Liberty, Manipulation, and Algorithmic Transparency

Reply to Franke

Klenk, Michael

DOI
10.1007/s13347-024-00739-7

Publication date
2024

Document Version
Final published version

Published in
Philosophy and Technology

Philosophy & Technology (2024) 37:48
https://doi.org/10.1007/s13347-024-00739-7

COMMENTARY

Liberty, Manipulation, and Algorithmic Transparency: Reply to Franke

Michael Klenk[1]

Received: 22 March 2024 / Accepted: 26 March 2024

Abstract
Franke, in Philosophy & Technology, 37(1), 1–6, (2024), connects the recent debate about manipulative algorithmic transparency with the concerns about problematic pursuits of positive liberty. I argue that the indifference view of manipulative transparency is not aligned with positive liberty, contrary to Franke's claim, and even if it is, it is not aligned with the risk that many have attributed to pursuits of positive liberty. Moreover, I suggest that Franke's worry may generalise beyond the manipulative transparency debate to AI ethics in general.

Keywords Transparency · Algorithms · AI · Manipulation · Power

1 Introduction

Algorithmic systems drive automated decisions in important fields such as medicine, politics, finance, and warfare. There are enormous benefits, but also grave ethical concerns (Buijsman et al. forthcoming). Chief among them is the call for automated decisions to be explainable and for decision sources and criteria to be transparent (Felzmann et al., 2020). However, Wang (2022, 2023) and Klenk (2023) argue that algorithmic transparency can be manipulative, thus revealing a dark side of manipulative transparency.

In a recent contribution, Franke (2024) raises a challenge for these accounts of manipulative algorithmic transparency. He claims that concerns about manipulative transparency align with positive liberty (to wit, the ability to live according to one's higher self), and pursuing positive liberty can lead to illegitimate, authoritarian regulation.

While I agree with the spirit of Franke's observation (cautioning against illegitimate moralising), I point out that his claim is unsubstantiated in the case of

✉ Michael Klenk
M.B.O.T.Klenk@tudelft.nl

[1] Department of Values, Technology and Innovation, TU Delft, Jaffalaan 5, 2628 BX Delft, The Netherlands

Published online: 02 April 2024

the indifference account of manipulation. Moreover, I suggest that his worry may overgeneralise.

2 Reconstructing Franke's Thesis

2.1 Algorithmic Transparency and the Manipulation Risk

Algorithmic transparency is achieved when information about the automated decision's input data and genesis is disclosed accurately, reliably, and comprehensively (cf. Durán & Jongsma, 2021). Wang (2022) and Klenk (2023) argue that the *means* to achieve algorithmic transparency can be manipulative.

To achieve algorithmic transparency, the deployers of algorithms have to inform users about an algorithm's decision by, for example, disclosing information on websites, creating instructional videos, and providing other forms of suitable explanations. In doing so, however, they may manipulate their audience. Since manipulation is a morally dubious form of social influence, the manipulation risk highlights a potential dark side of algorithmic transparency.

Notably, the views of Wang and Klenk differ substantially, although they align in their conclusion about the manipulative potential of algorithmic transparency. Wang (2022, 2023) relies on a broadly Foucauldian account and understands manipulation as an intentional effort to exploit vulnerabilities in the target. Klenk (2023) shows this 'vulnerability view' to be flawed and suggests the 'indifference view' of manipulation instead (see also Klenk 2020, 2021, 2022a, 2022b, 2024, Klenk & Jongepier 2022). According to the indifference view, the means to achieve algorithmic transparency can turn into manipulation when they are designed to be effective without regard to revealing reasons to the audience.

For example, an organisation may provide an explanation about a credit rating algorithm to meet legal demands and to build trust in users. However, if they did not design their explanation to reveal reasons for the system's trustworthiness, then the explanation is manipulative (cf. Klenk, 2023). Their explanation is, in the relevant sense, indifferent and thus manipulative: it is designed to effectively build trust in users while being indifferent about *how* that goal is reached.

2.2 Franke on the Paradox of Positive Liberty and Manipulative Transparency

Franke (2024) makes two novel claims about accounts of manipulative transparency. First, protecting against manipulative transparency is aligned with fostering positive liberty, and second, this is a cause of concern because there is a considerable risk of error when trying to promote positive liberty.

He leans on Berlin's (1969) distinction between negative and positive liberty. Negative liberty is to be free from external constraints. In contrast, positive liberty is living in line with one's 'higher self.' As Carter (2022) helpfully puts it, negative liberty is about how many doors are open to you, while positive liberty is about the reasons for which you go through those doors. Being free in

the positive sense would roughly mean going through those doors because your 'higher self' would do so.

Following Berlin (1969), Franke points out a risk with positive liberty. Insofar as it seems easier to identify constraints on people's behaviour (i.e. identify and curb threats to negative liberty) than constraints on their higher selves (i.e. identify threats to positive liberty), there is a greater risk of erring in promoting positive liberty. Moreover, bad actors can instrumentalise uncertainty about protecting the higher self by claiming they have privileged insight and proposing self-serving measures (Franke, 2024, p. 3). Berlin (1969), and others, realise both the value of positive liberty, and its inherent risk – hence the paradox of positive liberty (Carter, 2022).

Franke's (2024) original contribution to the debate manipulative transparency lies in drawing the connection between the paradox of positive liberty and the view that algorithmic transparency can be manipulative.

Franke finds that Wang's (2022, 2023) vulnerability view is "well aligned" with concerns about positive liberty because it cautions that people may fail to act in line with their true interests as a result of manipulative transparency (Franke, 2024, p. 4). He finds that caution should "perhaps be even higher" for the vulnerability view than the indifference view since the former is "more directly connected to positive liberty" (Franke, 2024).

I previously argued that the vulnerability view is a flawed view of the manipulation risks in algorithmic transparency (Klenk, 2023). Therefore, I will focus on Franke's discussion of the indifference view of manipulative transparency, and its relation to the paradox of positive liberty. About the indifference view (Klenk, 2023), Franke recognises that it is *not* "directly" aligned with concerns about positive liberty. But he maintains that there is indirect link (2024, p. 4):

> for what must be done to avoid being manipulative is to aim to "enhance the decision making capabilities of the users of the algorithm by revealing reasons to them" (Klenk, 2023, p. 14), which certainly amounts to increasing the positive liberty of users.

The indifference view prescribes that an 'influencer' aims to reveal reasons when they influence others. Franke interprets this prescription as being – ultimately – grounded in a concern with positive liberty and the higher self. This, argues Franke (2024, p. 4).

> *does* suggest caution when addressing transparency as manipulation by promoting positive liberty, for if Berlin is right, the risk of erring when promoting positive liberty is greater than that of erring when promoting negative liberty [and, in effect,] the goal of the critical account—to promote positive liberty by cultivating a more 'real', or 'ideal', or 'autonomous' self—while a worthy ideal, also requires interpretation in a way that may be hard to square with due process and rule of law.

Franke does not explain how, concretely, the caution he sees warranted should be incorporated in evaluating, regulating, and deciding about manipulative

transparency. However, his earlier comment on manipulative transparency suggests that he favours an open attitude about what – if anything – must be done about manipulative transparency (cf. Franke, 2022).

3 3. Evaluating Franke's Thesis

I commend Franke for laying a thought-provoking link between the debate about Berlin's two concepts of liberty and the manipulative transparency debate.

I am also sympathetic to Franke's general worry, which, I take it, concerns the dangers of moralising. When we "make moral judgments that don't take into account the complexity of the circumstances," we are swerving from moral analysis into the terrain of sophistry and moralising, as Samuel Scheffler put it (Politika, 2019). This is precisely why we need analyses of the boundaries of manipulation and its ethical repercussions that go beyond platitudes like 'manipulation undermines autonomy' and 'manipulation exploits vulnerabilities.' Otherwise, applying a term like 'manipulation' to e.g. algorithmic transparency is but a complicated way of expressing dissatisfaction and ethical concern about some type of influence. Such an approach would ill-serve the debate about manipulative transparency, and Franke's comment serves as a helpful reminder.

Moreover, by inviting further exploration of the links between views of manipulation and (the perils of) positive liberty, Franke suggests a fruitful avenue for further exploration. In that explorative spirit, however, I will now raise problems for his specific claim that indifference view of manipulation is a cause of caution for being (indirectly) aligned with positive liberty.

In short, I argue that Franke has not provided a plausible and convincing case for the (indirect) link between the indifference view of manipulation and the risks associated with the pursuit of positive liberty. In making my case, I aim to continue the debate Franke has so helpfully put in motion.

First, on Franke's own terms, the indifference view seems more aligned with negative liberty than positive liberty. Franke notes algorithmic transparency requires "sufficient quantity and quality of information" and that these "appear to be sufficient conditions for you to be free in the negative sense when acting on it" (Franke, 2024, p. 4).[1] Franke overlooks that the indifference view aligns very well with this concern about negative liberty. It requires 'influencers' like the deployers of algorithms to choose their means of influence (e.g. an explanatory text or video about the algorithm) depending on whether or not it reveals reasons to the target audience. Contrary to Franke's suggestion, the indifference view does not call for influencers like the deployers of algorithms to help their audience realise their higher selves.

Of course, Franke is right that there is room for interpretation about what it takes to "enhance the decision making capabilities" of an algorithm's user through

[1] The concern with providing *information of sufficient quality* is an apt concern in line with negative liberty because, to recall Carter's (2022) helpful illustration, you need to know where the doors are in order for them to be open to you in the sense relevant for negative liberty.

algorithmic transparency.[2] Nonetheless, it is sensible to count 'revealing reasons' as directly linked to the quality of information provided, thus aligning with Franke's understanding of negative liberty.

For example, an explanation of a credit rating algorithm that says, 'the output of the algorithm determines your credit rating' provides accurate information. However, the quality of information seems to depend, at least in part, on the targeted user. When the user has no idea what 'credit rating' means, or how it affects their life, the provided information is of low quality. It does not reveal reasons for the user to act, feel, or believe one way or the other. The indifference view's prescription to aim for an influence that reveals reasons to the target can thus be read as a prescription to aim to provide quality information. In effect, it serves the aim of negative liberty.

Therefore, the indifference view seems well aligned with concerns about negative liberty, which Franke considers unproblematic, and not with positive liberty, against which Franke urges caution. At the very least, there is room for debate as to whether the indifference view is even indirectly aligned with positive liberty, which would raise questions about the risk that Franke raises.

Second, even if we accept Franke's first claim (that the indifference account of manipulative transparency is (indirectly) aligned with positive liberty), his second claim about the risk of authoritarianism does not follow.

To begin with, Franke's (2024, p. 3) mistakenly suggests that the indifference account of manipulation aims to "cultivate a more 'real', or 'ideal', or 'autonomous' self." The indifference view is committed to a process ideal that presupposes that there are reasons for and against adopting a belief, desire, or emotion, and that social influence should be guided by the aim to reveal those reasons to others (Klenk, 2020, 2022a, 2024). It does not necessarily prescribe the contents of beliefs, desires, or emotions that someone's higher self would or should adopt. Hence, the indifference view is more aligned with a 'content neutral' version of positive liberty, which focuses on the *process* by which people form their desires, beliefs, or preferences rather than the content of these mental states (e.g. Christman, 1991). Insofar as content-neutral accounts of positive liberty escape the paradox of positive liberty (cf. Carter, 2022), then so does the indifference view.

Moreover, identifying manipulative transparency with the indifference view of manipulation does not prescribe a fixed moral response. Instead, the indifference view is, first and foremost, an account of what manipulation *is*, and it leaves open that manipulation may be, all things considered, morally permissible. Hence, we may attest that some means to achieve algorithmic transparency are manipulative (since, say, they are aimed at effectiveness but indifferent to revealing reasons to the audience) and still judge that they are permissible. Perhaps they are done for the wrong reasons, but they still do a good job informing users about the algorithm. This means that opponents of manipulative transparency have to provide an additional ethical argument to defend regulation and the restriction of manipulative algorithmic transparency. This should

[2] Further work should certainly explore this in more depth, and it would be helpful to keep in mind Franke's caution against the perils of positive liberty.

provide some defence against moralistic uses of 'manipulation' that Franke worries about.

Finally, going beyond Franke's comment, I wonder whether the positive and negative liberty distinction points to a more fundamental view of what we need from each other in social influence. When we think of each other as solitary beings that are best left alone, a focus on negative liberty makes sense. But the moment we realise that we depend on each other 'to know which doors are open to us' (to use Carter's metaphor again), we notice that there are farther-reaching demands on us than to provide objectively accurate information, and abstaining from interference. This, however, is a but brief suggestion that merits further discussion and exploration.

In summary, Franke's general worry is apt, but his specific claim that the indifference view is aligned with potentially problematic concerns about positive liberty is problematic. Thus, we should not conclude that the indifference view of manipulative transparency expresses a concern with positive liberty, nor – if it is – that it bears a risk of authoritarianism.

4 Generalising Franke's Thesis

Before concluding, I want to point out a possible generalization of Franke's thesis. If Franke's worry is apt, it seems to apply to the entire field of technology ethics and the ethics of AI, not just the debate about manipulative transparency.

Current approaches in the ethics of artificial intelligence follow an 'ethics by design' approach (Buijsman et al. forthcoming). The aim here is to consider ethical values early on in the design process rather than as a constraint considered only at the end. A core question concerns how to identify the relevant values that should play a role in the design process. The proliferation of 'AI ethics frameworks' that outline the importance of and, sometimes, provide some substantive idea of values like beneficience, autonomy, justice, and non-maleficience, amongst others, can be seen as reflections on which values are important in a design process. Importantly, there is ample room for (rational) disagreement about the relevant values and how to operationalise them in concrete design projects (Klenk, 2022a).

Here's the link to Franke's concern with positive liberty: These values are deemed important (just as the 'higher self' is deemed important). It is difficult to interpret these values (just as it is difficult to figure out the 'higher self'). This opens the door for moralistic uses and interpretations of these terms that are self-serving (just as we can worry about misuses of positive liberty).

As a result, Franke's cautionary tale about positive liberty may – independently of his remarks about manipulative transparency – be read as a cautionary tale about applied ethics, and especially about AI ethics.

5 Conclusion

Algorithms can automate decisions in important areas of life. It seems important that the decisions and their grounds are made transparent. However, making algorithms transparent, users may end up being manipulated by the deployers of algorithms. Franke offers a new spin on the debate, suggesting that concerns about manipulative transparency are linked to problematic concerns with positive liberty. In this contribution, I showed that the indifference view of manipulative transparency is, in fact, not aligned with positive liberty, nor with the risks that are commonly associated with it. Moreover, Franke's worry may generalise to the wider AI ethics debate. Both points are worthy of further exploration, and Franke has advanced the debate by laying a fruitful link between the manipulation- and the positive/negative liberty debate.

Authors' contributions N/A (single author).

Funding The author's work on this paper has been part of the project Ethics of Socially Disruptive Technologies that has received funding from the Dutch Organisation of Scientific Research.

Data Availability All data is available in the MS.

Declarations

Ethics approval and consent to participate N/A.

Consent for publication Consent for publication is given.

Competing interests No competing interests.

References

Berlin, I. (1969). *Liberty: Incorporating four essays on liberty*. Oxford University Press.

Buijsman, S., Klenk, M., & van den Hoven, J. (forthcoming). Ethics of artificial intelligence. In N. Smuha (Ed.), *Cambridge Handbook on the Law, Ethics and Policy of AI*. Cambridge: Cambridge University Press.

Carter, I. (2022). Positive and Negative Liberty. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy*: *Summer 2022* .

Christman, J. (1991). Liberalism and Individual Positive Freedom. *Ethics, 101*(343–359).

Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of medical ethics*. doi:https://doi.org/10.1136/medethics-2020-106820.

Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards Transparency by Design for Artificial Intelligence. *Science and Engineering Ethics, 26*, 3333–3361. https://doi.org/10.1007/s11948-020-00276-4

Franke, U. (2022). How Much Should You Care About Algorithmic Transparency as Manipulation? *Philosophy & Technology, 35*, 1–7. https://doi.org/10.1007/s13347-022-00586-4

Franke, U. (2024). Algorithmic Transparency, Manipulation, and Two Concepts of Liberty. *Philosophy & Technology, 37*, 1–6. https://doi.org/10.1007/s13347-024-00713-3

Klenk, M. (2020). Digital Well-Being and Manipulation Online. In C. Burr & L. Floridi (Eds.), *Ethics of Digital Well-Being: A Multidisciplinary Perspective* (pp. 81–100). Springer.

Klenk, M. (2021). Interpersonal manipulation. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3859178

Klenk, M. (2022a). (Online) manipulation: Sometimes hidden, always careless. *Review of Social Economy, 80*, 85–105. https://doi.org/10.1080/00346764.2021.1894350

Klenk, M. (2022b). Manipulation, injustice, and technology. In M. Klenk & F. Jongepier (Eds.), *The Philosophy of Online Manipulation* (pp. 108–131). Routledge.

Klenk, M. (2023). Algorithmic Transparency and Manipulation. *Philosophy & Technology, 36*, 1–20. https://doi.org/10.1007/s13347-023-00678-9

Klenk, M. (2024). Ethics of generative AI and manipulation: A design-oriented research agenda. *Ethics and Information Technology, 26*, 1–15. https://doi.org/10.1007/s10676-024-09745-x

Klenk, M., & Jongepier, F. (2022). Manipulation Online: Charting the field. In M. Klenk & F. Jongepier (Eds.), *The Philosophy of Online Manipulation* (pp. 15–48). Routledge.

Politika. (2019). Egalitarianism and Consequentialism. https://www.politika.io/en/article/egalitarianism-and-consequentialism. Accessed 19 March 2024.

Wang, H. (2022). Transparency as Manipulation? Uncovering the Disciplinary Power of Algorithmic Transparency. *Philosophy & Technology, 35*, 1–25. https://doi.org/10.1007/s13347-022-00564-w

Wang, H. (2023). Why Should We Care About the Manipulative Power of Algorithmic Transparency? *Philosophy & Technology, 36*, 1–6. https://doi.org/10.1007/s13347-023-00610-1