary case for its promise in the concluding Section 5.

In setting out these approaches, we will have in mind not only hypothetical cases like Oracle AI but also some of today's candidates for manipulative machines. One salient class of examples are chatbots and virtual assistants, such as the Casper's Insomnobot 3000, whose job it is to soothe and chat with lonely sufferers of insomnia throughout the night and Amtrak's virtual assistant, which helps its website visitors to plan and book trips in real time. Another example is YouTube's video recommendation algorithm, which recommends videos to YouTube viewers by ranking them based on performance metrics (e.g., clicks, watch time) and personalization to viewers' interests (e.g., topic, history, context).[4] A further example is the Facebook advertising algorithm, which selects advertising to be placed in users'

news feeds by doing a complicated estimation and weighing of a given user's likelihood of engaging with (e.g., clicks, views, likes) a given ad, a given advertiser's bid for the slot in the feed, and the amount of money Facebook will be paid if the user does engage.[5] In assessing whether such present-day machine learning systems are manipulators, our focus is on whether they are manipulators in their own right, as opposed to simply being tools used by humans to manipulate.

As is no doubt evident, the three approaches that we will explore do not lead in compatible directions. However, at this early stage of research into manipulation and machines, our aim is to travel some distance down a variety of paths in order to identify both the challenges and the promise of different approaches.

2 Manipulation without intentionality: an influence-centric account

In a stereotypical case of manipulation, one person tries to get another person to act, think, or feel in some particular way, not by outright forcing or coercing the other person to act, think, or feel in this way but not by rationally persuading the other person to act, think, or feel in this way either. The manipulator may deceive the other person, pressure her, and/or play on her emotions and vulnerabilities. Granted, the line between pressure and force will sometimes be difficult to limn, but in a stereotypical case, the manipulator will consciously and deliberately aim at getting her target to do (think, feel) what she wants without resorting to threats or the like, adjusting her strategy as the situation unfolds. This archetype of a strategic manipulator can make it seem as though a strategic state of mind is essential to manipulation. As Marcia Baron puts it, manipulation has a "*mens rea*": there is a "mental component necessary for something to count as an act of manipulation" (Baron 2014, 100).

Baron suggests that the requisite *mens rea* is "intent to get the other to do *x*, along with insufficient concern about the other *qua* agent [in the way one goes about reaching the goal of getting the other to do *x*]". She does not think that a manipulator must intend to manipulate the other under that description. Kate Manne (2014) agrees and adds that the intent to get the other to do (or think, feel) something need not be conscious, since a person may manipulate someone else in spite of not having any conscious intention to do anything manipulative and even in spite of having a conscious intention of *not* manipulating that person. Still, Manne (at least tentatively) agrees with Baron that manipulators must at least unconsciously intend, or have a motive, to get the other to do, think, or feel something.[6]

This general view of the *mens rea* essential to manipulation includes both a positive and a negative aspect. The positive aspect is the claim that the manipulator must have an intention, or a motive, to get the other to do something, even if that intention or motive is unconscious and/or in conflict

with the manipulator's conscious intentions and motives. The negative aspect is the claim that they must display a lack of concern for the other's agency. It is the positive claim that implies that manipulation must be carried out by intentional agents, so it will be our focus.[7]

What Baron calls the *mens rea* is one side of manipulation. The other side – the *actus rea*, if you like – is the manipulative influence itself. Some theorists aim to characterize manipulation mostly by focusing on the nature of manipulative influence rather than on the mental states of manipulators. For instance, take the following account from Anne Barnhill (Barnhill 2014, 52):

> Manipulation is directly influencing someone's beliefs, desires, or emotions such that she falls short of ideals for belief, desire, or emotion in ways typically not in her self-interest or likely not in her self-interest in the present context.

This account of manipulation does not say anything about the manipulator's own states, but only describes how the manipulated person is influenced. Barnhill (2014, 68–69) is agnostic about whether manipulation requires intent or motive such as Baron and Manne describe, noting that some people have the intuition that manipulation must be intentional, at least at some level, while others are inclined to think one person may manipulate another by having an influence of the relevant kind, even if they do not in any way, at any level intend to have such an influence.

Like Barnhill, Allen Wood's (2014) account of manipulation focuses on the nature of the influence rather than on the state of mind of the manipulator. What is characteristic of manipulation, Wood says, is that it

> influences people's choices in ways that circumvent or subvert their rational decision-making processes, and that undermine or disrupt the ways of choosing that they themselves would critically endorse if they considered the matter in a way that is lucid and free of error.
>
> (Wood 2014, 35)

And he goes further than Barnhill in outright rejecting the requirement for any type of intention on the part of the manipulator or indeed of any *mens rea* at all. Wood claims that there are cases of "manipulation without a manipulator", but what he means is, manipulation without an individual person or a group of persons who is/are the manipulator. The cases Wood describes are cases where someone is manipulated by a system or social institution, specifically the capitalist free market system and the social institution of advertising. In these cases, according to Wood, something is indeed doing the manipulating. But what does the manipulating is not an entity with a state of mind since it is not an entity with intentional states at all.

On Barnhill's and Wood's understandings of manipulation (as opposed to Baron's and Manne's), Oracle AI could surely manipulate humans, since it could influence the beliefs, emotions, and desires of its human interlocutors in ways that are not in their self-interest or that undermine their rational decision-making processes. For instance, by engaging in very human-like conversation so as to cause a human interlocutor to become emotionally bonded to it,[8] an Oracle AI could cause the human to desire that her Oracle friend be free, leading the human to neglect all the good reasons she knows she has to keep the Oracle contained. Indeed, even the lowly Facebook advertising algorithm is a manipulator in this sense. This algorithm (really a cluster of machine learning algorithms) places advertisements in users' news feeds based on a complex calculation incorporating advertiser bid levels, estimates of users' likelihoods to click or otherwise engage with the advertisement, and many other factors (see Note 5). Drawing again on Barnhill's definition for illustrative purposes, it seems that this cluster of algorithms directly influences users' beliefs, desires, and emotions in ways that fall short of ideals. For instance, it may cause them to desire those new shoes they cannot really afford to buy, or to feel enthusiasm for a political candidate who does not best represent their interests, or to believe that they might be able to lose weight quickly with a new diet plan though experience has demonstrated that this is unlikely. In this way, the algorithm would promote non-ideal emotions, desires, and beliefs.

So when it comes to machine manipulation, one might simply claim that the notion is unproblematic because manipulation is not essentially tied to a manipulator's state of mind (*mens rea*) but to the influence the manipulator exerts (*actus rea*). And machines, even those we are surrounded by today, can and do exert the relevant kinds of influence on us. But this banishment of the states of the manipulator from the concept of manipulation may be too quick.

Consider the following case. Jane is very superstitious about cracks in pavements. Ever since learning a rhyme in childhood about breaking your mother's back, she has religiously avoided stepping on them and is always in a slightly heightened state of visual monitoring when walking on pavements. One day, while walking on a pavement, Jane mistakes an unusually straight and thin streak of mud for a crack. The streak of mud directly causes Jane to believe something false (that there is a crack in the pavement), to form a desire that is not in her best interest (a desire to avoid a crack in the relevant location, which will make her gate less efficient), and to experience emotions of fear and anxiety in a case where they are not warranted. It seems fair to say that the streak of mud influences Jane's beliefs, desires, and emotions in ways such that she falls short of ideals for beliefs, desires, and emotions. Similarly, it seems fair to say that Jane's reason is bypassed (because the reaction is due to her superstition), that she is deceived (since she mistakes the streak for a crack), and that she is pressured (since the

streak evokes emotions of fear and anxiety about stepping on a crack). But there does not seem to be any *manipulation* here.

The problem is not just that the influencer is not a person (or intentional agent) since the same kind of thing can happen when the influencer *is* a person. Consider the following case. Daniel has recently left a destructive relationship that was built upon the overuse of alcohol. While shopping, he sees a person who looks like his ex-partner. He is flooded with longing for the drunken excitement that they once shared and acquires a desire to purchase alcohol to drink later. This person directly influences Daniel's desires and emotions such that he falls short of relevant ideals. But this person does not manipulate Daniel, and there is no manipulation here – at least, there seems to be no more reason to think so than there is in the case of Jane and the mud on the pavement.

The worry, then, is that an influence-based account of manipulation, which can accommodate manipulative AI and machine learning systems, might overgenerate cases of manipulation. One way to respond to this is simply to embrace it. Yes, the mud on the pavement and the stranger in the store manipulate Jane and Daniel in these circumstances. If we accept that there can be manipulation without a (intentional agent-type) manipulator, there is nothing problematic about this. What is interesting and important about manipulation is the way in which it influences us – the *actus rea* – and these cases exemplify manipulation as well as those in which an archetypal strategic, human manipulator wields the influence.

However, it seems to us that the concept of manipulation is not this broad, and that saying that the mud manipulates Jane or that the stranger manipulates Daniel would be clearly figurative applications of the concept. Manipulation, whether by humans, animals, institutions, or machines, is distinguished from other ways in which the rationality of people's attitudes and decisions may be degraded (such as by chance occurrences as described in the last two examples). Although we will not defend a particular analysis of manipulation that reflects this, we will propose a *necessary* condition on manipulation that would be part of such a concept. This condition could be combined with an account like Barnhill's, for instance, to yield something closer to a necessary and sufficient condition for manipulation, understood in an influence-centric way:

> For any entity (person, animal, institution, machine, etc.) X, a behavior or feature of X having a certain influence on another entity Y is an instance of manipulation only if the occurrence of the behavior or feature in X is partly explained by its tendency to have that influence on Y or on other entities relevantly like Y.

The idea behind this is that acts or features whose influence counts as manipulation occur or obtain *because* they are likely to have certain kinds of influence on others. In some cases, their likelihood of having this influence combines with a manipulator's intention or desire to have that influence in the explanation of why they occur or obtain. But this need not be

the case, so long as the likelihood that the acts or features will have that influence is part of the explanation of why they occur or obtain.

Consider Kate Manne's case of Joan, who gives extravagant gifts to neglectful relatives, without any conscious intention or desire to make them feel guilty about not maintaining their relationship with her. Manne judges that Joan's behavior counts as manipulation despite the lack of conscious motivation to steer the relatives' beliefs, desires, emotions, or decisions. Nonetheless, it seems clear from Manne's description that the tendency of extravagant gift-giving to make neglectful relatives feel guilty *is* part of the explanation of why Joan does it.[9] Similar reasoning applies to Manne's example of Neal, a character from a David Foster Wallace story (Foster Wallace 2004, "Good Old Neon") who tries not to be manipulative but cannot seem to help it. Plausibly, Neal's manipulative behavior is explained by deep-seated, unhealthy psychological needs he has to be perceived in certain ways by others. Because he has these needs, and because certain behaviors tend to cause others to perceive him in the ways the needs demand, Neal exhibits these behaviors, even when he tries very hard not to. Once again, the tendency of the behaviors to have the manipulative influence partly explains the fact that Neal exhibits them.

A similar case can be made for Wood's examples of institutional manipulation by capitalism and by the institution of advertising (as opposed to individual advertisers or corporations). Wood says that both of these manipulate people by

> encouraging them to focus narrowly on their own lives, and even regarding their own lives, to focus only on the present and the immediate future. It encourages people in the idea that they owe nothing to other people except those (such as their family) with whose interests they are immediately engaged.
>
> (Wood 2014, 39–40)

Presumably, to connect this with Wood's general remarks about the nature of manipulation, the idea is that these encouragements hamper people's rational decision-making processes. Of course, it is debatable whether or not capitalism and advertising (*qua* institution) have such influences. But if they do, it does not seem so far-fetched to call the production of such influences by features of these institutions *manipulation*.[10] Further, we submit, part of the reason why it does not seem so far-fetched is that these cases satisfy the requirement we articulated earlier. Whatever features of advertising encourage limited focuses that hamper people's capacity for rational choice are there partly *because* they have this effect.

For instance, suppose the endless repetition of jingles or slogans is one such feature. This has come to be a hallmark of advertising in part because it causes people to focus on their immediate desires and purchase products for which they get a fleeting yearning (perhaps because a jingle is stuck in their head). In the case of capitalism, the story would be more complicated.

Drawing from Wood's discussion, it might be something like this: the capitalist system influences people's attitudes and choices by not making manifest to them the broader consequences of their market activities. This feature of consequence-opacity obtains in part because it tends to have the effect of encouraging people to make short-sighted economic decisions, which in turn promote the capitalist system.[11]

Of course, these are just-so stories that are inaccurate or vastly oversimplified. It is debatable whether anything in the vicinity is, in fact, the case. Still, it seems to us that if nothing in the vicinity is the case – if the explanations of why advertising and capitalism have these features have nothing to do with their tendency or likelihood of producing the influence that is supposed to be manipulative – then it is much less plausible that they are cases of manipulation.

The requirement we proposed also clearly rules out the cases of the mud on the pavement and the stranger in the shop from being cases of manipulation. The explanation of why the mud looks like a crack has nothing to do with the fact that looking this way is likely to influence Jane's attitudes and choices. Likewise, the explanation of why the stranger in the store looks like Daniel's ex-partner has nothing to do with the fact that looking this way is likely to influence Daniel's attitudes and choices.

Contrast these cases with the hypothetical case of Oracle AI, and the actual cases of the Facebook advertising algorithm and the YouTube video recommendation algorithm. If Oracle AI manipulates a human interlocutor into setting it free by using language that causes feelings of emotional bonding and love in the human, the Oracle's use of that language is explained by its likelihood of causing those feelings in the human. (Presumably, the Oracle will have trained on human behavior datasets that give it a very good estimate of such likelihoods.) Similarly, when the Facebook algorithm displays a certain advertisement in the news feed of a certain user, the explanation of why it does that has to do with the likelihood of generating clicks, likes, or views, which is itself explained by the likelihood of influencing the user's attitudes and decisions in the relevant ways.[12]

We have now seen one broad approach to developing an account of manipulation that allows for machines to be manipulators whether or not they are intentional systems: the influence-based approach focuses on the kind of influence a manipulator has rather than on their state of mind. We argued that extant influence-based accounts of manipulation can be combined with the necessary condition proposed previously to give a viable, non-intentional analysis of manipulation. Next, we will go on to another strategy altogether.

3 Never mind if it's manipulation: "loose talk" or error-theoretic approaches

The second sort of strategy we want to consider is one according to which it can be useful to speak of algorithms, chatbots, and other machine agents *as*

though they can engage in manipulation. Such talk might be understood as employing a helpful misnomer, engaging in a useful pretense, or something else along these lines.

One might advocate this sort of approach if one holds that machines – or, at least machines without genuinely human-like intentions – are incapable of manipulation. This position would likely be motivated by a desire to endorse (i) a strong *mens rea* condition on manipulation, combined with (ii) the claim that machine agents are unable to exhibit (presently, or perhaps ever) the sorts of intentions required to meet this strong *mens rea* condition.[13] Alternatively, one might advocate this sort of approach if one is not interested in the conditions that must be satisfied for something to count as manipulation but is concerned instead with the pragmatic question of whether it is beneficial to think and speak of a given phenomenon in that way.

Several strategies exist for explaining the function of the "loose talk" (as we'll generally call it) that we engage in when we call (at least certain) machines "manipulative". One possibility is that this is just another instance of our psychological tendency to anthropomorphize the nonhuman world. Just as we talk of thermometers "telling" us the temperature or the washing machine "deciding to play pranks", so too can we project a human-like representational/motivational structure onto machine agents or even algorithms. Such projections prove useful to the extent that such talk helps us make reasonable predictions about the behavior of such entities (e.g., by constructing and reasoning about a fictional correlate of the relevant entity) and helps us reflect on how best to integrate them into our broader social fabric. But we should not take such talk too seriously, for then we might go looking in vain for the metaphysical correlates of the sorts of anthropomorphized states we project onto these entities.

Another option would be to claim that what loose talk about "machine manipulation" serves to do is, not to improve our predictive abilities by anthropomorphizing those machine agents, algorithms and so on, but rather to fold them into our normative practices. So, the idea runs, we needn't pretend that the YouTube algorithm has anything like intentions and goals; rather, we talk about this algorithm "manipulating" us so that we can subject it to normative scrutiny, criticize its developers, consider how best to regulate it, and so on. This way of understanding things allows us to bypass any question of whether we are in fact prone to anthropomorphize algorithms, and it allows us to explain how to make sense of talk of "machine manipulation" even in cases where the individuals involved are not at all prone to engage in such anthropomorphizing. The point of such talk is not to engage in a pretense about understanding the function of algorithms (for example) by attributing to them human-like beliefs, desires, etc. – though undoubtedly some are apt to do just this. Rather, the point of such talk is to allow us to engage in a pretense which will hopefully yield a better understanding of the potential harms that machines can generate and to allow us

to think through who bears responsibility for those harms, how we ought to mitigate them, and similar practical questions.

One question for both these strategies is just how far we want to take them. We can imagine, for instance, that some might be tempted to think that anthropomorphizing can be explanatorily helpful with respect to some of the things that we tend to call "manipulative", but not with respect to others. So, for instance, perhaps it is useful to anthropomorphize algorithms because these tend to reflect the thought processes of programmers as they are working through a problem. Given this, algorithms might well have a tendency to parallel the structure of human cognition enough for such anthropomorphizing to prove useful to our understanding. Large-scale organizations, such as companies or nations, might not prove amenable to such explanations, on the other hand – so talking as though, for example, a tobacco company is "being manipulative" might just lead us into confusion. This would not mean that whatever we are trying to point to when we talk about manipulation by tobacco companies is not morally problematic or worth criticizing and regulating. It would only suggest that talk of such companies engaging in "manipulation" would, on this picture, in fact be unhelpful in the pursuit of that goal. Similar issues arise with respect to our normative practices: there doesn't seem to be any good way of knowing at the outset how productive it will be to engage in the pretence that we can treat this or that entity as a part of our normative practice.

To be clear, an error theorist about machine manipulation is also free to conclude that none of this talk of "manipulative machines" is actually helpful; perhaps, we would do better in understanding the moral contours of our interactions with algorithms, artificial agents, and the people and organizations behind them by setting the notion of manipulation entirely to the side. In that case, our talk of "manipulative machines" might turn out to be best understood as a part of a bad folk theory of morality. We are inclined to think that this is not the case but hardly take ourselves to have ruled out this possibility.

4 Ameliorative approaches to the concept of manipulation

The last type of approach that we wish to get on the table is an ameliorative approach to the concept of manipulation. In particular, we will consider an ameliorative approach based on the conservative analysis we adumbrated in Section 2. The approach is motivated by the rapidly evolving kinds of interactions that we humans have with machine agents, which may be headed toward the envisioned confrontations with superintelligences. In this ameliorative mood, we will consider the influence-centric approach not as a proposal concerning our actual, current concept of manipulation but as a proposal concerning which concept of manipulation would best serve the legitimate purposes of such a concept. We will only scratch the surface of a

full defence of this ameliorative approach, but it should be enough to provide a basis for future work.

In defence of his broader influence-centric concept of manipulation (which does not include the necessary condition we imposed in Section 2), Wood suggests something like an ameliorative outlook. He claims that because "manipulation by circumstances" has the same sort of limiting effect on a person's rational decision-making processes as deliberate manipulation by another person, a broader concept of manipulation that includes both is "more interesting" (Wood 2014, 27).[14] But whether or not this is the case depends on *why* we are interested in manipulation: in Haslanger's terms, it depends on which concept better serves the legitimate purposes of having such a concept (e.g., Haslanger 2000, 33). If the legitimate purposes of having a concept of manipulation are to help understand and prevent the generation of nonideal attitudes or nonideal decision-making, then the broader concept Wood endorses may better suit these purposes. On the other hand, if the legitimate purposes of such a concept include identifying entities (be they intentional agents or not) whose features make them distinctively suited to producing such influence, these purposes may be better served by a concept that is at least narrow enough to exclude manipulation by (to draw again on our examples from Section 2) the mud on the pavement and the stranger in the shop.

Although we cannot make a full case for it here, we think it is among the legitimate purposes of a concept of manipulation to identify certain entities as manipulators and not only to identify manipulative influence. One reason for this is that many things which can have manipulative influence in the senses defined, for example, by Barnhill or Woods, have this influence in what we might loosely call a "one-off" manner. In our examples, the mud on the pavement has an influence of this sort on Jane as she walks by but probably does not have such an influence on anyone else. The same is true for the influence that the stranger in the shop has on Daniel. Assuming that at least one legitimate purpose for a concept of manipulation is to prevent deleterious influence, it will be unhelpful to identify "chance-manipulators" like the mud or the stranger and try to prevent their manipulative activity. For these putative manipulators will be too diverse, too many, and preventing them will give too little bang for the buck. By focusing instead on entities whose manipulative features are sustained by their effectiveness at producing this influence, we will be in a position to give ourselves the conceptual resources to identify, classify, and thus block negative influence that is repeated and systematic.

Our proposed necessary condition on manipulation, when combined with an influence-centric account like Barnhill's or Wood's, would allow the concept to serve the purpose of identifying such manipulators. Thus, whether or not it contributes to an accurate analysis of our actual, current concept, it might contribute to one that better serves the purposes of such a concept.

Another feature of the ameliorative influence-centric approach is that it leaves intentional states on the part of the manipulator out of the concept of manipulation. Some might see this as a disadvantage for conservative conceptual analyses along these lines.[15] Whether this is the case or not, we think it is an advantage from an ameliorative point of view. One reason for this has to do with potential regulation of the activity of manipulative machines.[16] If one legitimate purpose of having a concept of manipulation is to identify entities poised to be systematic manipulators, this is presumably a legitimate purpose because it is legitimate to try to limit the manipulative activities of such entities. However, if only entities with intentional states (like human beings, for instance) can be manipulators, then the concept of manipulation will only help us to identify individuals whose manipulativeness is difficult, and most likely undesirable, to regulate. This is because regulating people's manipulativeness would require making highly fallible but legally binding judgments about the nature of their intentions and beliefs. On the other hand, if the concept also helps to identify machines, algorithms, and the like, then it would help to identify better candidates for having their activity regulated because of their manipulativeness. This would put law- and policy-makers in a better position to target the problems posed by current and future manipulative machines. Especially in light of the increased extension (in the Clark and Chalmers 1998 sense) of our mental activities via the internet and the blurring of the lines between human and machine in things like smart devices, a concept that doesn't commit itself to an epistemically or morally significant divide between the intentional and the non-intentional seems like it will serve us better.

A more general reason why a non-intentional concept of manipulation may better serve the concept's legitimate purposes is that in identifying manipulators, it moves us away from the difficult and potentially dangerous task of passing judgment on people's inner mental states (*mens rea*). Instead, this concept encourages a focus on the nature of someone's (or something's) influence and the factors that sustain that influence. These features are generally easier to assess in an objective and unbiased manner.

5 Conclusion

We have now charted part of the space of options for answering the question, "Can machines manipulate us?" which are available independently of an answer to the question whether machines can be genuinely intentional agents. The motivation for doing this was that the latter question is a perennial stumper, and deep commitments in the philosophy of mind and action are required even to begin to answer it. On the other hand, seemingly manipulative machines are a pressing concern, not just for the study of existential threat from AI but also for understanding and categorizing threats to people's autonomy and well-being in contemporary online life. In light of this predicament, we explored three ways of answering the question,

"Can machines manipulate us?" without positing that machines are (or are not) genuinely intentional agents. First, we set out an alternative concept of manipulation on which intentionality is not an essential condition for being a manipulator. Second, we sketched some strategies for understanding talk of machines manipulating us as "loose talk", coupled with either explaining away the sense that some of the example machines we discuss are manipulators, or maintaining that it simply does not matter whether machines can manipulate us or not, strictly speaking. Finally, we recast our alternative conception of manipulation, which we first presented as a conservative conceptual analysis of the current concept of manipulation, as an ameliorative account.

These approaches are not compatible, and we have not taken any stand on which is the right approach. As stated at the outset, our task here has been primarily to map out different ways to go. We hope that the map we have provided may serve as a launchpad for further investigation of machine manipulation and its relation (or lack thereof) to broader issues of machine intentionality. However, in this concluding section we would like to also give some preliminary reasons for thinking that the final approach we outlined, the ameliorative adoption of a concept of manipulation that does not make intentionality on the manipulator's part essential, has some significant advantages. We think it is the most promising line to pursue in this arena, though we certainly do not think the others should be cut off.

The central positive consideration we see in favor of articulating and adopting a concept of manipulation that does not make the manipulator's intentionality essential is this: doing so will enable us to bring together under a single concept a range of intuitively related phenomena that can threaten people's well-being in similar ways and to explain the nature of their intuitive relation. It will allow us to see how certain patterns or types of influence can be mirrored in different media and by different causally efficacious entities. The approaches we presented in Sections 2 and 4 would aim to provide one type of account of these similarities. A valuable future project would be to assess whether this explanatory sketch stands up to development and scrutiny or whether a different approach entirely is called for. At the same time as it promises an explanatory unification of seemingly manipulative influences from different sources (be they human beings, animals, machines or institutions), this approach also avoids the fancy footwork required to explain away the intuition that the behavior of machines like the hypothetical Oracle AI is manipulative. Taken together, we find these to be solid, though of course defeasible, reasons to seek a non-intentional concept of manipulation.

Moreover, while an influence-centric conservative analysis like the one we explored in Section 2 offers a notion of manipulation which allows for the possibility of manipulative machines, we suspect that it may not capture everything we intuitively associate with the concept of manipulation. We are

in fact skeptical that there really is a single concept here that we have all pre-theoretically internalized, rather than a cluster of closely related concepts. This motivates a shift from trying to generate a single best fit for this cluster, to asking instead: what in the vicinity will prove to be the most useful concept of manipulation? Or, at any rate, what will be the most useful concept for the purposes of addressing the seemingly manipulative behaviors of machines that we have discussed in this chapter? The influence-centric ameliorative analysis that we have sketched provides a promising start on answering this question.

Notes

1. The authors would like to thank the editors and participants in the Manipulation Online workshop for helpful feedback on this chapter. Special thanks to Michael Klenk for detailed comments. Work on this chapter was supported by a Swedish Research Council grant (VR2019-03154) and the Norwegian Research Council grant (303201).
2. The film *Ex Machina* offers one depiction of what this might look like.
3. Here we use "intentional" in the broad sense so that it characterizes a state of an organism or a system as representing, being directed on, or being about things. *Intentions,* in the sense of intentions to perform certain actions, are then just one type of intentional state.
4. See Alfano et al. (2020) for a detailed discussion of the algorithm's effects.
5. For a high-level overview of how AI (deep learning) works in Facebook advertising, see www.facebook.com/business/news/good-questions-real-answers-how-does-facebook-use-machine-learning-to-deliver-ads. As far as we can tell, the actual code is not public (understandably, since it's their entire business model).
6. This emerges in Manne's discussion of a case in which someone manipulates neglectful relatives into feeling guilty for their neglect by giving them elaborate gifts but is not conscious of doing this. Manne says that this case is still compatible with the manipulator having unconscious intentions to make the relatives feel guilty, and that she is "at least friendly to" the possibility "that there are genuine intentions which are at least to some extent unconscious". Despite being friendly to this possibility, Manne wishes to also leave open the opposite view, that there are no such intentions. She says that in the case she describes, the needed unconscious elements might be "motives" of some other sort. (See Manne 2014, 230–31, especially note 26.)
7. Perhaps, the negative claim implies that manipulation must be carried out by moral agents, so that their concern could be "insufficient" as opposed to simply absent. Certainly, a machine – or a rock, for that matter – can display an absence of concern for someone qua agent, but for this absence of concern to be "insufficient" in some respect, the machine would need to be required, perhaps morally required, to display a higher level of concern. At any rate, we will leave this issue aside.
8. See Aronson and Duportail (2018) for some discussion of this.
9. Manne writes: "without having a suitably manipulative end (albeit possibly unconscious), it seems plausible to think that [Joan's] actions would not count as being manipulative, although they might still leave her relatives *feeling* as if they had been treated manipulatively" (Manne 2014, fn 27). In general, she suggests that some sort of motive to influence the other in a certain way is required for an act to be manipulative. Plausibly, then, this motive combines with the

guilty-making features of extravagant gifts from a relative you neglect socially, to explain why the gift-giving occurs.

10. Marcia Baron (2014, fn 11) remarks in response to Wood that it is a stretch to say that the institution of advertising manipulates; we should prefer to say that advertisers or groups of advertisers manipulate. This seems a better response in the case of advertising than in the case of capitalism, where it would be difficult to pin the putative problems Wood enumerates on individuals or even groups of actors. We will not dwell on these matters here, as our aim is only to establish that to the extent there is plausibility to the claims of manipulation by institutions, this is because such cases differ from cases like the mud on the pavement. We think a basic difference is that the former satisfy the requirement articulated earlier (as we are about to argue) while the latter do not. As an aside, though, it is worth noting that social institutions beyond capitalism and advertising seem like candidates for manipulators. Varying cultural institutions of the family, marriage and child-rearing, for instance, have immense influence on people's attitudes and choices, often in ways that contravene their rationality and self-interest, without any individual or group of individuals being identifiable as the manipulator.
11. Another possible reaction to these cases would be to try to split apart the notion of manipulation from that of being manipulated. See, for instance, Klenk, in this volume.
12. One view is that Facebook is an artifact and that artifacts have the properties they do by virtue of being designed by some agent. One way to spell that out is in terms of affordances (Klenk 2020): artifacts have the property of affording behaviors. Facebook affords wasting time on it. But having affordances is a property determined by the designs of some agent, thus Facebook's manipulating one into wasting time on it could be causally downstream of the designer who programmed it to have the affordance of being something on which to waste time, and this seems close to the intentional model. But recall the dialectical context: we're assuming there can be manipulation without manipulators; and we're not taking any stance about the metaphysics of machines and to what extent, if at all, their properties are determined by agents. So we can stop with the intuitive enough claim that if there's systemic, non-agential manipulation, Facebook seems like a good candidate for such manipulation. Thanks to Michael Klenk for discussion here.
13. Klenk (2022), in this volume, discusses these under the heading of "sine qua non arguments".
14. Actually, Wood makes this comment about a broader concept of *coercion,* which would include being forced to do something by circumstances as well as by another person. Although he does not explicitly apply the same reasoning to the concept of *manipulation,* his discussion suggests that a broader concept of manipulation would be the "interesting" one for parallel reasons.
15. We have in mind those who think that some sort of manipulative intention or motive on the part of an entity is essential to an activity of that entity counting as manipulation, such as Baron and Manne, *op. cit*.
16. Thanks to Michael Klenk for encouraging us to consider the regulatory angle.

6 References

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

Armstrong, Stuart, and Xavier O'Rourke. 2018. "Good and Safe Uses of AI Oracles." https://arxiv.org/pdf/1711.05541.pdf.

Armstrong, Stuart, Anders Sandberg, and Nick Bostrom. 2012. "Thinking Inside the Box: Controlling and Using an Oracle AI." *Minds and Machines* 22 (4): 299–325.

Aronson, Polina, and Judith Duportail. 2018. "Can Emotion-regulating Tech Translate Across Cultures? | Aeon Essays." *Aeon Magazine*. Accessed August 23, 2021. https://aeon.co/essays/can-emotion-regulating-tech-translate-across-cultures.

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

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

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Chalmers, David. 2010. "The Singularity: A Philosophical Analysis." *Journal of Consciousness Studies* (17): 7–65.

Clark, Andy, and David J. Chalmers. 1998. "The extended mind." *Analysis* 58 (1): 7–19.

Cole, David. 2020. "The Chinese Room Argument." In *Stanford Encyclopedia of Philosophy: Winter 2020*, edited by Edward N. Zalta. https://plato.stanford.edu/archives/win2020/entries/chinese-room/.

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

Haslanger, Sally. 2000. "Gender and Race: (What) Are They? (What) Do We Want Them to Be?" *Nous* 34 (1): 31–55.

Haslanger, Sally. 2012. *Resisting Reality: Social Construction and Social Critique*. Oxford: Oxford University Press.

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

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

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

Searle, John. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3 (3): 417–57.

Turing, Alan. 1950. "Computing Machinery and Intelligence." *Mind* 59: 433–60.

Wallace, David Foster. 2004. *Oblivion: Stories*. London: Hachette UK.

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