Philosophy & Technology (2024) 37:22
https://doi.org/10.1007/s13347-024-00713-3

COMMENTARY

Algorithmic Transparency, Manipulation, and Two Concepts of Liberty

Ulrik Franke[1,2]

Received: 10 January 2024 / Accepted: 2 February 2024

Abstract
As more decisions are made by automated algorithmic systems, the transparency of these systems has come under scrutiny. While such transparency is typically seen as beneficial, there is a also a critical, Foucauldian account of it. From this perspective, worries have recently been articulated that algorithmic transparency can be used for manipulation, as part of a disciplinary power structure. Klenk (*Philosophy & Technology* 36, 79, 2023) recently argued that such manipulation should not be understood as exploitation of vulnerable victims, but rather as indifference to whether the information provided enhances decision-making by revealing reasons. This short commentary on Klenk uses Berlin's (1958) two concepts of liberty to further illuminate the concept of transparency as manipulation, finding alignment between positive liberty and the critical account.

Keywords Algorithmic transparency · manipulation · Isaiah Berlin

1 Introduction

Today, many decisions are made by automated algorithmic systems. Some, e.g., recommendations for what to read, watch, or buy are comparatively trivial, while others, e.g., recommendations for who should be hired, granted a loan, or allowed bail are more serious. Steady advances in artificial intelligence suggest that such automated decision-making is still in its infancy (see, e.g. Sarker, 2021; Cobbe, 2019; Araujo et al., 2020, for a few different perspectivess).

As a reaction, there have been many calls for increased transparency of automated systems. Largely, this stems from advances in machine learning, where technologies such as deep neural networks offer impressive results but explaining particular out-

✉ Ulrik Franke
ulrik.franke@ri.se

1 RISE Research Institutes of Sweden, SE-164 29, Kista, Sweden

2 KTH Royal Institute of Technology, SE-100 44, Stockholm, Sweden

Published online: 16 February 2024

comes is very difficult (see, e.g.,Guidotti et al., 2018; Arrieta et al., 2020). However, transparency may be equally important in traditional algorithmic systems, and was called for well before the latest machine learning revolution (see Fleischmann and Wallace, 2005). Such transparency is typically seen at least as a *prima facie* good, though there is a debate about how it can be traded-off against other goods, such as achieving higher accuracy (London, 2019), or avoiding perverse effects of disclosure (de Laat, 2018; Prat, 2005).

However, in addition to this *informational account* of algorithmic transparency, there is a *critical*, Foucauldian account where transparency is part of a disciplinary power structure. From this perspective, Wang identified the possibility of *algorithmic transparency as manipulation*, where an explanation of an algorithm does not only confer neutral information, but also seemingly objective norms which may be imperceptibly internalized, undermining "individuals' cognitive capacity for critical thinking, leading to a situation where people follow the norms only because of ideological conditioning" (Wang, 2022, p. 17).

More recently, Klenk (2023) engaged with Wang's argument, suggesting that it depends on a problematic *vulnerability view* of manipulation (where vulnerabilities are exploited to steer your decisions towards a manipulator's ends), but that it can be salvaged by instead adopting an *indifference view* of manipulation (where a manipulator influences you in a way that aims to be effective, but not in order to reveal reasons to you):

> In short, algorithmic transparency may *not* be designed to enhance the decision making capabilities of the users of the algorithm by revealing reasons to them. If that is the case, then algorithmic transparency will be manipulative. (Klenk, 2023, p. 14)

Compared to the vulnerability view, this indifference view has additional explanatory power to shed light on how algorithmic transparency can amount to manipulation. In particular, it does not require intentions to exploit or harm those being manipulated, and is perfectly compatible with the existence of paternalistic or overall beneficial manipulation (Klenk, 2023, pp. 12–13).

This short commentary on Klenk (2023), uses Berlin's two concepts of liberty (Section 2) to illuminate the concept of transparency as manipulation, finding alignment between positive liberty and the critical account (Section 3). The paper concludes by discussing some implications in Section 4.

2 Berlin's Two Concepts of Liberty

Isaiah Berlin (1958), in his inaugural lecture as Chichele Professor of Social and Political Theory at Oxford University, famously made the distinction between *negative liberty* (freedom from) and *positive liberty* (freedom to).[1]

Negative liberty is freedom from oppression:

[1] Berlin uses liberty and freedom interchangeably (p. 169), and so do we.

> Political liberty in this sense is simply the area within which a man can act unobstructed by others. If I am prevented by others from doing what I could otherwise do, I am to that degree unfree; and if this area is contracted by other men beyond a certain minimum, I can be described as being coerced, or, it may be, enslaved. (Berlin, 1958, p. 169)

Importantly, negative liberty is not about general inability to do what you want, but about particular inability caused by others (p. 170). Furthermore, though negative liberty may need to be limited for the sake of others' equal liberty, those emphasizing its importance (e.g, Locke, Mill, Constant, and Tocqueville) typically defend some "minimum area of personal freedom which must on no account be violated" (p. 171).

Positive liberty, by contrast, is freedom to act autonomously:

> I wish my life and decisions to depend on myself, not on external forces of whatever kind. I wish to be the instrument of my own, not of other men's, acts of will. I wish to be a subject, not an object; to be moved by reasons, by conscious purposes, which are my own, not by causes which affect me, as it were, from outside. […] I wish, above all, to be conscious of myself as a thinking, willing, active being, bearing responsibility for my choices and able to explain them by reference to my own ideas and purposes. I feel free to the degree that I believe this to be true, and enslaved to the degree that I am made to realise that it is not. (Berlin, 1958, p. 178)

Importantly, positive liberty is not about being free to follow the spur of the moment. Positive liberty is to be free from "irrational impulse, uncontrolled desires" in favor of being true to some "'real', or 'ideal', or 'autonomous' self" (p. 179).

Berlin acknowledges that both of these liberties are worthy ideals to strive for and that they need not—logically—be in conflict. Nevertheless, they have "historically developed in divergent directions, not always by logically reputable steps, until, in the end, they came into direct conflict with each other" (Berlin, 1958, p. 179). More precisely, Berlin warns that promoting positive liberty sometimes becomes tyrannical, because identifying freedom with obedience to a higher self may lead to identifying it with obedience to those who can interpret this higher self:

> Once I take this view, I am in a position to ignore the actual wishes of men or societies, to bully, oppress, torture them in the name, and on behalf, of their 'real' selves, in the secure knowledge that whatever is the true goal of man (happiness, performance of duty, wisdom, a just society, self-fulfilment) must be identical with his freedom – the free choice of his 'true', albeit often submerged and inarticulate, self. (Berlin, 1958, p. 180)

3 Liberty, Transparency and Manipulation

It is illuminating to analyze algorithmic transparency as manipulation using Berlin's two concepts.

Under the informational account of transparency, the prime concern is the quantity and quality of the disclosed information (Wang, 2022, p. 4). If information about an

algorithm is insufficient in these respects (e.g., erroneous, misleading, or biased) it curtails your negative liberty, but if it is not, it does not. Sufficient quantity and quality of information appear to be sufficient conditions for you to be free in the negative sense when acting on it.

But under the critical account of transparency, even information which is not erroneous, misleading, or biased, can be manipulative. On Wang's account, such information may lead to norm-objectification: you may uncritically act towards a manipulator's ends rather than your own, unable to question this. You do not act in your "true interests" (Wang, 2022, pp. 18, 19, 20) and so are not free in the positive sense. Thus, the vulnerability view of transparency as manipulation is well aligned with Berlin's notion of positive liberty.[2]

On Klenk's critical account—the indifference view of transparency as manipulation—information is manipulative when not designed to reveal reasons to the users of the algorithm. Thus, the indifference view of manipulation focuses not on how you are affected, but on the aims of the agent providing the information. There is no direct appeal to any "true interests". Thus, the indifference account of transparency as manipulation is not directly aligned with Berlin's notion of positive liberty. Indirectly, however, there is alignment, 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 (compare Berlin's "positive doctrine of liberation by reason", p. 191).

To summarize, we have observed an alignment between critical accounts of transparency as manipulation on the one hand and Berlin's notion of positive liberty on the other. The critical perspective embraces and promotes positive liberty whether directly (under the vulnerability view) or indirectly (under the indifference view).

4 Discussion and Concluding Remarks

Berlin's warning against the dangers of positive liberty is sometimes interpreted as a rejection of its value. But as a pluralist, Berlin embraces both negative and positive liberty.[3] Similarly, the alignment discovered above between critical accounts of transparency as manipulation and positive liberty is not a rejection of these accounts.

However, the alignment *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. (Perhaps this caution should be even greater under the vulnerability view,

[2] It is instructive to compare Wang's claim that "[m]anipulation is morally objectional because it exploits individuals' vulnerabilities, and […] treats the manipulee as mere means, rather than an end in themselves" (p. 18) with Berlin's discussion first of why lies and manipulation deny human nature as autonomous beings in a Kantian sense (pp. 183–184) and then of how this idea can be transformed into an authoritarian one (pp. 198–199), through "steps which, if not logically valid, are historically and psychologically intelligible" (p.198).

[3] To clarify his position, he later wrote that "I am not offering a blank endorsement of the 'negative' concept as opposed to its 'positive' twin brother, since this would itself constitute precisely the kind of intolerant monism against which the entire argument is directed" (Berlin, 2002, p. 50, footnote 1).

compared to the indifference view, since the former is more directly connected to positive liberty.)

As an illustration, consider Klenk's example from political advertising (p. 11): if stereotypes of 'foreign-looking' people are used to ignite xenophobia rather than (implausibly) to reveal reasons for political deliberation, this is manipulation. To avoid this, we may regulate political advertising—but this could easily degenerate into oppressive censorship. Now, following Berlin, this risk seems greater if the law follows the critical account, e.g., forbids advertising that does not aim to reveal reasons or is prone to be norm-objectifying, compared to the case where the law follows the informational account, e.g., mandates disclosure of who paid for an ad or why it was shown to you.

One reason for this greater risk is that 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. A related worry is that the pursuit of high ideals may hinder actual, piecemeal, progress; that critical accounts of transparency as manipulation risk ending up asking too much, e.g., that the users of algorithms forswear their existing wants in favor of nobler ones, that providers of algorithms only act on motives so pure that they are nowhere to be found, or that the entire socio-economic system must be recast before non-manipulative information on credit-scoring can be offered.[4]

If Berlin is right, the risk of erring when promoting positive liberty under the critical account of transparency as manipulation is greater than that of erring when promoting negative liberty under the informational account. It is prudent to consider his warning when addressing issues of algorithmic transparency. But even if Berlin is *not* right, the fact that different people may judge these risks differently is yet another explanation of the observation made by Franke (2022) about different levels of constructionist commitment: even if intellectually convinced by a critical account of transparency as manipulation in some particular case, different people may end up having different ideas about what, if anything, should be done about it. More precisely, a person sharing Berlin's risk assessment will end up with less constructionist commitment.

Author Contributions Not applicable (single author).

Funding Open access funding provided by RISE Research Institutes of Sweden. The author received no external funding for this work.

Availability of data and material Not applicable.

Declarations

Competing interests The author declares no conflict of interest.

Ethics approval and consent to participate Not applicable.

[4] In the spirit of Berlin's value pluralism, it should of course also be acknowledged that there is a corresponding risk that the informational account of transparency ends up asking too little, i.e., that it allows too much manipulation and even exploitation by the powerful of the weak.

Consent for publication Yes.

References

Araujo, T., Helberger, N., Kruikemeier, S., & De Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & SOCIETY, 35*, 611–623. https://doi.org/10.1007/s00146-019-00931-w

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Berlin, I. (1958). Two Concepts of Liberty. Reprinted in Hardy, H (ed) *Liberty*, Oxford University Press, 2002, pp. 166–217. https://doi.org/10.1093/019924989X.003.0004, to which the pagination used here refers.

Berlin, I. (2002). Introduction. In: Hardy H (ed) *Liberty*, Oxford University Press, pp 3–54. https://doi.org/10.1093/019924989X.003.0001

Cobbe, J. (2019). Administrative law and the machines of government: judicial review of automated public-sector decision-making. *Legal Studies, 39*(4), 636–655. https://doi.org/10.1017/lst.2019.9

Fleischmann, K. R., & Wallace, W. A. (2005). A covenant with transparency: Opening the black box of models. *Communications of the ACM, 48*(5), 93–97. https://doi.org/10.1145/1060710.1060715

Franke, U. (2022). How Much Should You Care About Algorithmic Transparency as Manipulation? *Philosophy & Technology, 35*(92). https://doi.org/10.1007/s13347-022-00586-4

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR), 51*(5), 1–42. https://doi.org/10.1145/3236009

Klenk, M. (2023). Algorithmic Transparency and Manipulation. *Philosophy & Technology, 36*(79). https://doi.org/10.1007/s13347-023-00678-9

de Laat, P. B. (2018). Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability? *Philosophy & Technology, 31*(4), 525–541. https://doi.org/10.1007/s13347-017-0293-z

London, A. J. (2019). Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hastings Center Report, 49*(1), 15–21. https://doi.org/10.1002/hast.973

Prat, A. (2005). The wrong kind of transparency. *American economic review, 95*(3),862–877. https://www.jstor.org/stable/4132745

Sarker, I.H.(2021). Machine learning: Algorithms, real-world applications and research directions. *SN computer science, 2*(3),160. https://doi.org/10.1007/s42979-021-00592-x

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