Philosophy & Technology (2022) 35:92
https://doi.org/10.1007/s13347-022-00586-4

COMMENTARY

How Much Should You Care About Algorithmic Transparency as Manipulation?

Ulrik Franke[1,2]

Received: 23 September 2022 / Accepted: 27 September 2022 / Published online: 14 October 2022

Abstract
Wang (*Philosophy & Technology* 35, 2022) introduces a Foucauldian power account of algorithmic transparency. This short commentary explores when this power account is appropriate. It is first observed that the power account is a constructionist one, and that such accounts often come with both factual and evaluative claims. In an instance of Hume's law, the evaluative claims do not follow from the factual claims, leaving open the question of how much constructionist commitment (Hacking, 1999) one should have. The concept of acts in equilibrium (Nozick, 1981) is then used to explain how different individuals reading Wang can end up with different evaluative attitudes towards algorithmic transparency, despite factual agreement. The commentary concludes by situating constructionist commitment inside a larger question of how much to think of our actions, identifying conflicting arguments.

Keywords Algorithmic transparency · Constructionism · Hume's law · Acts in equilibrium

1 Introduction

In modern society, a steadily increasing number of tasks and decisions are automated. Advances in artificial intelligence (AI) and especially its machine learning (ML) subset are constantly expanding the realm of tasks which can be carried out by machines (see, e.g., Hirschberg and Manning, 2015 on natural language processing or Zhao et al., 2019 on object recognition). Even though the learning of these systems is relatively well understood—this is part of what underpins the impressive technical advances—it is often exceedingly difficult to explain particular outcomes in concrete cases, giving ML a 'black box' reputation.

✉ 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

As a result, transparent and explainable AI has become a prolific research area in recent years (see, e.g., Guidotti et al., 2018; Arrieta et al., 2020). While most believe that such transparency or explainability is at least a *prima facie* good, there is a debate about exactly how much transparency should be required and in which circumstances (for some different arguments and positions, see Fleischmann and Wallace, 2005; Holm, 2019; Zerilli et al., 2019).

However, Wang (2022) offers a different perspective—a Foucauldian analysis of algorithmic transparency as part of a disciplinary power structure. More precisely, Wang explains that algorithmic transparency can be understood from two complementary perspectives: an *informational account* where more transparency merely gives more information about how an algorithm works, and a *power account* where making an algorithm transparent "is not just about revealing objective information about how it works, but also about the interests of those who created it and their views about those who are to be subject to it" (Wang, 2022, p. 5). Thus, on the power account, explanations about the inner workings of algorithms such as the FICO credit scoring system used by Wang as a running example are not merely conferring new knowledge to the recipients, but constitute a display of seemingly objective norms which may be internalized by the recipients, 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).

The purpose of this short commentary is to further explore the question of when the power account is appropriate.

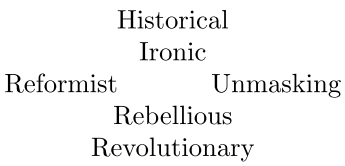
2 Constructionism and Hume's law

Even though Wang (2022) does not explicitly label the Foucauldian power account a *constructionist* account of algorithmic transparency, this seems like a reasonable and illuminating interpretation.[1] Such analyses of the 'construction of *X*' are critical of the status quo. More precisely, following Hacking (1999), p. 6, p. 12), constructionists claim at least that (0) *X* is taken for granted and appears inevitable, but that in fact (1) *X* need not have existed, is not determined by the nature of things and is not at all inevitable. It seems fair to say that Wang (2022) fits this general pattern, with *X* being (our reactions to) algorithmic transparency. Furthermore, still following Hacking , constructionists often also claim that (2) *X* is quite bad as it is and (3) that we would be much better off if *X* were done away with.[2] Wang at times hints at (2) and (3), but detailed exegesis is not our purpose here.

[1] Miller (2008), p. 251) opens a chapter on Foucauldian constructionism with the following remark: "Anyone who knows anything about Michel Foucault knows that he was a constructionist". Among Wang's key references in the running example about credit scoring, Burton (2007) studies "the construction of the trustworthy consumer", Burton (2012) investigates the "constructed nature of risk and credit scoring" and DuFault and Schouten (2020) address "the construction of the datapreneurial credit consumer identity".

[2] These are somewhat abbreviated versions of Hacking's claims. The full wordings of (1)–(3) are found in Hacking (1999), p. 6) and the full wording of (0) is found in Hacking (1999), p. 12).

Instead, we make a general observation: In an instance of Hume's law—that an 'ought' cannot be derived from an 'is'—it appears that evaluative conclusions such as (2) or (3) do not follow from the factual premises (0) and (1): You may be convinced about (0) and (1), yet not subscribe to (2) or (3). Thus, whenever we read and are *factually* convinced by constructionist research on some *X*, the additional question remains which *evaluative* position to adopt towards *X*. Indeed, Hacking (1999, p. 19) offers a taxonomy, and a partial order, of six different grades of constructionist commitment:



At the one end of the spectrum, the historical constructionist "can be quite noncommittal about whether *X* is good or bad" (Hacking, 1999, p. 19), whereas at the other end of the spectrum, the "activist who moves beyond the world of ideas to and tries to change the world in respect of *X* is *revolutionary*" (Hacking, 1999, p. 20, emphasis in original). This question of commitment and evaluative conclusions is the question we address in this commentary: Which attitude should we adopt towards the power account of algorithmic transparency offered by Wang? How much should you care about algorithmic transparency as manipulation?

3 Acts in Equilibrium

Without claiming to have devised a full solution to the question posed in the previous section, we propose that one important clue can be found in the notion of acts being in equilibrium. Following Nozick (1981, p. 349), an *act in equilibrium for a person* is defined as follows:

(a) he does (or wants to do) it, and
(b) if he knew the causes of his doing or wanting to do the act then he would still (want to) do it as much[3]

The equilibrium notion goes some way towards explaining why Hacking's (2) or (3) might seem like appropriate reactions to learning that some *X* is constructed:

> Clearly, it is desirable that the acts we do be in equilibrium, that they (and we) are able to stand and withstand knowledge of their causes. Would it not

[3] Nozick first gives the label (b) to a tentative version, but then goes on to offer a stricter version; this is our (b).

> be very distressing to learn or believe that if you knew why you were pursuing some major course of action, what was causing you to do so, you wouldn't then choose to do it? Even if we do not know an act's specific causes, can doing it withstand the (general) knowledge that that act is in disequilibrium? (Nozick, 1981, p. 349)

To make a concrete credit scoring example, suppose that you intend to buy some particular thing using your credit card. Now, suppose that you learn that the cause[4] of your wanting to do so is the belief that (i) your credit score will not be adversely affected, because (ii) it has been disclosed to you that the FICO scoring algorithm does not penalize some additional debt as long as you make your payments on time, because (iii) having customers using their credit and paying their bills—including a considerable interest—is how creditors earn their profit. Perhaps your intention to use your credit card will then turn out to be in disequilibrium: You will not want to pay with credit card as much and may indeed (want to) pay with your debit card or with cash instead, to avoid accumulating debt, especially if you also note that mathematical fact (iv) that your total cost will be lower if you pay upfront and avoid the interest.[5] If, upon closer inspection, you find that algorithmic transparency regularly tends to induce such acts in disequilibrium for you, it seems reasonable that you will gradually become more and more convinced about Hacking's (2) and (3)—that transparency about the FICO scoring algorithm is bad and should be done away with.

However, the equilibrium account is also illuminating in that it can explain why you may *not* become convinced about (2) and (3): "That there are bad motives for a position does not show there *couldn't* be good motives for it" (Nozick, 1981, p. 349, emphasis in original). This may lend additional force to the closing remarks in Wang's Section 5 (p. 17), where some limits to arbitrariness, discrimination, and unfairness are acknowledged. To make another example, suppose that you intend to pay your credit card bill on time. Now, suppose that you learn that the cause of your wanting to do so is the belief that (i) you should pay your bills on time, because (ii) it has been disclosed to you that the FICO scoring algorithm penalizes not paying on time, because (iii) if debtors stop paying their debt, the creditors who design the FICO scoring algorithm will go bankrupt. Perhaps your intention to pay on time will then turn out to be in equilibrium: You will still want to pay your credit card bill on time, especially if you also believe that (iv) responsibly used credit offers advantages not to be had in an economy without lending, and (v) credit without incentives to pay on time is unsustainable. If, upon closer inspection, you find that algorithmic transparency regularly tends to induce acts in equilibrium for you, it seems reasonable that you will reject Hacking's (2) and (3), even if you accept the factual

[4] The notion of acts in equilibrium needs to be refined with respect to partial knowledge of causes. Nozick discusses this in footnote 59 (pp. 714–715), concluding that we should "add the plausible condition that eventually there is full enough knowledge so that its effect is not changed by even wider knowledge, that is, that there isn't an infinite cycle of shifting of adherence to and from the act under wider and wider knowledge."

[5] Assuming, plausibly, a positive real interest rate.

constructionist claims that the FICO scoring system is not objective and inevitable, but is indeed influenced by interests and asymmetrical power relations.

Recall that the definition of an act in equilibrium pertains to a particular person. Thus, depending on factual and evaluative convictions, different acts will be in equilibrium or disequilibrium for different persons, who will thus differ in their constructionist commitment.

4 Discussion and Concluding Remarks

At least part of the appeal of the Foucauldian account of our actions in terms of conditioning and discipline is that it offers a novel and possibly important perspective on our lives, in the spirit of the Socratic dictum about the unexamined life not being worth living. In a parallel to the *subjunctive* notion of equilibrium, Nozick (1981, p. 351, emphasis in original) says of the *occurrent* situation "when the person does have the causal knowledge and the belief or action does stay unchanged, that the belief or action *socratizes*." Clearly, such reflection about our reasons for beliefs and actions is valuable.

On the other hand, our cognitive faculties are not sufficient to reflect about our reasons for beliefs and actions in every case. In an oft cited dictum, Whitehead (1911) reflects:

> It is a profoundly erroneous truism, repeated by all copy-books and by eminent people when they are making speeches, that we should cultivate the habit of thinking of what we are doing. The precise opposite is the case. Civilization advances by extending the number of important operations which we can perform without thinking about them. Operations of thought are like cavalry charges in a battle—they are strictly limited in number, they require fresh horses, and must only be made at decisive moments. (Whitehead, 1911, pp. 45–46)

Equally clearly, Whitehead makes an important point, perhaps especially valid in the original mathematical context: reflecting carefully about the Peano axioms every time we do arithmetic with natural numbers is a waste of time.

Wang's Foucauldian and constructivist account of algorithmic transparency as part of a disciplinary power structure highlights how the perspectives of Socrates and Whitehead, each *prima facie* appealing, are in conflict. We do not claim to have solved this problem in this short commentary, but we do claim to have found two relevant pieces of the puzzle. Hume's law and Nozick's notion of acts in equilibrium at least partially explain why different individuals can have different grades of constructionist commitment (Hacking, 1999), e.g., why one person reading Wang may end up believing that the FICO scoring algorithm and its documentation should be radically changed, while another person reading Wang may end up believing that it requires only minor tweaks.

In Section 2 we asked how much you should care about algorithmic transparency as manipulation. Our partial answer is that it is individually rational to care to the extent that algorithmic transparency induces you to do acts which are in disequilibrium. But

this answer is a limited one. For suppose that "[a]s we gain more knowledge, including scientific knowledge within psychology, sociobiology, and sociology, perhaps we will want to modify our current wants and character […] Perhaps we are only near the beginning of development with more to come, including moral development" (Nozick, 1981, p. 351). Such a process, Nozick reminds us, could go on for a long time, and we do not know whether it will converge or not.

Funding Open access funding provided by RISE Research Institutes of Sweden.

Data Availability Not applicable

Declarations

Ethics Approval Not applicable

Consent to Participate Not applicable

Consent for Publication Yes

Competing Interests The author declares no competing interests.

References

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion, 58,* 82–115.

Burton, D. (2007). *Credit and consumer society*. Routledge.

Burton, D. (2012). Credit scoring, risk, and consumer lendingscapes in emerging markets. *Environment and Planning A, 44*, 111–124.

DuFault, B. L., & Schouten, J. W. (2020). Self-quantification and the datapreneurial consumer identity. *Consumption Markets & Culture, 23,* 290–316.

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

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,* 1–42.

Hacking, I. (1999). *The social construction of what?* Harvard University Press.

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science, 349,* 261–266.

Holm, E. A. (2019). In defense of the black box. *Science, 364*, 26–27.

Miller, L. (2008). Foucauldian constructionism. In J. A. Holstein & J. F. Gubrium (Eds.), *Handbook of constructionist research* (pp. 251–274). New York/London: The Guilford Press.

Nozick, R. (1981). *Philosophical explanations*. The Belknap Press of Harvard University Press.

Wang, H. (2022). Transparency as manipulation? Uncovering the disciplinary power of algorithmic transparency. *Philosophy & Technology*, *35*(3), 69.

Whitehead, A. N. (1911). *An introduction to mathematics*. E-book by Project Gutenberg (originally Williams & Norgate, London).

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy & Technology, 32,* 661–683.

Zhao, Z.-Q., Zheng, P., Xu, S.-T., & Wu, X. (2019). Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems, 30,* 3212–3232.