Digital Well-Being and Manipulation Online

forthcoming in 'Ethics of Digital Well-Being: A Multidisciplinary Approach',

edited by Christopher Burr & Luciano Floridi (Springer).

Author information

Michael Klenk
Delft University of Technology
ORCiD 0000-0002-1483-0799
m.b.o.t.klenk@tudelft.nl

Abstract

Social media use is soaring globally. Existing research of its ethical implications predominantly focuses on the relationships amongst human users online, and their effects. The nature of the software-to-human relationship and its impact on digital well-being, however, has not been sufficiently addressed yet. This paper aims to close the gap. I argue that some intelligent software agents, such as newsfeed curator algorithms in social media, manipulate human users because they do not intend their means of influence to reveal the user's reasons. I support this claim by defending a novel account of manipulation and by showing that some intelligent software agents are manipulative in this sense. Apart from revealing a priori reason for thinking that some intelligent software agents are manipulative, the paper offers a framework for further empirical investigation of manipulation online.

Keywords

Digital well-being; persuasive technology; manipulation; intelligent software agents; digital ethics.

***

Social media usage is soaring globally: the average Internet user spends 2 hours and 15 minutes per day on social media (Global Web Index 2018), which amounted to about 30% of time spent online in 2017 (Young 2017). According to the (Internet World Statistics 2018), the number of social media users will rise to 3 billion people

in 2021, which would, based on current estimates, amount to almost 38% of the world's population using social media. So, more people than ever before are interacting with *intelligent software agents*, such as Facebook's newsfeed curator algorithm, regularly.

This article analyses the nature of our interactions with intelligent software agents and answers the question of whether some such interactions are manipulative, which has ramifications, as will be shown, for questions about digital well-being. In particular, I argue that manipulative software-to-human relationships are detrimental to well-being because they impugn the autonomy of human users.

The paper thereby aims at closing a research gap within digital ethics. Existing research on the ethics of social media has predominantly addressed two aspects: the nature of *user-to-user interactions online* and the *effects* of software-to-user interactions online.[1] However, the *nature* and *ethical status* of software-to-user interactions have received scant attention. With the advent of intelligent software agents (Burr and Cristianini 2019; Burr, Cristianini, and Ladyman 2018), that have a least some agency-like characteristics (Floridi and Sanders 2004), this is a significant omission. The paper explains and justifies the link between the nature of software-to-user interactions and digital well-being and provides conceptual clarity to the description and ethical evaluation of manipulative software-to-user interactions, which should enable further normative and empirical analysis of a ubiquitous form of online behaviour.

[1] Regarding the former, see, for example, discussions of the prevalence of user deception, e.g. Hancock and Gonzales 2013; Tsikerdekis and Zeadally 2014. Regarding the latter, see discussions of effects of social media use on user well-being, e.g. Huang 2017; Reinecke and Oliver 2016.

Finally, it aims to contribute to the study of manipulation itself, which, as one commentator puts it, is "in desperate need of conceptual refining and ethical analysis" (Blumenthal-Barby 2012, 345).[2]

I proceed as follows. After explaining and justifying the relation between digital well-being and online manipulation in more detail in section 1, I defend a novel account of manipulative action in section 2 using conceptual analysis.[3] According to the proposed account, a manipulative action consists of the attempt to exert directed but careless influence on someone (which will be made precise below). In section 3, I argue that we have reason to think that intelligent software agents are manipulative in this sense. Building on a framework for analysing software-to-user interactions introduced by Burr, Cristianini, and Ladyman (2018), I show that some intelligent software agents are manipulative because they aim to direct human users to specific actions while not intending to reveal reasons to them for doing so. A corollary of my argument is that it is *a priori* that some such interactions are manipulative because intentions to maximise a given behaviour necessarily crowd out intentions to reveal reasons for such behaviour.

In conclusion, the paper provides a novel analysis of manipulation, shows that some software-to-human interactions are manipulative, and explains how such interactions are detrimental to digital well-being.

---

[2] Though intelligent software agents indubitably exert an *influence* on human users, it is unclear whether that influence qualifies as, for example, persuasive, manipulative, or coercive; see (Alfano, Carter, and Cheong 2018). The concept of manipulation, in particular, is often used too coarsely, as something "in-between" persuasion and coercion; cf. (Faden, King, and Beauchamp 1986). In focusing on influences exerted by artificially intelligent software agents, the article goes beyond previous discussion of the ethics of persuasive technologies; cf. (Berdichevsky and Neuenschwander 1999; Fogg 1998; Spahn 2012).

[3] Note that the focus on manipulative action leaves open that an account of manipulat*ed* action contains further conditions that the one's defended here.

1 Digital well-being and online manipulation

Well-being in the broadest sense is what we have when we are living lives that are good for us (cf. Tiberius 2006). *Digital* well-being is concerned with the impact of technology on the extent to which we do and can lives that are good for us (Floridi 2014).[4] In this section, I explain and justify the relevance of online manipulation for well-being in three steps.

First, I motivate this paper's particular concern with software-to-human online interactions by sketching two preliminary reasons for thinking that intelligent software agents are manipulative. Second, I argue that the nature of software-to-human interactions, *in general*, is relevant (positively or negatively) for well-being. Finally, I suggest that manipulative interaction, *in particular*, is detrimental to well-being because of its effects on user autonomy.

In doing so, my aim is not to defend a particular theory of well-being; my aim is modest: To indicate how manipulation online is relevant for well-being on several prominent conceptions of well-being, which introduces the in-depth discussion in sections 2 and 3 of this paper.

1.1 Intelligent software agents and humans – signs of a troubled relationship

Our interactions with intelligent software agents are a source of ethical concern. As Burr, Cristianini, and Ladyman (2018, 756) note, the designers of the relevant technologies have themselves begun to raise warnings about their use. A statement by the ex-president of Facebook seems to show how the behaviour of

[4] It can also be asked whether technology may change what it *means* to live a good live; I address this question in an unpublished manuscript.

intelligent software agents was intentionally designed to manipulate (Pandey 2017, emphasis added, cited in Burr, Cristianini, and Ladyman 2018):

> The thought process that went into building these applications […] was all about *how do we consume as much of your time and conscious attention as possible*, and that means that we needed to sort of give you a little dopamine hit every once in a while, because someone liked or commented on a photo or a post or whatever, and that's going *to get you to contribute more content* […] It's a social validation feedback loop […] It's exactly the sort of thing a hacker like myself would come up with because you're *exploiting a vulnerability in human psychology*.

Design choices were made so that human users display the desired behaviour which, in the designer's own words, is exploitative of the users' vulnerability. More specifically, two features of intelligent software agents prompt an investigation of the software-to-user relationship with a view to possible manipulation, and its effects on user well-being.

First, intelligent software agents know a lot about the human user they interact with. It is now evident that intelligent software agents can learn a lot about human users from their behaviour online, even if no 'personal' information, such as one's name or address, are provided (cf. van den Hoven et al. 2018). Burr and Cristianini (2019) have argued that the reliability of inferences about users' beliefs, desires, personality, and behaviour is considerable.[5] The data trail that human users leave online certainly gives room to suspect that enough about the user's beliefs and desires is or can be known to make online manipulation credible (see also Buss 2005; cf. Alfano, Carter, and Cheong 2018).

[5] This does not necessarily mean that intelligent software agents have an accurate picture of a human user's true (digital) identity, which may be more fluid and performative, as Smith describes in this volume. Nevertheless, they have an accurate picture of what Smith calls a user's "corporatised identity." As Smith acknowledges, and as explained below, it seems that reliable inferences about a user's corporatized identity are sufficient for exerting considerable influence on that user.

Second, there is evidence that intelligent software agents can, to a considerable extent, steer the mental states and the behaviour of human users with whom they interact, plausibly based on the knowledge attained on them. Coupled with the widespread intuition that a manipulator likewise steers the behaviour of his victims, often suggested by the analogy that manipulators 'pull strings' of the manipulated patients, these findings also seem to suggest that software-to-user manipulation occurs (cf. Burr, Cristianini, and Ladyman 2018, 752ff). The influence on human users is not limited to behaviour but extends to emotions. By changing the presentation of content online, advertisers can purposefully influence the attitudes of human users (e.g. Kim and Hancock 2017).

Therefore, it is safe to say that some intelligent software agents have a tremendous amount of information about human users and that they can use that information to steer human users in desired directions. While this is indicative, it is admittedly speculative and still indeterminate as to the classification of the influence at hand (i.e. it leaves open whether the influence is e.g. manipulative). So, it needs to be determined whether the influence exerted by intelligent software agents qualifies as manipulation, and if so, how it could bear on user well-being.

1.2 How the nature of our interactions affects our well-being

I begin with the question of how manipulative action could bear on user well-being. Manipulation is an interaction with a specific nature. It can be shown that the nature of our interactions (that is, whether they are, for instance, persuasive, manipulative, seductive, or coercive) matters for (i.e. positively or negatively affects) our well-being. Hence, manipulative interactions may affect well-being.

Consider the four predominant philosophical approaches to well-being: hedonism, desire-based theories, objective list theories, and life-satisfaction theories (Parfit 1984, 493–502; cf. Tiberius 2006, 494). All allow that the nature of our interactions matters for well-being, either directly or indirectly.

On objective-list theories, such as Nussbaum's (2000) capabilities approach, some types of interactions, such as affiliations based on mutual care and respect, matter *directly* because they are final ends. That is, roughly, that we have reason to value such interactions for their own sake. For example, when someone makes an effort to persuade you of something rationally, they are appealing to your reasons and leave you the freedom to come to your conclusions; they show care and respect for you as a rational being, which directly matters for your well-being. So, proponents of objective-list theories of well-being can recognise the relevance of our interactions for well-being directly.

The relevance of our interactions for well-being, independently of its direct effects on pleasure or the satisfaction of one's desires, might seem more difficult to explain for hedonists, desire-based theorists, and life-satisfaction theorists.

However, the nature of an interaction plausibly matters instrumentally, even though it need not do so *directly* or *in the short run.* For example, the direct, short term effects of a given interaction might increase pleasure or satisfy one's desires, but there are indirect, long-term effects of the interaction, too. Different types of interaction may have different effects on our capacities for decision-making in the long run.[6]

[6] See Levy (2017).

For example, paternalistic interactions, in which one person makes decisions for another, may bring the paternalised person pleasure and desire-satisfaction in the short-run, but disable her ability to make fruitful decisions in the future. Hence, paternalistic interactions may be detrimental to pleasure indirectly, and in the long-run. In addition, empirical research has indicated that autonomy matters for people's life satisfaction and their well-being, hedonistically understood (cf. Reis et al. 2000).[7] Therefore, the relevance of our interactions for well-being should at least matter *indirectly* for hedonists and desire-based theorists, because it is plausible that the nature of our interactions influences its effects.[8]

Hence, the four prominent conceptions of well-being make some room for taking the nature of our interactions as a determinant of well-being seriously. The observations above suggest that determining the nature of software-to-human interactions matters for evaluating digital well-being.

1.3 Manipulative action is, more often than not, detrimental to well-being

We are left with the question of how, precisely, well-being is affected by manipulative interactions. I argue in this section that manipulative interactions are detrimental to well-being by undermining autonomy and so the threat of online manipulation is a threat to our well-being.

[7] See also Cavalo et al, this volume, specifically sections 1 and 2.

[8] I am assuming here that these claims about the relevance of the nature of interaction apply, ceteris paribus, to the relevance of the nature of software-to-human interactions, in particular. That inference might be faulty if relevant normative properties (e.g. the property of being manipulative) of interactions supervene on properties that are lacking, or cancelled out, in software-to-human interactions (e.g. the property of being intentional). Most importantly, we need to assume that intelligent software agents are, in the relevant sense, *agents* (cf. Floridi and Sanders 2004). For reasons of space, however, I cannot fully assess that assumption in this paper but the discussion in section 3 gives some reason to think that it is explanatorily useful to regard them as agents in the relevant sense, which may be sufficient in normative contexts, too.

Autonomy forms a part of many prominent objective-list theories of well-being (Ryff and Singer 1998; e.g. Nussbaum 2000). There are several reasons why proponents of objective-list theories consider autonomy as positively relevant for well-being. First, in treating someone so as to preserve her autonomy, we treat that person with respect for her *rationality* (Levy 2017, 498). That enables her, the thought goes, to develop and exert her capabilities as a rational being. Second, the relevance of autonomy for well-being is linked to moral *responsibility*. On many accounts of moral responsibility, being responsible is linked to being responsive to reasons (Fischer and Ravizza 1998). To manipulate someone is to fail to treat them as responsible agents and, therefore, detrimental to their reason-responsiveness.[9] But practising and developing rationality and responsibility are, on prominent objective list theories, what living a good life is all about (cf. Vallor 2016). Hence, autonomy in that sense is immediately relevant for well-being on objective list theories of well-being.

As before, hedonists and proponents of desire-based and life-satisfaction theories should recognise the relevance of autonomy *indirectly*. A lack of autonomy has been shown to affect happiness and life-satisfaction negatively, and it is at least not beneficial to the development of one's decision-making capabilities, which matters for desire-satisfaction (cf. Ryan and Deci 2000).[10]

[9] In addition, some have argued that autonomy is valuable independently from its relation to well-being; (cf. Sen 2011). So, even if the link between autonomy and well-being is doubted, there may be independent reason to be concerned about manipulation's impact on autonomy.

[10] More more detailed discussion about the concept of autonomy in relation to digital well-being, and specifically the concept of autonomy as understood within Ryan and Deci's Self-Determination Theory, see Calvo et al, in this volume.

To illustrate how manipulative action threatens to undermine autonomy, we need to get ahead of ourselves a little and preview some implications of my account of manipulation that I introduce in more detail in section 2.

According to my account of manipulative action, the manipulator *does not intend* to influence in such a way that he reveals his victim's reasons for doing as the manipulator wishes. In short, the manipulator *cares about something being done* but *does not care to show that there are good reasons to do it* (more on this in section 2).[11] If that account is valid, then we should expect that humans at the 'receiving ends' of manipulative relationships will have less opportunity to assess their reasons for acting, which, in turn, limits their capabilities, to assess and evaluate possibilities for thought and action (cf. Burr, Taddeo, and Floridi 2019). As shown above, that is bad for their well-being, on several prominent construals of well-being. So, we have reason to expect that manipulative action generally negatively affects autonomy, and, therefore, to generally negatively affect well-being.

Therefore, the relation of manipulation, autonomy and well-being suggests that it is an essential task for the scholarship on digital well-being to assess the degree to which (online) technologies are manipulative.

Admittedly, this assessment is merely preliminary because the precise nature of the relationship between manipulation and well-being depends on the true account of well-being. For now, it should suffice that we can find room for the

[11] Manipulation is sometimes defined in terms of autonomy (e.g. as autonomy undermining), but that is not an account that I defend (reference for criticism of autonomy account). But ISAs might undermine autonomy.

relevance of manipulation for well-being on several prominent accounts of well-being.

With a better view on the relationship between manipulation and digital well-being, we can now explain what manipulation is and show that it can be detrimental to well-being.

2 Manipulative action as directed and intentionally careless influencing

I defend the claim that manipulative action is intentional, *directed* influencing of a manipulatee, or patient of the manipulative action, coupled with *carelessness* about revealing the manipulatee's reasons for behaving as intended by the manipulator.[12] In other words, to act manipulatively is intending someone else to do something through a means that is not directed at revealing reasons to that agent.

More precisely, *directed influencing* means that the manipulator intends the patient of his manipulation to exhibit some particular behaviour – this is to exclude that *accidental* influences can count as manipulative.

Carelessness does not denote sloppiness, negligence, or failure about choosing a method that reveals reasons to the patient of one's manipulation but rather the utter disregard for even attempting to do so.[13] Manipulators are careless in the sense that they do not intentionally direct any effort at influencing their subjects

[12] The account is akin to a broader understanding of bullshitting (cf. Frankfurt 2005) applied to more than speech acts and de-coupled from truth. In that respect, the account is similar to Frankfurt's analysis of bullshit, because it makes do with a *disregard* for reasonability rather than requiring the intention to violate reasonability.

[13] Thus, carelessness ought to be understood in the sense in which someone might be careless or carefree about how people think about him, not directing any effort to trying to influence his public image in any way. The proposed sense of being careless is entirely compatible with actually and accidentally being crafty at creating a good public image.

in a way so that they can see the reasons for following suit – even though the manipulator might, actually and accidentally, reveal his patient's reasons to them.

We can now examine how the proposed account explains some paradigm cases of manipulation. Here are three examples of manipulative action:

Advertising The advertising manager for a home detergent wants to increase sales by conveying the product's superior cleaning power, compared to other home detergents, in illustrative video clips. That there is no evidence for the product's superior cleaning power is immaterial because the video clip is not aimed at revealing such reasons to clients anyway.[14]

Nudging The school board decides that the students of its schools should eat healthier and re-arranges the food display so that healthy foods are more cleverly displayed in school cafeterias.[15]

Children Little Daniel does not want to go to bed. His mother promises him chocolate the next day if he goes to bed now.

We have manipulative actions in all three cases, which, intuitively, is the correct result. The advertising manager, the school board, as well as Daniel's mother are acting manipulatively because the means of influence they chose, respectively, are not intended to reveal reasons to their patients (buyers of the detergent likely do not have such reasons, whereas pupils in the cafeteria, and little Daniel, likely do).

[14] As suggested earlier, the proposed account of manipulation purports to be morally neutral – in what follows, I do not consider whether and, if so, why manipulation is morally problematic.

[15] Recent meta-analyses on nudging effectiveness provides ample clues; see (Cadario and Chandon forthcoming).

The intuitive idea behind the account of manipulative action offered here is that to manipulate someone is to make that person do or believe something in a way that disregards any reasons that that person might eventually have for doing or believing so. So, as suggested by the Advertising case, it does not matter whether or not there are any reasons for the patient to act in the desired way. As suggested by the Nudging and Children case, the account does not require that there be no reason for the manipulated person to be doing something, only that the manipulator chooses a method of influence that is not intended to reveal any such reasons to the manipulatee.[16] More precisely:

Manipulative action: M aims to manipulate a patient S only if

a) M aims to have S exhibit some behaviour b through some method m

and

b) M disregards whether m reveals eventually existing reasons for b to S.

I will say that an agent acts manipulatively, or engages in manipulative action, or is manipulative whenever that agent meets the criteria for manipulative action. Manipulation, on this account, is a success term: The success criterion is whether or not a directed, careless influence is intended. It is not crucial whether the manipulator succeeds in fooling the patient, but only that he aims to do so. There can, in effect, be very bad manipulators.[17]

[16] I do assume that, until proven otherwise, we should regard manipulation as a unified concept; for criticism see (Ackerman 1995).

[17] In other words, agents that intend to manipulate (and, therefore, succeed) but fail to get their target victim to behave in their intended ways. Thanks to [redacted for blind review] for prompting me to clarify this point.

2.1 Details of the account

A few clarificatory remarks are needed about the notions of a method and how a method can reveal eventually existing reasons for the targeted subject. A method can be understood in a broad sense to include any action performed by M to make S exhibit some behaviour, such as a gesture or speech act. I use the term method rather than action because what might intuitively seem like non-actions, such as not reacting to another person's call, also count as a method in the relevant sense. In the right context, not acting is a bona fide form of influence as it provokes particular behaviour in others.

The requirement to 'disregard whether m reveals eventually existing reasons for b to S' is a tad more complicated and it helps to proceed step-wise. First, 'eventually existing reasons' is a modal term that suggests that there might or might not be reasons for S to exhibit some behaviour b. The critical point is that the manipulator does not aim to reveal any such reasons.

To explain this part of the second requirement, I focus on non-manipulative action and sketch manipulative action as its negation. During persuasion, or non-manipulative influence in general, one typically intends one's action to show to the target of one's behaviour that what one wants from them is reasonable. Persuasive interactions have not only causal, behavioural aims (e.g. to get someone to do or believe something), but normative ones too: they are aimed to get others to, in a sense, *see* that what one asks of others is reasonable or true.

The most obvious cases of persuasive influence are arguments (understood as actions). By 'providing an argument', one aims that the other takes up a certain belief and that one's action reveals reasons to the other. In such typical cases of

persuasion, the method m at once is intended to make a causal difference and a normative one, too, because it reveals reasons to the target for exhibiting the intended behaviour.

For persuasive action on my account, it is not required that the patient subject becomes aware of the reasons he or she has for performing the intended behaviour through that route – it is sufficient if the persuader intends so — many interactions and attempts at influence in the epistemic realm work through references to testimony. Consider, for example, claims like "You should stop smoking because I read that all the experts agree that smoking causes cancer," or that "You ought to believe that the moon landing took place because all credible sources say so." In both cases, a speech act is how one aims to exert directed influence, *and* that speech act is intended to convey reasons for complying. Hence, the action is not manipulative.

Manipulative action thus understood requires the manipulator to intentionally employ some way of influencing the target to effect a target behaviour and lack an intention about revealing to the patient any reasons that might exist to act by the manipulator's aims.

A typical manipulative action according to my account can be glossed as expressing the manipulator to be thinking roughly as follows: 'I want you to perform behaviour b, so I do m, and I would have chosen m even if it did not reveal your reasons for doing b to you.' The manipulator intends his influence to have a particular effect on the manipulatee, and chooses his influence accordingly, but he is oblivious to whether his chosen means of influence reveals any reasons for exhibiting the intended behaviour to the victim (it does not follow that the

manipulator is oblivious to whether there are reasons, for the manipulation patient, for following suit).

Some manipulators like parents and liberal paternalist choice architects do care about a manipulatee's reasons for exhibiting the intended behaviour, but they do not care about revealing these reasons through their chosen method of influence. Parents may not care because their children do not sufficiently grasp reasons yet. Paternalist choice architects may not aim for it because other means of influencing are more effective. In both cases, whether or not there are reasons for the manipulation patient to act, and whether they are revealed through the chosen method, is a mere side-effect.

2.2 Advantages of the account

Why should we accept the proposed account of manipulative action in analysing the behaviour of intelligent software agents? One reason is that the account offers advantages over alternative accounts of manipulation in the philosophical literature on manipulation.

To begin with, the account does not require any form of *deception* to be involved in manipulation. It is a common misconception that deception is always a component of manipulation. Because the account does not require deception to be involved, this is a reason in favour of the account. Barnhill's Open House case illustrates that speech acts are not needed for manipulation (Barnhill 2014, 58):

> **Open House**: Your house is for sale. Before holding an open house for prospective buyers, you bake cookies so that the house will smell like cookies, knowing that this will make the prospective buyers have more positive feelings about the house and make them more inclined to purchase the house.

Manipulative actions need not involve speech act (thus they need not involve stating falsities) and even making true claims that lead manipulatees to behave rationally can sometimes be manipulative rational and rational claims may be manipulative (Gorin 2014, 75; Barnhill 2014, 80). In cases like Barnhill's, the chosen method of influence (olfactory influence) does not reveal reasons to the manipulatee for showing the target behaviour (buying the house), but the manipulator chose that method nonetheless. Hence, this is a case of manipulative action.

Moreover, the account does not require the manipulatee to behave in less than ideal ways. Hence, it is possible to classify nudging as manipulative, which is plausible, at least in the non-moralised sense in which I use the term manipulation here. Nudges often do lead to behaviour that is closer to the ideal than 'un-nudged' behaviour, and other accounts of manipulation cannot make sense of the intuition that nudging is manipulative nonetheless. Hence, the view is an improvement over another popular view in the philosophical literature on manipulation, which entails that manipulation necessarily involves (attempting to) make the patient behave in less than ideal ways (Noggle 1996; Noggle 2018; Scanlon 1998).

Another reason for adopting the proposed account of manipulation is that it jibes well with concerns about *autonomy*, *harm*, and (frustrations of) *self-interest* that pervade attempts at analysing the concept of manipulation, without making these concerns necessary elements of manipulation.

The account captures the intuition that to persuade (as opposed to manipulating) is to take a certain interest in enabling the other person to

deliberate reasonably. In that sense, the account jibes well with previous discussions of a close link between autonomy and manipulation (cf. Coons 2014; Wood 2014; Frankfurt 1971), even though it does not spell out manipulation as the submission of autonomy.

Moreover, it explains how manipulative behaviour does not aim to help someone see how acting as the manipulator intends may be "keeping with their rational assessments of [an] outcome" (Kligman and Culver 1992, 186–87). It also explains how manipulative action often leads to harm and that it violates the self-interest of its victims because intending to reveal reasons for some behaviour is a (minimal) way for preventing one's target from performing harmful actions and from living up to its self-interest. Again, however, it does not make harm or self-interest part of the definition of manipulation, because that would exclude cases such as Nudging or Children from counting as manipulative.

Similarly, the account explains why manipulative actions often lead to the manipulated subject violate norms or rules (Noggle 1996), because this may be a side-effect of not revealing the reasons a subject has for performing certain actions. Again, however, that is not required for manipulative action, as illustrated by the cases discussed above.

Finally, the account does not require that the manipulator disregards whether S *has reasons* for doing x or believing y. Instead, the emphasis is on the method the manipulator uses and whether that method reveals eventually existing reasons to S.

I take the preceding considerations to show that the proposed account of manipulation has sufficient plausibility, and advantages over alternative

accounts, to take it as revealing some crucial elements of manipulation. This is sufficient to employ it in a study of software-to-human interactions.

It should be clear, however, that, on this broad account of manipulative action, many ways of interacting with others and influencing them count as manipulative action, perhaps more than what we are commonly used to expect.[18]

We can now use this account in investigating in evaluating the manipulativeness of ISAs (or, rather, the extent to which ISAs perform manipulative actions).

3 Intelligent software agents manipulate human users

Thus far, I have sketched a broad account of manipulative action. We can now put together observations about the behaviour of intelligent software agents and a clearer view of what manipulation is. I will show in this section that at least some intelligent software agents act manipulatively toward human users. I will also suggest that there are a priori reasons for thinking so. The argument can be formalised as follows:

1. If intelligent software agents attempt directed and careless influencing of human users, then intelligent software agents manipulate human users.

[18] I do not think, however, that this account implies that unreasonably many actions are instances of manipulation. People plausibly do engage in manipulative actions toward children (as argued above), but that need not be a morally problematic instance of manipulation. In many other instances where we do not take the care to muster a persuasive interaction and instead retort to a manipulative one, it seems correct to suggest that we are manipulative in these cases. In such cases, there seems to be a slight blemish in manipulating others: all things being equal (including, for example, the effectiveness of the method), having used a persuasive form of influence would have been preferable.

2. Some intelligent software agents attempt directed and careless influencing of human users

3. So, some intelligent software agents manipulate human users.

Premise 1 follows from the account of manipulative action given in section 2. The focus is now on defending premise 2.

3.1 An agent-based framework to study manipulation online

Following Burr, Cristianini, and Ladyman (2018), I analyse intelligent software agents as players in a game. Burr et al. suggest that there are three features of every such interaction (2018, 736):

> 1. The ISA has to choose from a set of available actions that bring about interaction with the user. For example, recommending a video or news item; suggesting an exercise in a tutoring task; displaying a set of products and prices, and perhaps also the context, layout, order and timing with which to present them.
>
> 2. The user chooses an action, thereby revealing information about their knowledge and preferences to the controlling agent, and determining the utility of the choice the ISA made.
>
> 3. The cycle repeats resulting in a process of feedback and learning.

The system is programmed to seek maximum rewards (to wit, to maximise its utility function) and its utility typically depends on the actions of the human user. Burr, Cristianini, and Ladyman (2018, 737) note that, in some cases, the utility function depends on the so-called click-through rate of the human user, which "expresses the probability of users clicking through links." They argue that there can be situations of cooperation and situations of competition, depending on whether the utility functions of intelligent software agents and human users are aligned or not, respectively (Burr, Cristianini, and Ladyman 2018, 740).

They go on to describe several different types of interaction between software agents and humans, amongst them coercive, deceptive, or persuasive interactions.[19] The important point is that some intelligent software agents aim to maximise user engagement: thus, their utility function depends on the probability of a user clicking on a given link. As Burr, Cristianini, and Ladyman (2018) indicate, learning about an intelligent agent's utility function is sufficient to learn about its intentions and beliefs.

These claims should be accepted conditional on the claim that this is indeed the right model of the intentions of intelligent software agents. As Burr et al. note, their model is an assumption about the way that intelligent software agents work, which may be challenged. For now, however, I take this to establish that some intelligent software agents are aiming at maximising user engagement. In other words, they indent to maximise a certain behaviour of human users.

3.2 Applying the agent-based framework

Based on this framework, and the account of manipulation defended above, we can ask whether the system can be engaged in the pursuit of maximising user engagement *while* intending to reveal thereby the human user's reasons for acting in this line.[20] If the answer is 'No' for a given intelligent software agent, then that

[19] Burr et al. break down persuasive behaviour into 'nudging' and 'trading' behaviour, and regard the former as more manipulative than the latter. Trading is defined as being a mutualistically beneficial interaction. My account of manipulation supports the statement that nudging is *more* manipulative than nudging because in order to trade one represents another's reasons for acting, which is not necessarily the case for nudging.

[20] The argument relies on a claim about intentions are mutually exclusive in the sense that one cannot, as a matter of conceptual possibility, intend things that are metaphysically exhaustive, thus, one cannot intend at the same time two things that are mutually exclusive. One cannot, for example, intend to leave the room and intend to remain in the room. That is because the meaning of an intention is to build up action potential in a certain direction, and that cannot be done in mutually exclusive directions.

agent acts manipulatively. Since being aware of and guided by one's reasons for acting is a component of (or at least correlated with) well-being, as illustrated in section 1, we would be missing a chance of enhancing user well-being through such interactions.

3.3 Empirical evidence for manipulation online

Work on existing intelligent software agents, reviewed by Burr, Cristianini, and Ladyman (2018) strongly suggests that some ISAs aim for maximum engagement and, therefore, they are unlikely to intend to adjust their behaviour to reveal reasons to the user.[21] Still, it is possible, of course, that the ISA reveals through its method of influence to the user reasons for doing as suggested by the ISA.

For example, the nutrition app *Cronometer* sends emails to customers that explicitly state that behaving in a certain way (e.g. logging at least one activity in the first week of using the app) reliably leads to a certain behaviour with the majority of their users (e.g. continued use of the app). Thus, the app's method of influence is to provide the user with a reason, and it can reasonably be argued that revealing this reason to the user is intentional. It might, however, be purely accidental that revealing reasons for acting in a certain way happen to be the most reliable method of getting the user to act. However, in that case, the intention is

[21] One might agree in principle with the claim that revealing reasons would often be superfluous to the initial aim of maximising utility for some ISAs (e.g. those that deploy some form of deception or nudging), but insist that revealing reasons would nevertheless be compatible with other forms of interaction (i.e. trading), especially if it helps establish trust with the user. The point is well taken: ISAs might instrumentally reveal reasons to the humans they interact with. After all, revealing reasons can be a valuable means, for example to increase human users' trust. ISAs are unlikely, however, to aim for reason-revealing as a final end. Thus, they are acting manipulatively. As discussed in section 1, that may not *directly* impact a user's well-being (at least not on hedonist or desire-satisfaction theories), but indirectly. I return to this point below. Thanks to Christopher Burr for raising this objection.

more aptly characterised as 'choosing a method that maximises the likelihood of desired action' rather than as 'choosing a method that reveals reasons for acting.'

Existing work on the aims and behaviour of intelligent software agents, therefore, suggests that at least some of them are manipulative vis-à-vis human users simply because the aim of maximising engagement crowds out the aim of revealing reasons.

That claim, however, may seem to rest on unsure footing (and, consequently, premise 2 of the argument defended in this paper remains open to criticism, too). An initial worry might be that intelligent software agents do not have real aims or intentions.[22] If they do not have intentions, then they cannot be manipulative. At the very least, however, their behaviour can be described as exhibiting intention-like states, such as aims, which is sufficient for the behaviour to count as manipulative.

A more serious problem relates to the empirical evidence base. Thus far, very few works have engaged with the intentions of the designers of intelligent software agents or the intelligent software agents themselves.[23] However, the account of manipulation defended in this paper requires that empirical investigations of online manipulation must focus on the intentions of intelligent software agents (the 'supply-side' of possible manipulative behaviour, so to speak), rather than the effects on human users (the 'output' side, so to speak). With current work predominantly focusing on the 'receiving' end of potentially manipulative action

[22] Some accounts of intentional action require to agent to be able to give an account of his intentions.

[23] The work of Burr, Cristianini, and Ladyman (2018) being an exception.

(i.e. the user), we lack more detailed accounts of what the potential 'supply-side' of software-to-human interactions intends.[24]

Thus, given that the aims of the manipulator are relevant, the directive should be to establish the intentions of intelligent software agents and, derivatively, their creators. The creators of intelligent software agents may be seen to use ISAs as a *method* of influencing human users.

In that sense, there might thus even be multi-layered accounts of manipulation. Such investigations can take many angles. One of them would be to analyse the business models of institutions that create intelligent software agents (Joseph 2018; Niu 2015). Analyses of a company's expressed aims, as well as their actual strategies, may allow reasonable inferences about the intentions or aims of intelligent software agents. The advertising-based business model of many companies that offer allegedly free services online, while also employing intelligent software agents, such as Facebook and Google, suggests that engagement maximisation is indeed their actual intention.

Thus, though the evidence base is currently still building, it seems reasonable to conclude that there is empirical support for the claim that some intelligent software agents attempt directed and careless influencing of human users. Thus, at least some intelligent software agents are manipulative.

[24] Of course, this is not to denigrate the importance of evidence about how users are affected by digital technologies. The present account of manipulative action leaves open the conditions for manipulat*ed* action, and studying the 'output' side will be crucial to ascertain whether users are manipulated by intelligent software agents. See the chapter by Calvo et al, in this collection, for a very useful case study of the impact of Youtube's recommender algorithm on user autonomy.

3.4 A priori evidence for manipulation online

In addition to the reviewed empirical considerations, a priori reasons are suggesting the manipulativeness of some intelligent software agents. That is because it is doubtful that intelligent software agents can intend to reveal reasons to human users because it is doubtful that they have a grasp of what reasons are. Their understanding of human behaviour can plausibly, and parsimoniously, be described as reduced to correlations between properties.

For example, an intelligent software agent may grasp that users with a property 'aged between 25 and 30,' often or reliably display a property 'interested in travelling,' as well as the property 'likes experiences not had before.' On that basis, the software agent can predict various decisions, for example that the prospects of travelling to new places will excite the user.[25] However, grasping such relations does not amount to grasping reasons, at least if reasons are understood as irreducible to and partly independent from the actual desires of an agent. On all but the most subjective accounts of reasons (which are caricatures), an agent's reasons are not reducible to the agent's present desires. 'Robust realist' account of reasons locate them outside the agent's desires (Scanlon 1998; Parfit 2011), and even though current subjectivists locate them in an agent's desires, they take only what might be called 'considered' desires to ground reasons (Schroeder 2007). In neither case is it possible to 'read off' an agent's reasons from the agent's behaviour, or the behaviour of comparable agents.

Thus, if it is true that intelligent software agents cannot grasp an agent's reasons for acting (or believing), then they cannot aim to reveal such reasons to

[25] Thanks to Stephan Jonas for discussion and helpful input on this point.

the agent. Since there are reasons to suspect that intelligent software agents cannot, in principle, grasp reasons for action, there is reason to suspect that their interactions with human are necessarily manipulative.[26]

It is worth emphasising that this does not mean that manipulative intelligent software agents are necessarily morally bad. The conceptual analysis of the term manipulation can and should proceed on the assumption that the concept is not completely moralised to begin with (cf. Wood 2014).

For that reason, a full moral evaluation of the actions of intelligent software agents is still outstanding. I have been deliberately careful in writing that the concept of manipulative action is not *completely* moralised, because it seems true to say that it is slightly moralised at least in the following sense. Given the direct and indirect effects of manipulation on well-being (on various prominent accounts of well-being, as discussed in section 1), it seems true that manipulative action is, ceteris paribus, worse for another's well-being than non-manipulative action.

Though well-being and moral goodness are distinct, it also seems plausible that manipulative action is, ceteris paribus and defeasibly, morally worse than non-manipulative action. That is because it fails to respect agents' rationality, or indirectly negatively affects final goods such as happiness.[27]

[26] See also the chapter by Smith, this volume, whose argument implies that intelligent software agents represent the identity of human users inaccurately in such way that the autonomy of human users is compromised by the interaction. Since there seems to many to be a close link between autonomy-subversion and manipulation, Smith's argument may provide another angle to argue for the necessary manipulativeness of intelligent software agents. However, given the account of manipulative action defended here, and more general considerations about manipulated behaviour that are beyond the scope of this paper, I doubt that there is such a tight conceptual link between autonomy-subversion and manipulation.

[27] Thanks to Christopher Burr for prompting me to clarify this point.

Thus, whenever an agent has the means to achieve a particular goal (such as getting someone else to do or believe a particular thing) and a non-manipulative method would be as efficient and safe, it seems that the non-manipulative method is to be preferred. Choosing a manipulative method instead would, therefore, constitute some moral failing, even though the degree of that failure is to be determined.

4 Conclusion

Social media use is soaring globally. Plausibly, some of that rise in popularity is due to the actions of intelligent software agents, such as newsfeed curators or recommendations engines. After defending an account of manipulation as the directed and careless influencing of manipulatees, the paper argued that there are both empirical as well as a priori reasons for thinking that at least some intelligent software agents are manipulative vis-à-vis human users.

This argument has ramifications for the debate about digital wellbeing. Insofar as manipulative action, by definition, lacks the intent to reveal to others the reasons for their action, the victims of manipulative action are at greater risk of action unreasonably or, even if they act reasonably, of being unaware of why they are acting reasonably. They might, therefore, miss out on valuable aspects of life. In conclusion, the nature of at least current software-to-human interactions is not conducive to digital well-being.

Future work should deepen the empirical insight into the intentions of (the designers of) intelligent software agents to determine the actual extent of

manipulation, for which the account of manipulation introduced in this paper offers a suitable starting point.[28]

[28] I am grateful to Christopher Burr for insightful comments on a previous draft, and to audiences at the 2019 Media Ecology Conference in Toronto and the Digital Behavioural Technologies Workshop in Munich for discussion of a previous version of this paper. My work on this paper was supported by a Niels Stensen Fellowship. In addition, work on this project was part of the project ValueChange that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 788321..

References

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

Alfano, Mark, J. Adam Carter, and Marc Cheong. 2018. "Technological Seduction and Self-Radicalization." *J. of the Am. Philos. Assoc.* 4 (3): 298–322. https://doi.org/10.1017/apa.2018.27.

Barnhill, Anne. 2014. "What Is Manipulation?" In *Manipulation: Theory and Practice*, edited by Christian Coons. Oxford: Oxford University Press.

Berdichevsky, Daniel, and Erik Neuenschwander. 1999. "Toward an Ethics of Persuasive Technology." *Commun. ACM* 42 (5): 51–58. https://doi.org/10.1145/301353.301410.

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.

Burr, Christopher, and Nello Cristianini. 2019. "Can Machines Read Our Minds?" *Minds & Machines* 83 (5): 1098. https://doi.org/10.1007/s11023-019-09497-4.

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

Burr, Christopher, Mariarosaria Taddeo, and Luciano Floridi. 2019. "The Ethics of Digital Well-Being: A Thematic Review." https://doi.org/10.2139/ssrn.3338441.

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

Cadario, Romain, and Pierre Chandon. forthcoming. "Which Healthy Eating Nudges Work Best? A Meta-Analysis of Field Experiments." *Marketing Science*. https://doi.org/10.2139/ssrn.3090829.

Coons, Christian, ed. 2014. *Manipulation: Theory and Practice*. Oxford: Oxford University Press.

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

Fischer, John Martin, and Mark Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge studies in philosophy and law. Cambridge: Cambridge University Press.

Floridi, Luciano. 2014. *Fourth Revolution: How the Infosphere Is Reshaping Human Reality*. Oxford: Oxford University Press USA.

Floridi, Luciano, and J. W. Sanders. 2004. "On the Morality of Artificial Agents." *Minds and machines* 14 (3): 349–79. https://doi.org/10.1023/B:MIND.0000035461.63578.9d.

Fogg, B. J. 1998. "Persuasive Computers: Perspectives and Research Directions." In *Making the Impossible Possible: 18 - 23 April Los Angeles ; CHI 98 Conference Proceedings*, edited by Clare-Marie Karat, 225–32. New York, NY: ACM Press.

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

Frankfurt, Harry G. 2005. *On Bullshit*. Princeton, NJ: Princeton University Press.

"Global Web Index Social." 2018. Unpublished manuscript, last modified May 03, 2019.

Gorin, Moti. 2014. "Towards a Theory of Interpersonal Manipulation." In *Manipulation: Theory and Practice*, edited by Christian Coons, 73–97. Oxford: Oxford University Press.

Hancock, Jeff, and Amy Gonzales. 2013. "Deception in Computer Mediated Communication." In *Pragmatics of Computer-Mediated Communication*, edited by Susan C. Herring, Dieter Stein, and Tuija Virtanen, 363–83. Handbooks of pragmatics / eds. Wolfram Bublitz; Andreas H. Jucker; Klaus P. Schneider ; Volume 9. Berlin: de Gruyter Mouton. Accessed May 29, 2019.

Huang, Chiungjung. 2017. "Time Spent on Social Network Sites and Psychological Well-Being: A Meta-Analysis." *Cyberpsychology, Behavior, and Social Networking* 20 (6): 346–54. https://doi.org/10.1089/cyber.2016.0758.

Internet World Statistics. 2018. "World Internet Users Statistics and 2018 World Population Stats." Accessed June 02, 2018.

Joseph, Sarah. 2018. "Why the Business Model of Social Media Giants Like Facebook Is Incompatible with Human Rights." http://theconversation.com/why-the-business-model-of-social-media-giants-like-facebook-is-incompatible-with-human-rights-94016.

Kim, Sunny Jung, and Jeff Hancock. 2017. "How Advertorials Deactivate Advertising Schema: MTurk-Based Experiments to Examine Persuasion Tactics and Outcomes in Health Advertisements." *Communication Research* 44 (7): 1019–45. https://doi.org/10.1177/0093650216644017.

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

Levy, Neil. 2017. "Nudges in a Post-Truth World." *Journal of medical ethics* 43 (8): 495–500. https://doi.org/10.1136/medethics-2017-104153.

Niu, Evan. 2015. "This Company Has the Best Business Model in Social Media." Accessed November 30, 2017.

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

Noggle, Robert. 2018. "The Ethics of Manipulation." In Zalta 2018.

Nussbaum, Martha Craven. 2000. *Women and Human Development: The Capabilities Approach*. The Seeley lectures 3. Cambridge: Cambridge University Press.

Pandey, E. 2017. "Sean Parker: Facebook Was Designed to Exploit Human "Vulnerability"." https ://www.axios .com/sean-parke r-faceb ook-explo its-avulne rabil ity-in-human s-25079 17325 .html.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Clarendon.

Parfit, Derek. 2011. *On What Matters*. Oxford: Oxford University Press.

Reinecke, Leonard, and Mary Beth Oliver, eds. 2016. *The Routledge Handbook of Media Use and Well-Being*. New York, NY: Routledge.

Reis, Harry T., Kennon M. Sheldon, Shelly L. Gable, Joseph Roscoe, and Richard M. Ryan. 2000. "Daily Well-Being: The Role of Autonomy, Competence, and Relatedness." *Pers Soc Psychol Bull* 26 (4): 419–35. https://doi.org/10.1177/0146167200266002.

Ryan, Richard M., and Edward L. Deci. 2000. "Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being." *American Psychologist* 55 (1): 68-78. Accessed September 09, 2019.

Ryff, Carol D., and Burton Singer. 1998. "The Contours of Positive Human Health." *Psychological Inquiry* 9 (1): 1–28. https://doi.org/10.1207/s15327965pli0901_1.

Scanlon, Thomas. 1998. *What We Owe to Each Other*. 3rd print. Cambridge, MA: The Belknap Press; Harvard University Press.

Schroeder, Mark Andrew. 2007. *Slaves of the Passions.* New York, NY: Oxford University Press.

Sen, Amartya. 2011. *The Idea of Justice.* Cambridge, MA: Harvard University Press.

Spahn, Andreas. 2012. "And Lead Us (Not) into Persuasion…? Persuasive Technology and the Ethics of Communication." *Sci Eng Ethics* 18 (4): 633–50. https://doi.org/10.1007/s11948-011-9278-y.

Tiberius, Valerie. 2006. "Well-Being: Psychological Research for Philosophers." *Philosophy Compass* 1 (5): 493–505. https://doi.org/10.1111/j.1747-9991.2006.00038.x.

Tsikerdekis, Michail, and Sherali Zeadally. 2014. "Online Deception in Social Media." *Commun. ACM* 57 (9): 72–80. https://doi.org/10.1145/2629612.

Vallor, Shannon. 2016. *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting.* New York, NY: Oxford University Press.

van den Hoven, Jeroen, Martijn Blaauw, Wolter Pieters, and Martijn Warnier. 2018. "Privacy and Information Technology." In Zalta 2018.

Wood, Allen W. 2014. "Coercion, Manipulation, Exploitation." In *Manipulation: Theory and Practice*, edited by Christian Coons, 17–50. Oxford: Oxford University Press.

Young, Katie. 2017. "Social Media Captures over 30% of Online Time." Accessed November 30, 2017. https://blog.globalwebindex.net/chart-of-the-day/social-media-captures-30-of-online-time/.

Zalta, Edward N., ed. 2018. *Stanford Encyclopedia of Philosophy*. Summer 2018.