Published online Internet Policy Review (2019), at https://policyreview.info/articles/news/autonomy-and-online-manipulation/1431

Autonomy and Online Manipulation

Michael Klenk[1] & Jeff Hancock[2]

***

The public is increasingly concerned with the abilities of data collectors, like Facebook and Google, to understand and influence individual users. Most data collectors rely on online technologies to do so. We define online technologies as connected data-gathering software, like social media algorithms, or hardware, like smartwatches, that interact with users. For example, by sending users push-notifications or compiling content based on user preferences.

This public concern has resonated in academia. More and more researchers argue that online technologies manipulate human users and, therefore, undermine their autonomy. We call this the MAL view on online technology because it argues from Manipulation to Autonomy-Loss. MAL enjoys public visibility and will shape the academic discussion to come.

This view of online technology, however, fails conceptually. MAL presupposes that manipulation equals autonomy loss, and that autonomy is the absence of manipulation. That is mistaken. In short, an individual can be manipulated while being fully personally autonomous.

Internet policy researchers should be aware of this point to avoid looking in the wrong place in future research on manipulative and harmful online

[1] M.b.o.t.klenk@tudelft.nl. Niels Stensen Fellow & Visiting Postdoctoral Scholar, Social Media Lab, Stanford University.

[2] hancockj@stanford.edu. Professor of Communication, Director of the Social Media Lab, Stanford University.

technology. Showing that manipulative online technology leads to autonomy-loss requires empirical testing, or so we will argue.

1 Reconstructing the Manipulation to Autonomy-Loss (MAL) view

We will illustrate MAL in more detail by discussing a recent article by Daniel Susser, Beate Roessler, and Helen Nissenbaum in this journal (2019). Their article presents a well-informed and lucid account of the potentially harmful effects of online technology. Since they articulate their assumptions about the relationship between manipulation and autonomy lucidly, their article helps to illustrate what is mistaken about the MAL view of online technology.

Susser et al.'s argument involves three crucial claims. (1) A claim about the influence of online technologies on users (call this INFLUENCE). (2) A claim about the manipulativeness of online technologies (call this MANIPULATION). (3) The MAL claim, according to which manipulation equals autonomy loss. Schematically, their argument goes as follows:

INFLUENCE:	Online technologies influence human users.
MANIPULATION:	Online technologies are manipulative.
MAL:	If an influence is manipulative, then it is autonomy undermining.
CONCLUSION:	So, online technologies are autonomy undermining.

In support of INFLUENCE, they note how data collectors compile our online traces "into enormously detailed profiles," which can then be used by "advertisers and others engaging in behavioural targeting […] to detect when and how to intervene in order to most effectively influence us" (p. 6, page numbers refer to

Susser et al.'s article). Moreover, they suggest that "digital surveillance enables detection of increasingly individual- or person-specific vulnerabilities," including the exploitation of cognitive biases and other needs (ibid.). We take INFLUENCE to be well supported.

They defend MANIPULATION by defining online manipulation as follows. Manipulation is "the use of information technology to covertly influence another person's decision-making, by targeting and exploiting decision-making vulnerabilities" (p. 6). They then argue that, given *Influence*, online technologies plausibly manipulate users.[3] Importantly, they claim that to exploit individuals' decision-making vulnerabilities is to fail to "encourage individuals to slow down, reflect on, and make more informed choices" (ibid.).

Finally, in support of MAL, they write that "manipulation violates its target's autonomy" (p. 8). To unpack this claim, need to introduce their account of autonomy and then explain how manipulation jeopardises it. They define personal autonomy with two conditions. (1) One has the competencies (cognitive and affective) to consider one's choices and to act upon them. (2) One reflectively endorses the ends (e.g. goals) and grounds (e.g. reasons) of one's actions (pp. 7-8). They then establish the connection between manipulation and autonomy as follows. (Online) Manipulation, they write, "undermines a target's autonomy in two ways: first, it can lead them to act toward ends they have not chosen, and second, it can lead them to act for reasons not authentically their own" (p. 9).[4]

[3] Though their definition of manipulation raises a number of critical conceptual questions (for example, whether covertness is a necessary condition of manipulation, and whether manipulation is *pro tanto* bad, as their definition suggests), the important point for our argument is how they link manipulation and autonomy.

[4] They also note that autonomy-loss may lead to further harms. This claim is not subject to our criticism.

In conclusion, Susser et al. argue that online technologies frequently manipulate and, therefore, undermine users' autonomy, which they consider morally wrong in most cases. Thus, they claim that potential autonomy loss explains why "online manipulation poses such a grave threat" (p. 9).

2 Clarifying the MAL view of online technology

Susser et al. do not make the nature of the manipulation-autonomy connection explicit. What they write leaves open two options. A *contingent* reading of the claim (roughly, manipulating S often or mostly undermines S's autonomy). A *necessary* reading (roughly, manipulating S always undermines S's autonomy).

The contingent reading is the weaker claim, because it allows for more exceptions, and thus the more charitable reading of their argument. However, they explicitly define manipulation as covertly influencing someone so that they *fail* to "slow down reflect on, and make more informed choices" (p. 6). So, it becomes hard to see how there could be genuine cases of manipulation (on their account) without autonomy loss (again, on their account of autonomy).

Moreover, they consider it a sign of manipulation that "one did not understand one's motivations" (p. 4) and that one was "directed, outside one's conscious awareness, to act for reasons one can't recognise, and toward ends one may wish to avoid" (p. 4).

There are thus clear signs that support interpreting Susser et al.'s endorsement of MAL as a necessary conceptual link between manipulation and autonomy-loss.

3 Challenging the MAL view of online technology

The move from manipulation to autonomy-loss does not stand up to scrutiny. To see why it helps to look at the conditions for MAL to be true. MAL is a view of a *conceptual link* between manipulation and autonomy-loss. It says that whenever one finds something that is manipulative, one has found something that is autonomy-undermining.

There are many such conceptual links. For example, whenever one encounters a bachelor, one encounters an unmarried man – the concept 'bachelor' implies the concepts 'unmarried' and 'man.' However, there is a danger of jumping to conclusions here. We should be wary of letting contingent empirical observations confuse our claims about conceptual necessity. For example, it is an empirical fact that, say, many or most bottles are plastic. Nevertheless, we cannot conclude that the concept 'bottle' implies the concept of 'plastic.' The relation is empirical, not conceptual.

The lesson is this. In an argument about bachelors, we only need to show that someone is a bachelor to get the result that he is unmarried 'for free,' by courtesy of a conceptual link. However, in an argument about bottles, we do not get the claim that a given bottle is made from plastic 'for free,' because there is no conceptual link between 'bottle' and 'plastic.'

The MAL view makes the same mistake. It suggests that there is a necessary conceptual link between manipulation and autonomy-loss. But that is mistaken. There are cases of manipulation that are not autonomy-undermining, on *any* plausible understanding of personal autonomy.

Susser et al. understand personal autonomy in broadly *externalist* terms. According to externalist approaches, personal autonomy comes down to the extent to which the agent can appreciate and endorse her reasons for acting. The intuition behind externalist approaches is as follows. A person cannot wholly 'own' her actions, or act for reasons "authentically their own" (p. 9), insofar as she does not take some appropriate attitude like endorsement or understanding toward her reasons for acting. We will look at the two most influential externalist accounts in philosophy.

One prominent externalist conception of autonomy goes as follows. The ability to assess and chose an action is fleshed out as an agent's ability to evaluate her motives based on whatever else she believes and desires, and to adjust her motives in response to these evaluations (Christman, 1991). For example, indoctrinated people are not autonomous. Their indoctrination prevents them from evaluating doctrine in light of their own (potentially) critical beliefs and emotions. Susser et al. credit this conception as the basis of their account.

There is an alternative externalist conception. On this view, the ability to assess and chose an action has been fleshed out as an agent's ability to appropriately respond to a sufficiently wide range of reasons for and against behaving as she does (Fischer & Ravizza, 1998). For example, there are reasons for and against pursuing a challenging career (e.g. personal reward vs less family time). One acts autonomously when one can 'feel the pull' of both reasons for and against a particular act.

Neither conception of externalist autonomy implies that manipulation is incompatible with autonomy. Consider the following example:

Breakthrough: Johannes cherishes autonomy above everything else, and he wants others to be autonomous, too. He creates a self-optimisation app called *Breakthrough* that helps users to free themselves of societal expectations and conventions and to determine for themselves the lives they want to live. *Breakthrough* reminds users of their goals. It points out how societal expectations may have contributed to their choice. It also creates opportunities for users to reflect on and potentially revise their motives and goals. It does so in light of the user's motives and also in light of what the app's advanced algorithm deems good reasons for doing something, e.g. eating healthier. The ultimate aim is for users to *breakthrough*. To abscond any habitual, unconsidered, socially-influenced action so that they take any action with full emotional and cognitive endorsement, in line with all their ends and grounds. Cordula is an avid user of the app and eventually *breaks through*. She would not have thought how much the app would change her life. Amongst other things, she stops seeing several long-term friends, to whom her relationship seemed merely conventional and not genuine, to focus prepping for a triathlon. She is ok with that, however, because she prefers being fully autonomous to her former life.

Cordula is autonomous according to either externalist conception of autonomy. She responds well to reasons (e.g. reasons for eating healthier) and to reasoning (e.g. to eat healthier, given that she wants to be healthier and committed to that goal). Indeed, that is the very aim of the Breakthrough app and the very reason that Cordula is using it. Nevertheless, Cordula seems to be manipulated by the *Breakthrough* app. There is a sense in which she lives a life

that is authentically hers, because of the way she reflects on and endorses her motives and reasons. However, there is also a nagging sense that she may have given up too much of her life to the *Breakthrough* app. The app seems to exert an overpowering and illegitimate influence on her behaviour. She seems manipulated by *Breakthrough*. Therefore, manipulation need not undermine externalist autonomy, contrary to Susser et al.'s argument.

That observation generalises and thus relies less on potentially problematic intuitions about particular cases. If someone, like Cordula, reflectively endorses an action (like eating healthier), we can always ask how she arrived at her endorsement. We can then ask whether *those* grounds are authentically hers. And so on – into a regress. Manipulation can sneak in anywhere in that line. Externalist accounts must allow it on pain of raising the bar much too high for autonomous action (cf. Gorin, 2014, p. 89).

Let us recap. Susser et al. defend the MAL view, the view that online technology manipulates and, therefore, undermines autonomy. They defended this view on the assumption that manipulation equals autonomy-loss and they understood autonomy externalistically. We suspect that Susser et al. are not the only ones who embrace the MAL view on online technology. Other scholars also operate with a broadly externalist conception of autonomy and suggest that manipulative online technology undermine autonomy, though often less explicitly. For example, Frischmann and Selinger see autonomy in an externalist light as they link it to unhampered "self-reflection" and the ability to "determine one's own intentions" (Frischmann & Selinger, 2018, 18, 153). They see that type of

autonomy in jeopardy as "we're being conditioned to obey" by online technologies (2018, pp. 4–6).

However, as we have shown, both intuitive cases and general theoretical considerations suggest that manipulation does not necessarily undermine autonomy on an externalist understanding. Manipulation *does not* equal autonomy-loss. The MAL view on online manipulation fails.[5]

4 Implications for internet policy research

The failure of the MAL view of online technology has three crucial implications for internet policy researchers interested in online technology and autonomy-loss.

First, one should do *better conceptual work* to understand manipulation in such a way that manipulation does not equal autonomy-loss (cf. Klenk, forthcoming). The consequences of classification are not merely terminological but practical. *Manipulative* technologies would, and should, be subject to different policies than *non-manipulative* technologies.

Second, one could do *additional conceptual work* to identify conceptual links to go from the influence of online technology to autonomy-loss. The concept of manipulation will not be able to do this work. But there might be others, like coercion.

Third, one should do *empirical work* on the experiences that lead people to feel their autonomy compromised in the context of online technology. MAL depends entirely on the conceptual link between manipulation and autonomy-loss.

[5] MAL's presupposition that manipulation equals autonomy-loss also fails on coherentist accounts of autonomy, which are the main alternative to externalist accounts. Unfortunately, we have no space to explain in detail how coherentist autonomy is compatible with manipulation. See Klenk (forthcoming) for further discussion.

Since we cut that link, we need new ways to show that online technologies subvert autonomy, if they do. This goes to show that this is not this just a semantic worry about the meaning of the word manipulation or the concept of manipulation. At stake is the genuine problem of how online technologies affect autonomy.

5 Conclusion

Online technology can manipulate us without compromising our autonomy. It is plausible that manipulation is compatible with autonomy, and that autonomy-loss can come by other means than manipulation. Hence, the MAL view of online technology, and Susser et al.'s argument that depends on it, fail.

Several other scholars (e.g. Zuboff, 2019) make an equally problematic assumption about the link from Autonomy-loss to Manipulation (what we call the ALM view). If our argument in this paper is any indication, the ALM view is ripe for a reality check, too.

Going forward, we will need more *conceptual* work on the concept of (online) manipulation, and more *empirical work* to test its links to autonomy(-loss).[6]

[6] Thanks to Sunny Xin Liu for helpful comments on an earlier draft of this paper. Michael Klenk's work on this paper was funded by a Niels Stensen Fellowship.

References

Christman, J. (1991). Autonomy and personal history. *Canadian Journal of Philosophy*, *21*(1), 1–24.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge: Cambridge University Press.

Frischmann, B. M., & Selinger, E. (2018). *Re-engineering humanity*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316544846

Gorin, M. (2014). Towards a theory of interpersonal manipulation. In C. Coons & M. Weber (Eds.), *Manipulation: Theory and practice* (pp. 73–97). Oxford: Oxford University Press.

Klenk, M. (forthcoming). Digital well-being and manipulation online. In C. Burr & L. Floridi (Eds.), *Ethics of Digital Well-Being: A Multidisciplinary Approach*. Retrieved from https://philpapers.org/rec/KLEDWA

Susser, D., Roessler, B., & Nissenbaum, H. (2019). Technology, autonomy, and manipulation. *Internet Policy Review*, *8*(2), 1–22. https://doi.org/10.14763/2019.2.1410

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. New York, NY: PublicAffairs.