Freedom in the Digital Age

Designing for Non-Domination

Maas, J.J.C.

DOI
10.4233/uuid:b9c83022-f54d-42a4-abda-afcd776279de

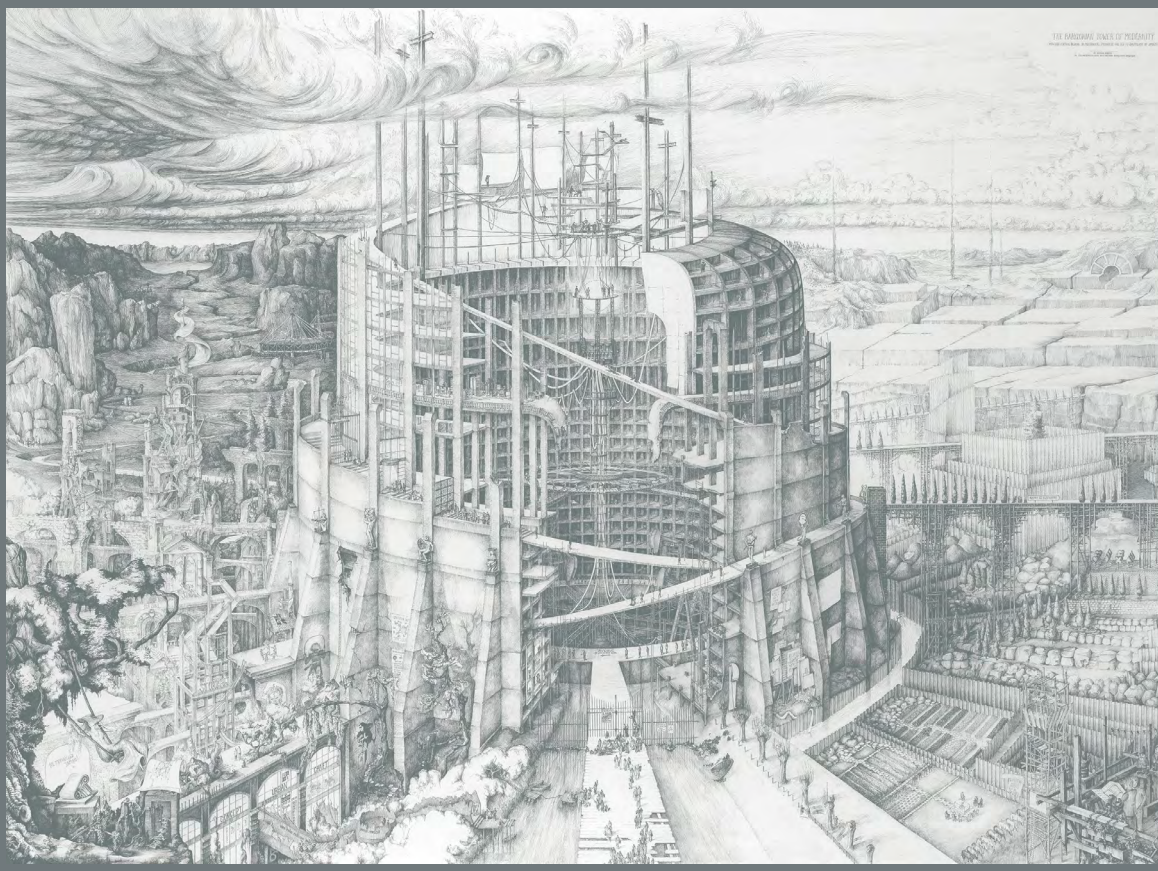
Publication date
2025

Document Version
Final published version

Citation (APA)
Maas, J. J. C. (2025). *Freedom in the Digital Age: Designing for Non-Domination*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:b9c83022-f54d-42a4-abda-afcd776279de

Freedom in the Digital Age: Designing for Non-Domination

Jonne Maas



Simon Stevin Series in the Ethics of Technology

Freedom in the Digital Age: Designing for Non-Domination

Freedom in the Digital Age: Designing for Non-Domination

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus Prof.dr.ir. T.H.J.J. van der Hagen
chair of the Board for Doctorates
to be defended publicly on
Wednesday 8 January 2025 at 15:00 o'clock

by

Jonne Julia Clara MAAS
Master of Science in Philosophy of Science, Technology and Society,
University of Twente, the Netherlands
born in Amsterdam, the Netherlands

This dissertation has been approved by the promotors.
Composition of the doctoral committee:

Rector Magnificus	chairperson
Prof.dr. M.J. van den Hoven	Delft University of Technology, promotor
Prof.dr. C.M. Jonker	Delft University of Technology, promotor
Dr. J.M. Durán	Delft University of Technology, copromotor

Independent members:

Prof.dr.ir. I.R. van de Poel	Delft University of Technology
Prof.dr. M.V. Dignum	Umeå University
Prof.dr. F. Santoni de Sio	Eindhoven University of Technology
Dr. U. Aytaç	Utrecht University
Prof.dr. S. Roeser	Delft University of Technology, reserve member

Keywords: freedom, domination, AI systems, Responsible AI, AI design, power relations, republicanism
Printed by: Grefo Prepress
Cover by: Carlijn Kingma, https://carlijnkingma.com/The-Babylonian-Tower-of-Modernity

The research reported in this dissertation was partially supported by the EU H2020 ICT48 project „Humane AI Net“ under contract # 952026. The support is gratefully acknowledged.

The Simon Stevin Series in the Ethics of Technology is an initiative of the 4TU Centre for Ethics and Technology. Contact: info@ethicsandtechnology.eu

ISBN: 978-94-6366-987-0
ISSN: 1574-941X

Copies of this publication may be ordered from the 4TU.Centre for Ethics and Technology, info@ethicsandtechnology.eu or at the TU Delft repository, https://repository.tudelft.nl/
For more information, see http://www.ethicsandtechnology.eu

Table of Contents

Acknowledgements

I am incredibly grateful for everyone who has shaped this dissertation in various ways. First and foremost I'd like to thank my promotor Jeroen van den Hoven, without whom this dissertation would not have come together in the way it has now. Thank you so much for the opportunity to work closely with you these past years. I'd also like to thank my other promotor, Catholijn Jonker. Your enthusiasm for the project was truly a strong motivational force, and your suggestions during the monthly meetings shaped much of my thinking. Thank you, and thank you also for introducing me to the Hybrid Intelligence community. Finally, I'd like to thank my co-promotor, Juan Durán. I have learned a lot from your analytic approach, and it has urged me to be more precise in my own thinking. I am incredibly grateful for all your advice and feedback.

I'd like to thank Ibo van de Poel, Virginia Dignum, Filippo Santoni de Sio, Uğur Aytaç, and Sabine Roeser for being on my examination committee. I very much appreciate the time you have taken to engage with my dissertation. A double thanks to Virginia Dignum for all the work you've done for Humane AI, and specifically pulling WP5. It has been a pleasure working together with you on various projects these past years.

I also am incredibly grateful for the professional community in which I have been surrounded the past years. The Humane AI network showed me what it is like to operate with a broad European Consortium, allowing me to visit excellent places during my PhD. Thanks also for funding many of my conferences and research visits. I have also very much enjoyed spending time with the Hybrid Intelligence consortium members as an affiliate researcher. Thank you specifically to Rineke Verbrugge for organizing the PhD meetings, and for making me feel included even though my funding came from elsewhere. I have really appreciated it.

Besides these big consortia, the Delft TBM department has been a wonderful homebase for the PhD. Thanks to everyone, and in particular Nathalie van den Heuvel, Brenda van der Sar, Lissy Mayer, Monica Natanael, and Sophia Wessels for your invaluable organisational help these past years. A big thanks to Olya also for being such a supportive colleague.

I have immensely benefitted from the Delft Digital Ethics Centre. Thanks to all involved for your company and feedback on my work. A special thanks to my paranymph Giorgia Pozzi. I'm incredibly grateful that I had the opportunity to share

the PhD experience with you, I could not have done it without you. Thank you for being my rock in academic life.

I've also benefitted immensely from my two research visits at ANU. Thanks in particular to Seth Lazar for sponsoring me, to Philip Pettit for insightful conversations, and the grad community for your warm welcome.

There is some work without whom, literally, the papers wouldn't exist. Thanks to all my co-authors for their collaboration: Herman Veluwenkamp, Marianna Capasso, Lavinia Marin, Jeroen van den Hoven, and Aarón Moreno Inglés.

Besides this rich professional community, I also praise myself lucky with being surrounded by my lovely friends and family. Thanks to the Kaaskartel: Demi, Floor, Hannah, Karlijn, Patricia, Saskia, Bram, Dennis, Jochem, Joris, Jurre. A special shout out to Bram for keeping my life somewhat organized at the Krib; Floor, for all your support these past years; Patricia, for sharing Manaslu—my physical and mental peak still—and for all our other adventures. Thanks to the flying bitches: Cece, Mila, Julia. Hiking mountains with you really does make me feel like a flying bitch. Julia, if anyone has seen my ups and downs these past years, it's you. Thank you for everything. Charlotte, thanks for your wonderful friendship.

My close family has provided me with an invaluable base during the PhD: my helpful brother and paranymph Timo, my thoughtful sister-in-law Kaat, my wonderful niece and nephew Lena and Melle, and my always supporting parents Harro and Geerte. Thank you all for providing me with the most stable foundation for writing a PhD during Covid (less stable but more lively with the babes), for all your support, and for not only being my family, but for making us *be* a family.

Finally, Darryl, my dear goofball. You challenge me in all the right ways. I have no words to express my gratitude for your presence in my life. Thank you for your kindness, joyfulness, patience, reflectiveness, and all your other wonderful qualities. I love you dearly.

Executive Summary

Artificial Intelligence (AI) raises numerous ethical concerns on society, such as violating privacy, undermining autonomy, and producing biased, discriminatory outputs. Increasingly, the field in AI Ethics has discussed concerns related to the power dynamics underlying the development and deployment of AI systems. However, the literature does not provide a concrete conceptualization of these power dynamics, which hinders a comprehensive normative evaluation. In this thesis, I provide a concrete conceptualization of these dynamics. I show how AI systems deployed in core societal sectors, such as healthcare and public administration, constitute a power relation between the 'shapers' (the developers and deployers) and the 'affected' (the direct and indirect end-users and society more broadly construed) of an AI system. I furthermore argue that these power relations generate and are expressive of the moral wrong of *domination* as understood in republican political theory.

An instance of domination occurs when someone is subjected to an arbitrary and uncontrolled power (e.g., a slave subjected to a master). For republicans, domination constitutes a moral wrong because it implies that the dominated agent is not in control of their own choices, and is therefore undermined in their ability to flourish as a person. Republicans hold such control as the true source of one's freedom. Unlike the more commonly endorsed liberal conception of freedom as non-interference (where one is free when not interfered with), republicans thus endorse a conception of freedom as non-domination (where one is free as long as any potential interference occurs under one's own control).

In the context of AI systems, I argue that such digital domination arises for three reasons. First, the shapers have superior decision-making powers which affect end-users and society more broadly. Second, technological limitations such as AI's black box character result in accountability gaps. Third, underlying societal structures enhance the power of the shapers and inhibit the power of the stakeholders by undermining their ability to contest and challenge these decisions. These three points combined result in that AI developers and deployers have arbitrary and uncontrolled power over end-users and society more broadly. Consequently, the freedom of the 'affected' is undermined.

Drawing on the field of conceptual engineering, I propose that in order to safeguard freedom in the digital age, we must rethink the conception of freedom from a liberal to a republican one. Specifically, safeguarding freedom requires to actively

design AI systems that adhere to the value of non-domination, or what I call *design for non-domination*. Such designing for non-domination requires a broad societal perspective and cannot be done by focusing on the system itself (e.g., making it more explainable or incorporating stakeholders in a naive and voluntary way). It fundamentally requires a socio-technical design approach that actively considers social, political, legal, and economic structures.

Samenvatting

Kunstmatige intelligentie (AI) roept tal van ethische zorgen op voor de maatschappij, zoals het schenden van de privacy, het ondermijnen van autonomie en het produceren van bevooroordeelde en discriminerende uitkomsten. In recente jaren besteed het vakgebied AI-ethiek steeds meer aandacht aan de machtsdynamiek die ten grondslag ligt aan de ontwikkeling en inzetting van AI-systemen. De literatuur biedt echter geen concrete conceptualisering van deze machtsdynamiek, wat een normatieve evaluatie belemmert. In dit proefschrift geef ik een concrete conceptualisering van deze dynamiek door te laten zien hoe AI-systemen die worden ingezet in kernsectoren van de maatschappij, zoals gezondheidszorg, justitie en openbaar bestuur, een machtsrelatie vormen tussen de 'vormgevers' (de ontwikkelaars en de inzetters) en de 'getroffenen' (de directe en indirecte eindgebruikers en de maatschappij in bredere zin) van een AI-systeem. Ik beargumenteer dat deze machtsrelaties het morele onrecht van *dominantie* bewerkstelligen zoals gedefinieerd volgens de republikeinse politieke stroming.

Een geval van dominantie doet zich voor wanneer iemand wordt onderworpen aan een willekeurige en ongecontroleerde macht (bijvoorbeeld een tot slaaf gemaakte die wordt onderworpen aan een meester). Voor republikeinen is dominantie een moreel onrecht, omdat het impliceert dat de gedomineerde agent geen controle heeft over zijn eigen keuzes en daarom wordt ondermijnd in zijn vermogen om als persoon te floreren. Republikeinen beschouwen dergelijke controle als de ware bron van iemands vrijheid. In tegenstelling tot de meer algemeen onderschreven liberale opvatting van vrijheid als non-interferentie (waarbij men vrij is wanneer er niet wordt geïnterfereerd), onderschrijven republikeinen een opvatting van vrijheid als non-dominantie (waarbij men vrij is zolang potentiële interferentie plaatsvindt onder iemands eigen controle).

In de context van AI-systemen betoog ik dat dergelijke digitale dominantie om drie redenen ontstaat. Ten eerste hebben de vormgevers superieure beslissingsbevoegdheden die eindgebruikers en de samenleving beïnvloeden. Ten tweede resulteren technologische beperkingen zoals het black box-karakter van AI in hiaten in de verantwoordingsplicht. Ten derde versterken onderliggende maatschappelijke structuren de macht van de vormgevers en remmen ze de macht van de belanghebbenden doordat ze hun vermogen ondermijnen om deze beslissingen te betwisten en aan te vechten. Deze drie punten samen resulteren erin

dat AI-ontwikkelaars en inzetters willekeurige en ongecontroleerde macht hebben over eindgebruikers en de samenleving in bredere zin. Dit ondermijnt de vrijheid van de 'getroffenen'.

Geïnspireerd door het vakgebied van conceptuele engineering stel ik voor dat we het concept van vrijheid moeten heroverwegen van een liberale naar een republikeinse conceptie om de vrijheid in het digitale tijdperk te waarborgen. Dit vereist het actief ontwerpen van AI-systemen die vasthouden aan de waarde van non-dominantie, of wat ik *design for non-domination* noem ('ontwerpen voor niet-dominantie'). Zulk *design for non-domination* vereist een breed maatschappelijk perspectief en kan niet worden bereikt door te concentreren op het systeem zelf (bijvoorbeeld door AI uitlegbaar te maken of belanghebbenden op een naïeve manier bij de ontwikkeling te betrekken). *Design for non-domination* vereist dus een sociaal-technische benadering voor AI ontwerp die actief rekening houdt met sociale, politieke, juridische en economische structuren.

Introduction: Freedom in the Digital Age

In recent years, digital technologies—specifically, online platforms such as Google, Facebook, Uber, and Amazon—have disrupted society in various ways. In 2018, the world was shocked by the Cambridge-Analytica scandal, in which data of millions of Facebook users was collected without their knowledge and used to profile millions of other Facebook users in an attempt to influence the US 2016 presidential elections (Ma 2018). When in 2018 a person was killed by one of Uber's automated vehicles, questions regarding responsibility and accountability became inevitable (Nyholm 2023). In 2020, the Covid-19 pandemic reinforced already existing worries about the duties of online platforms with regard to content moderation, given the spread of mis- and disinformation about vaccines, the disease, and potential cures (Nemitz 2018; Sander 2019; Naeem, Bhatti & Khan 2021; Gisondi et al. 2022). In late 2022, OpenAI released the large language model ChatGPT into society which is now forcing educational institutions to rethink their ways of assessing students and taking exams (Lo 2023; Heaven 2023).

Events like these have increasingly raised concerns about the significant social, economic, and political power[1] online platforms have in society which provides them with the ability to interfere and shape society more broadly. Specifically, concerns relate to the lack of sufficient safeguards and checks and balances to ensure this interference and shaping happens in the 'right' way. For instance, Francis Fukuyama and colleagues (2021) compare the power of Big Tech to a loaded gun on the table, just waiting to be picked up and used by the wrong person. Paul Nemitz (2018) is concerned that Big Tech companies undermine constitutional democracy because they do not seem to be willing to be subjugated to the rule of law, a fundamental aspect

[1] Power is a highly contested and ambiguous concept. In Chapter 2 of this dissertation, I provide a technical definition of power in terms of power-dependence relations between the powerful and powerless. For now, I use power to refer to an agent's position to significantly shape societal life (e.g., influencing presidential elections, deciding what content is allowed to be shared in online spaces, or transforming educational institutions). Consequently, when an agent is powerful they have a relevant degree of power to shape society.

of a well-functioning democracy.[2] Shoshana Zuboff (2019a) has argued that we now live in the age of *Surveillance Capitalism*, where data collection and analyses allow corporations to maximize their own profits at the detriment of individual freedom and autonomy, thereby undermining the democratic order. Recently, Yanis Varoufakis (2023) has gone even further, claiming cloud capital has killed capitalism, replacing it with an economic order he calls *Technofeudalism*. Similar to how vassals and serfs served their overlords in medieval England, individuals and companies are now serving online platforms either by providing companies with their service (e.g., Uber drivers) or with their data (e.g., Facebook users).

For Fukuyama et al., Nemitz, Zuboff, and Varoufakis, and numerous other scholars (e.g., Benn & Lazar 2022; Schaake 2024) the power of online platforms is currently not exercised in the 'right' way. Their concerns, worries, and criticisms raise at least two relevant conditions of what is 'wrong' with the current power. First, there is the aspect of *uncontrolled* power. This is captured by Fukuyama et al.'s fear that the wrong person picks up the gun, and Nemitz' critique that these highly powerful platforms are able to circumvent laws or at least shape laws according to their preferences, wishes, or desires. The power to interfere with and shape society is thus uncontrolled as there are no (effective) checks and balances, and no concrete way to intervene in case the 'wrong' person picks up the gun. In other words, these platforms can exercise their power at their will and at their discretion. Second, there are concerns related to what the power of online platforms allows these companies to do to people (microtargeting their users by data collection and analysis), and what the consequences of such social power are (undermining individual freedom and democracy more broadly). Yet, while these scholars all strongly agree on both conditions, what is missing is a clear connection of how uncontrolled power in the digital realm precisely undermines freedom.

In this dissertation, I explore how these two conditions relate to each other. Specifically, I ask how the freedom of citizens in modern, digital societies is

[2] Although in the time between Nemitz' concern the Digital Markets Act (DMA) and Digital Service Act (DSA) have gone into effect, which may address some of his concerns, it is no secret the amount of lobbying power (i.e., power to shape political and regulatory initiatives) these tech companies have over the development of regulation (Dignam 2020; Erman & Furendal 2024). Although these companies have now come to recognize they cannot do without regulation, the combination of their epistemic authority, influence with regard to regulation, and remaining circumvention of other laws (e.g., taxation) strongly suggests that the influential position of tech companies remains a threat to constitutional democracies. I address the influence of tech companies on regulatory initiatives in Chapter 6 of this dissertation.

undermined by the uncontrolled power of developers and deployers of digital technologies to interfere in their lives. This question has the following three cornerstones. It requires (1) an account of freedom, (2) how freedom relates to uncontrolled power, and (3) clarity on how I conceive of the 'digital realm.' Although the focus so far has been on online platforms, the digital realm extends beyond online platforms and includes various sectors such as healthcare, the judicial system, and public administration. These sectors are fundamental for societal life and—as I will argue shortly—are of main interest for my analysis on citizen freedom in the digital age. In what follows, I first elaborate on these cornerstones, after which I provide my thesis overview and discuss more specific research questions.

Freedom in the digital age: setting the stage

Let me start by explaining how I conceive of freedom, and how freedom and power are related. Zuboff provides a good starting point: our individual freedom is undermined because our ability for *self-governance* is reduced, if not hijacked. According to Zuboff, the pinnacle of one's freedom is autonomy, the ability to make one's own life choices. If we are constantly monitored and nudged in a particular direction—especially without us knowing—this aspect of freedom is undermined because such profiling removes what Zuboff calls the 'right to a future tense' (2019a, ch. 11). In the future tense, we live in uncertainty that allows for self-development or, in Hannah Arendt's words, for *action*. Arendt (1958) describes action as the activity that allows humans to develop themselves as unique human beings and, therefore, is an activity at the heart of realizing freedom. However, as Zuboff (2019b, 38) writes: "If the right to the future tense is abrogated, the miracle of human action is subordinated to others' plans that favor others' certainty." We can thus reformulate Zuboff's claim about the reduction of our freedom in the following way. Our freedom is undermined because we are subjugated to processes that affect how we develop ourselves according to other people's wishes instead of our own.

While I fully support Zuboff's conclusion to the effect that our freedom is reduced—and I second that the subjugation to external processes is a fundamental aspect of the reason *why* our freedom is reduced—Zuboff makes a conceptual error in assessing the moment our freedom is undermined. For Zuboff, the reduction in freedom is a *consequence* of uncontrolled power. She writes:

> It [surveillance capitalism] is an unprecedented market form that roots and flourishes in lawless space. It was first discovered and consolidated at Google, then adopted by Facebook, and quickly diffused across the Internet. Cyberspace was its birthplace because, as Google/Alphabet Chairperson Eric Schmidt and his coauthor, Jared Cohen, celebrate on the very first page of their book about the digital age, "the online world is not truly bound by terrestrial laws…it's the world's largest ungoverned space" (Zuboff 2016).

The fact that the digital space was unregulated set the stage for Google to use people's data as a way to increase their profit. Where initially Google treated its users as an end in themselves, aiming to ensure that the users would have the most optimal search experience, it changed to treating humans as a *means* to their own end (i.e., profit-maximization) through personalized advertising. Nudging people into specific directions by means of personalized advertising affected people's future tense. This undermined people's freedom because they became subjugated to Google's and its advertisers' wishes instead of their own.

The conceptual error I believe in Zuboff's analysis is the disconnect of the uncontrolled power from freedom. In Zuboff's view of freedom in the digital age, freedom is reduced when one is actually interfered with (such as when Google captures our data). It is at this moment that our ability for 'action' becomes subject to another's plan, and is therefore a causal consequence of Google's uncontrolled power—i.e., what Zuboff refers to as the lawless space in which Google initially thrived. However, rather than seeing the lack of freedom as a consequence of this uncontrolled power, I suggest that the uncontrolled power is *constitutive* of the reduction in freedom. This distinction is small but significant. It relates to a difference in viewing freedom as contingent on a specific act or as a relational entity where one's freedom depends on one's relations to others.

Freedom as a relational concept is found implicitly in the 'overlord' and 'serf' relation suggested by Varoufakis in the context of online platforms. Although Varoufakis does not elaborate on freedom in his analysis of the problem with digital platforms, the analogy with 'overlord' and 'serf' captures a specific relation between tech companies and tech users. Similarly, the concerns of Nemitz or Fukuyama et al. (i.e., concerns of companies escaping the rule of law or companies picking up the loaded guns) capture a certain vulnerability in which society finds itself. *If* the wrong person picks up the gun, society is in big trouble. Consequently, society stands in a continuing vulnerable relation to tech companies.

This conceptual error between freedom in terms of acts and freedom in terms of relations reflects a current debate in political philosophy between different conceptions of freedom. On the one hand, there is the more common liberal, negative

construal of freedom, i.e., freedom as non-interference, that treats freedom as contingent on acts (Berlin 1969). On the other hand, there is the republican conception of freedom as non-domination where one's freedom is reduced when one is subjected to the arbitrary or uncontrolled power to interfere (Pettit 1997).[3] Here, a person's freedom is contingent on their relations with others.

Adopting the 'right' conception of freedom matters for finding appropriate means to safeguard democracy. Liberal democracies, and particularly the US, as Zuboff herself highlights, draw heavily on a negative conception of freedom.[4] However, such a negative construal omits highly relevant aspects of freedom as it excludes the decision-making power regarding on whose terms the interference occurs. Focusing primarily on the *act* presents a one-dimensional view of freedom that misses the depth to secure the *robustness* of this freedom. Therefore it fails to address the vulnerable state of society. In my view, Zuboff's conceptual error is that she treats freedom according to a negative understanding (e.g., people's freedom is reduced when targeted by Google) rather than as a relational concept (e.g., the relation between serf/tech user and overlord/tech company). The republican conception is the one I endorse in this thesis (and, as I will argue shortly, one I believe Zuboff implicitly supports as well).

Rethinking freedom: from liberal to republican

Neo-republican theory, revivified by philosophers and historians such as Philip Pettit and Quentin Skinner, is a reaction to the common liberal understanding of freedom as non-interference, as captured in Berlin's (1969) conception of negative freedom. According to this negative conception of freedom, one is free if one is not interfered with. Interference, here, I take to be the imposition of negative social constraints and coercion.[5] If we are to take such non-interference seriously, however, there are some

---

3 With republican theory, or *republicanism*, I do not refer to the US Republican political party but to a particular political theory that has its root in ancient Athens and Rome. Republican freedom is therefore also by some (e.g., Skinner 2008) referred to as *neo-roman* freedom. I will be using the republican terminology as most prominently developed by philosopher Philip Pettit (1997; 2012), as this is the most widely accepted.

4 See e.g., Stiglitz (2023) for how a narrow interpretation of negative liberty has come to take hold over US ideologies.

5 What I mean with *negative* is that the interference can prevent the agent who is interfered with from realizing their goal. This is done either by obstruction or removal of options or not providing the necessary means to realize one's goal. What I mean with *social* is that the constraint is not from natural limitations, but always stand in relation with another agent. List and Valentini (2016) give the

important and counterintuitive implications. For instance, in a master-slave relation, according to an extreme interpretation of freedom as non-interference, the slave would theoretically be free if the master never interfered in the slave's life. However, it seems intuitively wrong that a slave can be free (List and Valentini 2016). The issue is not so much that the master actually interferes. Rather, the slave remains vulnerable to the *potential* interference by his master. Although the master might not decide to interfere with the slave right *now*, there are no checks and balances that secure this non-interference in the future. The slave is thus subjected to his master's arbitrary power that makes interference always a possibility.

Such an existence of (the possibility of) arbitrary power is referred to as *domination* in the republican tradition (Pettit 1997; Lovett 2010). For republicans, to enjoy real freedom is to be free from domination, free from someone else's arbitrary power to interfere, regardless of whether that power is exercised. Non-arbitrary power requires that the power exercised is forced to track the subordinate agent's best interests (Pettit 1997). In order to make sure that these best interests are indeed tracked, non-arbitrary power entails that the power must be controlled—either directly or indirectly—by the people subject to that power relation (Pettit 1997; 2012). After all, only they can indicate what their best interests are. A core feature of non-domination is thus the ability for self-governance, which is also why republicanism is inherently supportive of democracy.[6]

Earlier I mentioned that Zuboff is committed to the republican conception of freedom. To see this, consider again her 'right to the future tense'. According to Zuboff, the right to the future tense encapsulates the idea "that I can project myself into the future and thus make it a meaningful aspect of my present" (Zuboff quoted in Laidler 2019). What Zuboff implies is that because our future is *increasingly* determined by behaviour modification (i.e., nudging, steering, manipulation), it is *decreasingly* part of us. In other words, Zuboff's worry is that behaviour modification shapes us to conform according to someone else's plan, the future is increasingly less 'ours' to shape according to our own ideas and wishes.

---

example that I am unable to jump ten meters in the air, simply because that is physically impossible. However, it is not that the earth 'interferes' with me. In addition, interference also includes coercing an interfered-with-agent into doing something they otherwise would not have done (e.g., paying taxes). This can be done by means of penalties (e.g., a fee when late in paying taxes) or (threats of) physical or emotional abuse.

6 See Lovett (2012) for an opposing view.

Such a right to the future tense is reflected in paradigm cases in republican theory that emphasize the relational aspect of freedom. The idea is this. When one is dominated, one lives in a vulnerable, insecure state that makes it impossible to plan ahead. Consider a husband-wife relationship in a sexist patriarchy, where the wife depends on the husband's permission to do anything of significance (e.g., spending time with friends). She cannot make any meaningful or reliable plans for the future as her husband may change his mind at any time whether he allows her to go out.[7] This reflects the idea that in order to make my future self part of my present self, I must be in control of my own life and not be dependent on the whim of another whether my plans, hopes, and dreams will succeed. Zuboff's phrasing suggests that what makes me free requires me to be *in control*, to have the ultimate say. Such control is not an 'on' or 'off' switch. It is an ongoing relation with others.[8]

Zuboff's conceptual error thus lies in the following tension. On the one hand, Zuboff strongly focuses on *what* the interference of online platforms does to their users and less so on the fact what these companies *could* do so in the first place. Here, Zuboff's conceptual framework builds on a narrow conception of freedom as non-interference. On the other hand, the fundamental problem why people are undermined in their freedom seems to be precisely in people's lack of control over the interference. This requires a relational conception of freedom, which better corresponds to a conception of freedom as non-domination. By reasoning from a negative conception of freedom, Zuboff disconnects the uncontrolled power of these companies from the concept of freedom itself. This conceptualization of freedom, however, obscures the real threat to liberal democracies. What Zuboff, Varoufakis,

---

7 A fictional though telling example of what I have in mind can be found in the Māori film *Once Were Warriors* (Tamahori 1994) based on the similarly titled book. In the movie, Beth made several plans she could not execute due to her husband Jake's last-minute interferences or changes in decisions. Although the story depicts a relatively empowered Beth towards the end, throughout the movie she is fully dependent on Jake's will who prevents her from realizing any meaningful plans for the future.

8 Some clarification is in order. People *always* depend on others to achieve their dreams, hopes, and aims. The keyword is 'arbitrary' dependence. Although this point will become clearer in the dissertation itself, ultimately what establishes 'control' is whether I have the ability to influence, direct, and contest decisions. A 'latent' form of control here suffices. As long as the horse gallops in my desired direction, I do not need to intervene. Yet as soon as the horse changes course, I have the possibility to intervene and alter its course to where *I* wanted to go. This is different from me sitting on a horse over which I have lost all ability to steer its course. I do not decide where to go: the horse does. For republicans, I will be free as long as the horse goes in my preferred direction. Of course, in a society things are much more complicated, and 'my' direction generally becomes the 'common' direction of the public (Pettit 2012).

Nemitz, and Fukuyama et al. all seem to hint at is that this threat to freedom is not interference. It is uncontrolled power. Therefore, ironically, liberal democracies and their advocates misconceive the nature of the phenomenon that poses the biggest threat to them.

To successfully safeguard freedom and democracy, it is essential to find the correct way to conceptualize the root source of the problem. In Zuboff's (2016, 4) words: "every successful vaccine begins with a close understanding of the enemy disease." For Zuboff, this is surveillance capitalism. For Varoufakis, this is technofeudalism. For me, throughout this thesis, the enemy disease is domination. Republican theory and its conception of freedom as non-domination provide the normative and conceptual framing suitable to understand the concerns present in the digital age, as well as to address them. This does, however, require an update of the conceptual background of freedom that informs liberal democracies.

Let us take stock. So far, I have addressed the first two cornerstones of this dissertation. These were how we should understand freedom and how freedom relates to uncontrolled power. The freedom I am interested in is republican freedom, a conception that is necessarily linked to power relations. As long as the power is controlled by those subjected to the superior power, they are free. I now turn to the third and final cornerstone, i.e., what I mean precisely by the digital realm.

The digital realm: not just online platforms

Republicanism proves useful to assess the power of online platforms, and captures worries raised by Nemitz, Fukuyama et al, Zuboff, Varoufakis, and others who have emphasized similar concerns (e.g., Benn & Lazar 2022; Schaake 2024). Indeed, in recent years, we have seen increasing interest in how the digital realm creates relations of domination and hence undermines our freedom (Susskind 2022; Aytaç 2022; Hoeksema 2023; Muldoon & Raekstad 2023, Muldoon 2023). James Muldoon and Paul Raekstad (2023), for instance, are interested in how digital domination occurs in the context of the Gig Economy where gig workers are subject to algorithmic control. Uğur Aytaç argues we are dominated *qua citizens* as social media affects our capacity to partake in political life. This makes Aytaç's area of interest the relation between citizens and the state. Finally, Bernd Hoeksema (2023) focuses on how online platforms more generally lead to domination because of underlying societal structures.

In this dissertation, I am interested in how digital domination applies to relations of power amongst citizens of modern, digital societies[9] and how digital domination extends to digital technologies beyond online platforms. After all, the digital realm is not *just* online platforms. Let me first elaborate on digital technologies beyond the online realm, after which I will discuss the relations of power in more detail.

Increasingly, AI systems are used in core societal sectors, such as healthcare, public administration, and the judiciary. The AI systems I have in mind are (semi-)automated decision-making systems such as the allocation of childcare benefits or assessments of whether people need health care. In Rawlsian terms, such AI systems shape our background structure and basic institutions (Gabriel 2022). With the increased deployment of AI systems in core societal sectors, we also increasingly become dependent on these systems to function in society. This is different from online platforms. We can still avoid Google or Facebook and function reasonably well in society (although surely it will make our life more difficult). Yet, we simply cannot function in society without subjection to the systems that make up our technological and institutional background structure.[10] The choice is either be part of society, and thus be subjected to these AI systems, or live elsewhere (which increasingly becomes less of a possibility). Whenever we talk about freedom in the digital age, we should thus not limit ourselves solely to the online realm.

As of yet, republican theory is largely undiscussed beyond online platforms. Jamie Susskind (2022) is an exception. However, he moves between online platforms and 'core' AI systems (i.e., AI systems used in core societal sectors), making it unclear why core AI systems constitute relations of domination. For example, in his positive proposal to mitigate digital domination, he focuses solely on social media platforms, not providing an account of how we can mitigate digital domination beyond online platforms. This narrow focus on online platforms is a missed opportunity. Given the unreasonably high exit costs with regard to 'opting-out' of the core societal sectors, questions of freedom and power extend beyond online platforms.

Moving on to the relations of power, the AI Ethics literature has been concerned with questions of power in the context of core AI systems for some time. In order to situate myself within this literature, let me take a step back. As is well known, core AI

---

9 I will focus primarily on Western societies such as we find in Europe and North-America, although some of my arguments can be applied more globally.

10 Of course, some might argue that such online platforms have become part of our social structure. See e.g., Fuchs (2015). I am sympathetic to this reasoning, but whether this holds is unproblematic to my overall argument.

systems give rise to ethical concerns, most notably unfair discrimination. Generally, by now, it is widely recognized that automated decision procedures in various societal sectors negatively affect already marginalized groups more than non-marginalized groups (Eubanks 2018; Barocas & Selbst 2016). For instance, in the Netherlands, the use of an AI system for the detection of fraudulent behaviour amongst applicants for childcare benefits has been criticized for indirect discrimination against people from a particular ethnic group (*Netherlands Institute for Human Rights 2023*). In the US, an AI system was designed to allocate health care to patients, but it failed to provide people from the black community with their required health care due to a poorly chosen proxy (Obermeyer et al. 2018; Benjamin 2019). Concerns for ensuring 'Fair' AI are thus commonly conceived as one of the top priorities within the AI Ethics literature[11] (Johnson 2021).

The AI Ethics literature can be divided into two strands, or what has also been referred to as 'waves' (Pasquale 2019). The aim of the first wave is to improve already existing AI systems through initiatives that aim to develop AI systems aligned with ethical principles and values. These aims are conceptualised in numerous guidelines (see for an overview Jobin et al. 2019), pledges by companies developing the systems,[12] and technology fixes to ensure 'Fair' or 'Explainable' AI (Hirota et al. 2022; Gunning et al. 2019). Given that the first wave is interested in improving systems, it is predominantly concerned with the outcome and behaviour of the system.

The second wave takes on a broader perspective and can be seen as a critique of the first wave. The criticism is grounded in how the first wave remains too narrow as it overlooks decision-making power regarding these systems. Indeed, we can design a 'Fair' AI, but how we define fairness matters for who is affected in what way by the system. Questions of who is in charge of deciding which systems should be built and which problems must be tackled have led scholars to emphasize power asymmetries between developers and deployers (e.g., programmers and CEOs of AI companies) and the people that are subject to these systems. Such work points to the need for effective accountability mechanisms that ensure any AI system that affects people's lives supports democratic values such as contestability, accountability, and representation of the public interest (Powles and Nissenbaum 2018; Kalluri 2020; Zimmermann et al. 2020; Crawford 2021; Sloane et al. 2022; Birhane et al. 2023). The second wave is thus generally represented by addressing broader questions

[11] Another top priority is the question of explainability which helps to address responsibility and accountability concerns.

[12] See e.g., https://openletter.svangel.com/

regarding whether these systems should exist in the first place, and if so, who gets to decide and who gets to govern these systems.

Although the second wave explicitly deals with issues of power, what is noteworthy about the general focus of these second wave scholars is the *target* of who should have a say. Given that marginalized groups have the least say and are most affected, much attention in the second wave goes to increasing the say of marginalized groups specifically (Kalluri 2020; Zimmermann et al. 2020). The power asymmetry these authors are interested in is thus the power asymmetry between the non-marginalized and marginalized groups in society.

While I applaud the attempts in the second wave to provide marginalized groups with a voice, it is important to explicitly distinguish between two forms of power relations. One concerns the power asymmetry between marginalized and non-marginalized. The other concerns the power asymmetry between what I will refer to as the 'shapers' (i.e., developers and deployers) of an AI system and the 'affected' (i.e., end-users) more broadly.[13] This second power relation has received limited attention (see for exceptions e.g., Binns 2018; Dobbe et al. 2021). This is surprising given its fit within the question of who gets to decide and govern these systems, which becomes ever more urgent as AI systems increasingly fulfil roles within core societal sectors. My dissertation will provide more substantive evaluations and analyses on these power asymmetries as a more novel (and urgent) contribution to the second wave.

There is a strong republican reason why it is important to explicitly distinguish between these different types of power relations. While Kalluri (2020) is correct in pointing out how the developers and deployers of AI systems stand in a particularly powerful relation with regards to those affected, her analysis is specifically interested in those that, *in fact,* face the negative effects, i.e., members of marginalized groups. However, any concern for republican freedom in the digital age is whether someone *could* have interfered (e.g., by denying me healthcare or social benefits) and not just whether they did. In other words, a slave to a kind master is as much a slave as one to a cruel one. Work done in the second wave is indispensable to provide citizens in digital societies with, to speak in republican terms, a "voice worth hearing and an ear worth addressing" (Pettit 2001, 350). Yet these voices and ears apply equally to those negatively affected (those subject to a cruel master) as to those who do not experience negative consequences (those subject to a kind one). Thus, besides purposely moving beyond online platforms and focusing my discussion of freedom in the digital age on

[13] In Chapter 2, I elaborate on these concepts.

the core societal sectors, I specifically focus on the power asymmetry between the shapers and the affected of AI systems.

The question I asked at the beginning of the introduction that informs the general background of this dissertation is how our freedom is undermined by uncontrolled power in the digital realm. This required me to provide the following three cornerstones: (1) a relevant understanding of freedom, (2) how this understanding of freedom relates to uncontrolled power (i.e., arbitrary power relations), and (3) an understanding of the 'digital realm'. Regarding the first two, the freedom I am interested in is a republican conception of freedom as non-domination that is necessarily linked to the power relation between a superior and subordinate agent. When the subordinate agent has no control over whether and how the dominant agent may exercise their power, the subordinate agent is subjected to an uncontrolled power. This constitutes domination, making them unfree. Think again of the dominated wife who lacks control over whether she will be able to leave the house, as such decisions fully depend on her husband's will. Regarding the digital realm, I am specifically interested in the broader scope of the digital realm, including—and especially—those technologies that are not part of online platforms *per se*, but are unavoidable for people to interact with (either passively/actively/directly/indirectly) in order to live in modern, digital societies. In addition, I am interested in the power relations between the shapers and affected of AI systems. With these cornerstones in mind, I am now ready to discuss the thesis overview.

Thesis overview

The thesis I defend in this dissertation is that AI systems used in core societal sectors lead to relations of domination in society between those who shape a system and those affected by it. Specifically, I argue that all members of modern, digital societies are at risk of such domination, even when they are unlikely to be actively interfered with (i.e., being constrained or coerced in one's choices), making digital domination a problem for all. In order to defend this thesis, I will first argue that there is a power relation between the shapers and the affected. Then, I will claim this power relation is uncontrolled or arbitrary, hence constituting domination. Finally, I will argue that this relation applies to all members of modern, digital societies, not just those actually facing interference. If I am successful in arguing for my thesis, I can conclude that we are unfree according to a republican conception of freedom in the context of AI

systems. Moving forward, I can then assess what is necessary to safeguard freedom in the digital age.

To anticipate the main conclusion of this dissertation, I propose that in order to achieve a non-dominating technology sector (and therefore freedom in the digital age), we must *design for non-domination*. A design for non-domination strategy requires addressing direct power relations between the shapers and affected of AI systems by means of design-for-values approaches such as value-sensitive design or participatory design. However, designing for non-domination goes beyond these direct power relations. As I argue throughout my dissertation, broader societal structures and context are part and parcel of why certain people obtain a particular powerful positions, as well as why such powerful people are able to exercise their power in an arbitrary, unconstrained way. This includes, specifically, the lawless space in which digital companies initially thrived, as this space allowed companies to develop and deploy these digital technologies into society in the first place. Designing for non-domination, thus, must address broader societal norms, values, and structures that enable and uphold relations of domination in society. This will require, I (controversially) claim, to reconsider a presumed freedom to innovate we most vividly find in neo-liberal societies.

This final claim makes explicit an underlying point of my thesis. I do not merely wish to state that we are unfree on a republican conception of freedom. I believe the republican conception of freedom better captures moral concerns raised by authors like Zuboff and Varoufakis. I intend to make the bolder claim that we need to rethink the conceptual and normative background of freedom in the digital age in general. The relevance in doing so is that one's conception of freedom informs strategies to *safeguard* this freedom as well. Reconceptualizing freedom according to a republican definition thus proves a more suitable vantage point to depart the journey for safeguarding freedom in the digital age.

I will defend my thesis—and provide suggestions on how to move forward—in a total of six chapters, each with an interlude prior to them.[14] These interludes serve as a moment to reflect on the previous chapter and provide some reflection and background for the next. I reserve the more substantive reflections and objections for the conclusion. In the remainder of this introduction, I discuss each chapter in turn, show how they fit into my overall thesis, and discuss which research gap they are meant to answer.

---

[14] Some of these chapters are co-authored. At the start of each chapter, I will state whether the chapters are solo or co-authored.

Chapter 1

The first chapter, *Technology as a Driver for Morally Motivated Conceptual Engineering*, provides the methodological background for my dissertation. Here, I argue that there is value in rethinking concepts in light of technological advancements by drawing on three examples. These are how self-driving vehicles have challenged the notion of human control, how social media platforms invite us to reconsider the concept of critical thinking in light of content moderation, and, specifically relevant to this thesis, how big tech companies challenge our notion of freedom. The chapter addresses the more specific research question related to my thesis:

> **RQ1:** What is the value of rethinking freedom in the digital age?

I answer this question by drawing on the field of conceptual engineering. One aspect of conceptual engineering is to assess whether certain conceptions of concepts are the 'right' ones to deploy. For instance, JUSTICE is a concept, and Rawls' theory of justice as fairness is a particular conception of this concept. Questions that are interesting for scholars engaging in conceptual engineering are whether justice as fairness is the 'correct' interpretation (i.e., conception) of JUSTICE, or if perhaps another conception of justice better fits the general aim of what JUSTICE is meant to achieve.[15] Conceptual engineering thus takes on a broader perspective and steps out of a particular normative context or conceptual bubble.

The chapter *Technology as a Driver for Morally Motivated Conceptual Engineering* shows that technologies may give us reasons to critically assess societal norms and values that are dominant in a specific society. Yet, although online platforms have motivated other authors to argue that relations of domination exist in society, none have explicitly made the connection to what this precisely entails for our concept of FREEDOM in society. It is important to make explicit how conceptual engineering and technological advancements are often intertwined. Failing to update our conceptions as technologies change society is what can lead us to misidentify the root source of the problem. This chapter provides me the set-up to go further than previous authors and argue that such relations of domination challenge the normative context in which we operate.

---

[15] In the chapter, I expand on this discussion.

Chapter 2

Having laid out the general structure and methodological approach I adopt in this thesis, the second chapter, *Machine Learning and Power Relations*, provides an analysis of 'power relation' that later is needed for the subsequent chapters. To repeat, my thesis is to argue that core AI systems (AI systems used in core societal sectors) lead to relations of domination in society between the 'shapers' (developers and deployers) and 'affected' (end-users) of a system, and that this is a problem for all members of modern, digital societies. The first step in justifying this claim is to conceptualize the power relations underlying such technologies. In other words, I need to address the following question:

> **RQ2:** How should we understand the power dynamics underlying AI development and deployment between those who shape a system and those affected by it?

In *Machine Learning and Power Relations*, I provide a workable conceptualization of the power asymmetry underlying AI systems between the 'shapers' and 'affected' of a system. This power relation exists due to the fact that these systems affect the people (i.e., end-users) subjected to them. In turn, the shapers of the system decide how the system behaves (e.g., by favouring a particular conception of fairness over another, setting error rates, and deciding that the system will be developed and/or deployed). In order for the end-users to obtain their desires and goals, they depend on the system's output. Yet because the system's output, in turn, depends on its shapers, the end-users indirectly depend on the shapers of the system. Thus, there is a power relation between shapers and affected *via* the AI system.[16] My proposed conceptualization of the power dynamics underlying AI development and deployment is fundamental for the overall thesis as it provides the foundation to assess the potential for domination of AI systems used in core societal sectors.[17]

---

[16] There will be unavoidable consequences of the AI system on the end-user that were not intended by the developers and deployers (e.g., unnoticed biases in training data due to historical social injustices). This emphasizes the need for stakeholder involvement regarding the question whether the system should exist in the first place, which remains a deliberate choice by the shapers. I discuss the need for stakeholder involvement in more detail in Chapter 5.

[17] Note that AI systems are necessarily socio-technical. They are developed by people with a specific background and deployed (and interpreted) by people in a particular context (Hildebrandt 2021; Dobbe et al. 2021; Crawford 2021; Sartori & Bocca 2023; Noorman & Swierstra 2023; Johnson & Verdicchio 2024).

Chapter 3

For an instance of domination to occur, it is not sufficient to merely argue there is a power asymmetry between the shapers and the affected of an AI system. For domination, I need to argue that this power relation is *uncontrolled*. I present more detailed arguments for why this is the case in Chapters 4, 5, and 6. In Chapter 3, *A Neo-Republican Critique of AI Ethics*, I elaborate on the conceptual background that informs AI Ethics and claim that this background provides insufficient means to cater to power relations more generally. This has led me to the following research question:

> **RQ3:** How does the conceptual background of freedom that informs the AI Ethics literature steer the focus away from the power dynamics underlying AI development and deployment?

In the chapter, I argue that AI Ethics is strongly rooted in Mill's harm principle, which in turn is grounded in a conception of freedom as non-interference. How to precisely interpret the harm principle is strongly debated (Folland 2022). Furthermore, it has been argued that Mill's harm principle, in fact, includes relations of arbitrary power (Urbinati 2002). This debate is something I do not explicitly engage in within the chapter (and I will expand on this limitation in the accompanying interlude). However, what is evident for Mill is that *harm* is the problem, as opposed to the robustness of that harm, which is the concern in republican theory. Because the conceptual background informing the AI Ethics literature is rooted in Mill's harm principle along with a conception of freedom as non-interference, the literature risks not being able to sufficiently cater to power relations. The reason is because the field, by virtue of its specific conceptual background of freedom, leans more towards mitigating harmful acts (focusing on the 'interference') rather than addressing underlying power asymmetries (focusing on the 'domination').[18] Based on the analysis provided in Chapter 2, I furthermore argue that this conceptual background is problematic because of the potential for domination to occur.

Chapter 4

Having argued that freedom as non-domination is insufficiently considered in the AI Ethics literature, it is another point to show that the AI Ethics literature insufficiently

[18] The second wave in AI Ethics does address power relations. Chapter 3 therefore primarily addresses the first wave in AI Ethics that is concerned with improving systems.

caters to power asymmetries between the shapers and affected. So far, the primary focus has been on how these power relations play out directly. But these relations are necessarily embedded in a broader societal context that either enables or inhibits certain actors from exercising their power in unconstrained ways. In chapters 4 and 5, I discuss to what extent the AI Ethics literature insufficiently considers this broader societal context, and why this focus is essential. Chapter 4, *Opening the Black Box of AI, Only to Be Disappointed*, provides the initial steps and responds to the following question:

> **RQ4:** Why should the AI Ethics literature consider a broader societal perspective?

I provide two main answers to this question. The first relates to a concern of how a narrow sense of AI Ethics[19] could lead to 'political naivete'. With political naivete, the idea is that AI systems can be developed in accordance with certain ethical principles and values, yet when deployed, miss achieving the goal of being ethical in a meaningful way because the broader political context was overlooked. Think, for instance, of a content-moderation system that is meant to safeguard free speech by avoiding echo chambers yet that is deployed in an authoritative regime that censors its citizens. Such a system aiming to avoid echo chambers loses its goal (i.e., reducing echo chambers) given that people under state censorship live in an echo chamber regardless (e.g., citizens in North Korea).

The second concern is that many of these systems are not (yet) properly embedded in politico-legal institutions. In a sense, they are 'free-floating' without a guarantee that what they promise is what they deliver.[20] In the chapter, I call such politico-legal embedding 'deontic provenance', referring to the origin of the system (e.g., who its developers and deployers were). I argue that the ability to trace back a system's deontic provenance improves accountability as it provides people in society with the possibility to contest the decisions made during the development of the system. The possibility for contestation is essential because it provides a form of public control (via

---

19 Such a narrow sense of AI Ethics is often discussed in the context of the first wave of AI Ethics. However, as I argue in Chapter 5 some initiatives in the second wave risk overlooking broader societal structures as well.

20 The infamous *Ashley Madison* dating site for married people had several checks of security on their website, providing the false indication data provided to the website was safely secured. These checks, as we now know, were self-made using photoshop. In other words, they were not embedded in broader politico-legal institutions that (1) provided a guarantee of security and (2) that could have given the user legal backing when their data was leaked.

contestation), which increases the likelihood that AI systems track the best interests of the public. This is also what Langdon Winner refers to as 'political ergonomics'—the fit between technology and society. To anticipate the conclusion of this dissertation, tracking the public's best interests (i.e., being politically ergonomical) is a core requirement for an AI system to be non-dominating. Thus, tending to the concern of political naivete and ensuring we can trace a system's deontic provenance are necessary conditions to design for non-domination.

Chapter 5

Chapter 5, *Beyond Participatory AI*, continues the focus on the broader societal context. While much debate of AI Ethics is concerned with participatory AI to mitigate unjust power asymmetries (especially the second wave), this chapter shows how current initiatives are insufficient. It proposes suggestions on how to move forward by actively focusing on the broader societal context. The research question informing this chapter is as follows:

> **RQ5:** How can we move beyond current initiatives in AI Ethics in order to successfully address the power relations between the shapers and affected underlying AI development and deployment?

The chapter provides two suggestions on how to move beyond limitations of current initiatives in AI Ethics. For one, there must be a form of *robust* integration of stakeholders in order to address the arbitrariness of power relations. Currently, stakeholders depend on the goodwill of AI developers and deployers to be included during the design process. This, however, does not sufficiently provide the affected with the control necessary to address the arbitrary power of the shapers. Moreover, relating to the broader context mentioned in Chapter 4, I propose that participatory AI should address the broader political economy of which AI systems are part. One of the aims of participatory AI is to effectively empower stakeholders. However, current socio-economic structures necessarily empower some and disempower others. In order to empower stakeholders, we thus need to actively integrate the socio-economic structures underlying AI development and deployment within AI Ethics initiatives concerned with power dynamics.

Chapter 6

At this point in the dissertation, Chapters 1 through 5 will have led to the following main claims. I will have argued that there is a power relation between the shapers and the affected of an AI system (Chapter 2), that this power relation can be morally problematic from a republican perspective (Chapters 2-3), and that this moral concern is insufficiently addressed in current initiatives in AI Ethics (Chapters 3-5). As mentioned, the AI Ethics literature is limited as it is predominantly concerned with preventing harm, as opposed to other moral wrongs such as domination. Furthermore, the AI Ethics literature insufficiently addresses broader societal contexts, such as how AI systems are embedded in political and legal institutions, and how 'participatory AI' that genuinely empowers stakeholders requires to address the broader political economy in which AI is developed. These claims, however, require elaboration in order to defend my thesis that core AI systems dominate all citizens in society. So far, I have remained ambiguous on (1) who is dominated, (2) who does the domination, and (3) when precisely the domination occurs. Combining the points made in the previous chapters, in the sixth and final chapter, *Making Sense of Digital Domination*, I provide a coherent account to argue why all citizens of modern, digital societies are dominated. The overall question I aim to answer is the following:

> **RQ6:** How should we understand domination in the digital age?

I propose there are at least three different ways to understand digital domination. To be precise on which account is the correct one matters for how we should mitigate relations of domination in the digital age. I argue in favour of what I call the *socio-economic perspective*. This understanding of digital domination holds that anyone subject to AI systems faces what Dorothea Gädeke (2020) refers to as *interpersonal* domination. One way to address such interpersonal domination is by providing effective regulation and legislation, such as specific regulation concerned with stakeholder inclusion. However, from this socio-economic perspective, the root source of the domination has to do with broader socio-economic structures that allow for technological innovations to be developed and deployed beyond public control. This refers back to the 'lawless space' that gave rise to Zuboff's surveillance economy discussed previously.

As mentioned, Zuboff primarily took issue with the *consequences* of this lawless space. I, on the other hand, take issue with the lawless space itself. I see the lawless space as a space beyond public control that provides tech companies to develop (and exercise) their power. Robust freedom, i.e., freedom from domination, requires not only 'fixing up' instances of domination, but it also requires addressing power structures that allow

for such domination to occur. The socio-economic perspective I propose suggests that domination in the digital sector is, in fact, a *symptom* of a more problematic socio-economic order. This socio-economic order which combines a lawless space to innovate with a (neo-)liberal ideology that strongly supports people's freedom to innovate is problematic as it give rise to dominating societal structures. The reason the socio-economic order gives rise to domination is because a lawless space is beyond public control. Non-domination, on my proposed socio-economic perspective, requires an innovative space that is subject to public control that forces companies to track the best interest of the public. This chapter then provides a forward-looking glance to further research, namely how to shape our socio-economic structures so that innovations are under public control.

Conclusion

In the conclusion, I provide a summary of the chapters, and propose how we can address relations of domination in the digital age based on the more detailed analysis provided in Chapter 6. In order to improve our freedom in the digital age, we need to design for non-domination. What is necessary to mitigate relations of domination is improving both the regulatory initiatives with regard to AI systems as well as ensuring that innovations more generally are subject to public control. This requires a fundamental restructuring of underlying socio-economic structures.

***

I conclude my dissertation with the idea in mind that innovation *itself* poses concerns of domination. The neo-liberal ideology enabled many online platforms to take advantage of the lawless cyberspace, which, as Zuboff shows, transformed in many cases into surveillance capitalism. Domination was present already within the lawless space of AI development and deployment, before companies could abuse this space for their own profit maximization or before achieving their status of overlords by renting out their cloud capital. Rather than addressing the *surveillance* in surveillance capitalism, as Zuboff does, or the overlords that Varoufakis does, my dissertation suggests that we need to rethink whether this neo-liberal ideology that provides a form of "anarchic innovation" (Hussain 2023, 127) is desirable in the first place.

This thought connects to recent works in economics such as *The Road to Freedom* by Joseph Stiglitz (2023) and *Power and Progress* (2023) by Daron Acemoglu and Simon Johnson. Both books show how neo-liberal and libertarian ideologies have led

innovations astray, favouring those who do the innovations rather than society at large. My dissertation strengthens such works by conceptualizing precisely *why* innovations should be done under public control: it is required to ensure a free society for all. Innovations done beyond public control pose an existential threat to liberal (or, more correctly, *republican*) democracies. This threat need not become a reality. However, simply designing *one* AI system according to non-dominating principles (e.g., including stakeholders in the design process to provide them with the possibility to contest decisions) will not be sufficient to resolve the structural domination associated with neo-liberal and libertarian ideology. To successfully design for non-domination, we must integrate the broader societal structures in development and deployment processes of technological innovations.

Interlude 1

The first chapter discusses the general idea that technologies can be a source of rethinking certain conceptions in society. This chapter is co-authored with Herman Veluwenkamp, Marianna Capasso, and Lavinia Marin. My contribution to this chapter has been on the discussion of how the digital age gives us reason to rethink our conception of freedom. The research question I address in this chapter that informs the overall dissertation is the following:

> **RQ1:** What is the value of rethinking freedom in the digital age?

In order to address this question, consider first the co-evolvement of concepts and underlying societal norms and values more broadly. For instance, the concept of RAPE initially excluded marital rape (Banerjee & Rao 2022). This contributed to a continuing status of inferiority that women endured in society which, during the feminist movements throughout 20th century, became too outdated for the norms and values upheld in modern societies. By expanding the concept of RAPE to include marital rape, women who were married conceived the conceptual tool (i.e., RAPE) to express a particular moral wrong. An upshot of rethinking—and if necessary updating—dominant concepts in society is that it allows for a more targeted intervention to resolve potential concerns. For instance, in the case of expanding rape to include marital rape, the updated concept enabled married women to file charges against their husbands on account of rape.

A similar reasoning applies to conceptions that require revisiting due to technological advancements. The co-evolvement between technologies and societal norms and values potentially outdates conceptions that no longer suffice to capture societal change. In this way, conceptual engineering occurs as a response to moral intuitions that are insufficiently captured by dominant (or what we call in the paper 'operative') conceptions. The following chapter presents the argument that technological innovations can give us reasons to reflect on (and, if necessary, reconceptualize) moral concepts in society. Such reflection makes salient the particular normative context in which we operate and allows society to update this context if it insufficiently meets normative standards. This helps in providing means to mitigate moral wrongs. The example mentioned in the chapter relevant for my dissertation is how society takes for granted a conception of freedom as *non-interference*. This conception of freedom as non-interference, however, does not capture the moral

intuition of what precisely is wrong with the power of online platforms. We propose to adopt the republican conception of freedom as *non-domination*, which is better suited to capture these moral intuitions. In Chapters 2 and 3, I expand this idea to core AI systems.

Chapter 1: Technology as Driver for Morally Motivated Conceptual Engineering[21]

Abstract New technologies are the source of uncertainties about the applicability of moral and morally connotated concepts. These uncertainties sometimes call for conceptual engineering, but it is not often recognized when this is the case. We take this to be a missed opportunity, as a recognition that different people are working on the same kind of project can help solve methodological questions that one is likely to encounter. In this paper we present three use-cases where philosophers of technology implicitly engage in conceptual engineering. We subsequently discuss these cases to make clear what methodological choices are, and should be, made when doing this kind of conceptual work. We have two main goals. We first want to contribute to the literature on conceptual engineering by presenting concrete examples of conceptual engineering in the philosophy of technology. This is especially relevant, because the technologies that are designed based on the conceptual work done by philosophers of technology have crucial moral and social implications. Secondly, we want to make explicit what methodological choices are made when doing this conceptual work. Making explicit what methodology we are employing allows for a conscious reflection on this methodology. Ultimately, our hope is that conscious reflection leads to an improvement of the used methods.

Keywords Technology, conceptual engineering, moral conflict, freedom, critical thinking, control

1. Introduction

New technologies are the source of puzzlement and considerable moral uncertainty. How should we think about the technological issues, which principles apply, which values are salient, what do we owe to those affected by the negative consequences of our innovations? In addition to the problems about deciding which ethical principles and theories apply, there is a problem of how to conceptualize the parts of the world that we are dealing with. The introduction of the mechanical ventilator and the early pregnancy test have, for example, caused strong disagreement about the application

[21] Published as Veluwenkamp, H., Capasso, M., Maas, J., & Marin, L. (2022). Technology as driver for morally motivated conceptual engineering. *Philosophy & Technology*, 35(3), 71. For readability I have moved all references of this and the following chapters to the bibliography section at the end of this dissertation. Besides this, I have made no changes to the chapters.

conditions of what philosophers term "thick descriptive concepts" such as DEATH and PREGNANCY (Baker 2019; Leavitt 2006). These technology-induced uncertainties about the applicability of moral and morally connotated concepts are becoming very prominent. A Jewish student once asked Richard Feynman whether electricity is fire. That is a religiously and morally relevant question, since religious Jewish laws prohibit 'making fire' on Sabbath. Turning on the light could be forbidden depending on whether electricity is considered fire or not. The worldwide adoption of CRISPR-CAS genetic engineering techniques has given rise to fierce moral and legal debates in Europe about what counts as a 'natural way' of altering the genome. Can one have a thousand friends online, or do we conclude that those who claim such an impressive circle of friends must have a radically different conception of friendship? Blockchain and quantum encryption can mathematically guarantee that parties will comply with norms, promises and contracts. Now we can make our interaction 'failsafe', have we therefore established *trust* between parties? Many of the problems we have with technology are of this type and they call for conceptual engineering. In this paper, we analyse examples of how modern technology calls for conceptual engineering. Conceptual engineering, however, is a field that has only recently received considerable attention, and many important methodological questions remain open (Eklund 2020). We use our analysis of technology-induced examples of conceptual engineering to see how these methodological questions are answered in the technological domain.

Conceptual engineering is the design and implementation of concepts. This practice wouldn't be justified if we already employed the best possible concepts[22], but there are multiple reasons for doubting that this is so. The analytical tradition in philosophy is based on the idea that we should not just make do with the concepts we have been handed. It is shown - sometimes in great detail - how our language leads us astray, into paradoxes, puzzles, contradictions, absurd conclusions, or plain nonsense. Even if our preferred frameworks of philosophical concepts are generally fit for purpose (as explanatory tools), we still might find concepts or even subsets of concepts that are defective in some ways. A concept can be defective in different ways. Herman Cappelen (2018) identifies two major variants of conceptual deficiency. Firstly, the semantic value of a concept can be defective. Early emotivists, for example, held that our normative concepts are meaningless, and should therefore be revised or even abandoned. But the inconsistency of a concept is also used as a reason for revising the

[22] The "best possible concept" should be understood as the all-things-considered best possible concept. See for a related discussion Eklund (2012; 2017)

concept. *Truth*, *Freedom*, *Knowledge* and *Race* have all been called inconsistent or incoherent, and alternative conceptions have been proposed. The second kind of conceptual deficiency Cappelen identifies is that the concept is morally, politically or socially problematic. The concept *Marriage*, for example, is not semantically defective. However, if it excludes same-sex couples, then that might have objectionable moral, political and social consequences. One of the consequences that motivated a conceptual change in this case, is that people could come to see same-sex relationships as inferior. Sally Haslanger's ameliorative projects are also explicitly politically motivated. Her stated goal is the elimination of what she calls women. She believes that "it is part of the project of feminism to bring about a day when there are no more women (though, of course, we should not aim to do away with females!)" (Haslanger 2000, 46).

The kind of deficiency we are interested in is of the second kind. In the case of semi-autonomous cars, we see that the application of our old conceptions of control and responsibility causes moral uncertainty. It is not clear whether they introduce responsibility gaps (Matthias 2004; Sparrow 2007) and how this affects public policy options regarding transportation. Moreover, even if they do not introduce responsibility gaps, it is not clear what the proper distribution of responsibility should be, as Google and Uber's different reactions to a semi-autonomous car crash show (Nyholm 2018; Hindriks & Veluwenkamp ms). In this context it is not the case that there is moral uncertainty because we do not know which moral theory to apply - as Nickel et al. (forthcoming) had pointed out. It is also not the case that the application conditions of a thick descriptive concept have changed. Instead, we see a conflict between the technology, our moral judgements and the relevant moral concepts. One way of resolving this conflict is by proposing new or revised moral concepts or conceptions.

We see that this kind of morally motivated conceptual engineering steadily increases in importance. That this work is a kind of conceptual engineering is, however, not often recognized. We take this to be a missed opportunity, as a recognition that different people are working on the same kind of project can help solve methodological questions that one is likely to encounter. In this paper we have two main goals. We first want to contribute to the literature on conceptual engineering by presenting concrete examples of conceptual engineering in the philosophy of technology. This is especially relevant, because the technologies that are designed based on the conceptual work done by philosophers of technology are morally and socially significant. Secondly, we want to make explicit what methodological choices

are made when doing this conceptual work. Making explicit what methodology we are employing allows for a conscious reflection on this methodology. Ultimately, our hope is that conscious reflection leads to an improvement of the used methods.

To accomplish this all we proceed as follows. We first provide a framework for understanding conceptual work done by philosophers of technology as conceptual engineering (section 2). We then present three use-cases through the lens of this framework (section 3). Finally, we discuss the use-cases to make clear what methodological choices are made (section 4).

2. Conceptual engineering

In the following sections we introduce three case-studies to distinguish four methodological conclusions about morally motivated conceptual engineering. However, to make these distinctions we first have to introduce some terminology. We will distinguish between a concept and its conception(s). Conceptions stand to concepts in a one to many relation. That is, different conceptions can be *of* a concept. In his *A Theory of Justice* (1999, 5), John Rawls introduces this distinction as follows:

> Men disagree about which principles should define the basic terms of their association. Yet we may still say, despite this disagreement, that they each have a conception of justice. That is, they understand the need for, and they are prepared to affirm, a characteristic set of principles for assigning basic rights and duties and for determining what they take to be the proper distribution of the benefits and burdens of social cooperation. Thus it seems natural to think of the concept of justice as distinct from the various conceptions of justice and as being specified by the role which these different sets of principles, these different conceptions, have in common.

What Rawls seems to have in mind here, is that there is a specific role or function (providing "principles for assigning rights and duties") and that different conceptions perform this function to a certain extent. The relevant question now becomes which conception performs this function best, given the context under consideration. And, of course, Rawls famously argued that in the context of western liberal democracies, justice as fairness is the best conception of justice.

In line with this take on the concept/conception distinction, we propose to understand concepts and conceptions as entities that have meanings as their content.[23]

[23] Philosophers of language typically distinguish between representationalist and inferentialist theories of meanings. Representationalists define meaning in terms of what a conception purports to

The meanings of concepts are in some sense indeterminate, and there are different ways of making a concept precise. If we make a concept precise without a change of topic, then we have a conception of that concept. *Justice as Fairness* can now be understood as one way of making *Justice* precise. Not all ways of making a concept more precise will be without a change of topic. If we propose, for example, a conception which identifies justice with being a blue sky, then we will no longer be talking about our old concept of justice anymore. Two conceptions are of the same concept if they are similar enough. In some cases, two conceptions are similar enough if they play the same role, as the above interpretation of Rawls suggests. However, to be maximally inclusive in our definition of the relation between concepts and conceptions, we leave open the possibility that other factors determine what counts as "similar enough" (see e.g. Lalumera 2013 for different interpretations of the concept/conception relation).

We understand the practice of conceptual engineering as trying to find out what the correct conception of a concept is in a specific context. It is important to note that when we say "correct", we don't mean to ask which conception we are currently using. That is, we are not engaging in conceptual analysis. Instead, we are asking the question which conception we ought to use: which conception is best. This distinction is important, because often conceptual change is called for, exactly because there is something wrong with the conception we are currently using.

Given this way of understanding conceptual engineering in the philosophy of technology, there are several open questions. First there is a methodological question: what is it that makes a conception best? In the literature about conceptual engineering there are two main candidates: a metaphysical and a pragmatic approach (Thomasson 2020). It is, however, not evident which method is most suited for philosophers of technology. A second, prominent question when engaging in conceptual engineering is whether we are changing the topic when a new conception is proposed. This is an accusation that is sometimes made, and it is unclear what the best response to such a challenge is. The third question relates two different views about conceptions that are

---

represent. So if two conceptions purport to represent the same things, then they are identical (e.g., an equilateral triangle and an equiangular triangle). Inferentialists, on the other hand, individuate conceptions in terms of conceptual or inferential role (which can be spelled out in terms of what one is disposed of or in terms of what one is normatively committed to infer). The fundamental distinction between those theories is one of explanatory priority, as most representationalists hold that conceptions have an inferential role and inferentialists maintain that conceptions purport to represent things. We are most attracted to inferentialist theories of meaning, but for this paper we opt to remain neutral on this issue.

sometimes defended: conception relativism and conception imperialism. The relativists hold that no conception is better than another, while the imperialist holds that if a conception is the best conception of a concept in one context, it is also best in other contexts. What we are interested in is if these views are reflected in the conceptual work that is done by philosophers of technology. The final question we will assess is whether we always have to design new concepts, or whether there are other options. Before discussing how these questions are answered by philosophers of technology, however, let us present the three use-cases.

3. Cases

In this section we explore three use-cases that exemplify how technology introduces new contexts in which our old conceptions conflict with our moral judgements. The cases have been selected because the technology and the concepts involved have attracted much media attention and are currently still the topic of powerful moral debates.

3.1. Control

3.1.1. 'Old' conception

It is philosophical orthodoxy to define moral responsibility in terms of an epistemic and a control condition (Rudy-Hiller, 2018). That is, a moral agent is typically taken to be responsible for an action if and only if that agent is adequately aware of the consequences of performing the action and possesses a sufficient degree of control over the action. The conception of control has been, and still is, debated with regards to its connection to determinism: "is control possible if the world is governed by deterministic rules?". Apart from this fundamental philosophical question, it has been thought that the conception of control is relatively unproblematic.[24] The engineering conception of control says that as long as one is physically capable of intervening in a system's operation, one has operational control over that system.[25]

---

[24] But see (Fischer & Ravizza 1998; Himmelreich 2019) for a philosophical discussion of control.

[25] But note that also in mining and building destruction we talk about 'controlled explosions'. We think we have control by making sure that an end result is achieved, but once the process is underway we cannot possibly intervene.

3.1.2. Conflict with moral judgements

Recent research in ethics of technology has insisted, however, that traditional common-sense and engineering conceptions of control should be problematized in the light of developments in digital technologies. One of the reasons for this is that the traditional engineering conception of control leads to verdicts that conflict with our moral judgements about semi-autonomous vehicles. For example, suppose we consider someone behind in the driver seat of a car which has adaptive cruise control switched on. As long as this person is physically able to intervene in the operation of the car, this person has operational control and is therefore potentially responsible when something goes wrong. It also leads to the claim that fully autonomous vehicles should be forbidden. The motivation for this assertion is that according to the old conception of control there is no moral agent for whom the control condition is satisfied. Such a scenario has been called a responsibility gap: a situation where no one can be blamed for the harms of an autonomous system (Matthias, 2004).

Digital platforms are another context in which we see that the old conception of control conflicts with our moral judgements. In the documentary "The Social Dilemma" we see on full display the harm that social media companies such as Facebook and Twitter in fact do to their users. The makers of this documentary, however, seem reluctant to blame the creators of the algorithms that these companies use. Instead, these developers are the protagonists of the documentary, explicitly portrayed as ordinary people caught in a bigger game, who make nervous small talk before the actual filming starts. *They* are not in control of the algorithms! So, given the popular idea that responsibility requires control, one might think that the developers are not to blame for the harm done. However, it is also implausible to hold the users, or the algorithms themselves, responsible. Thus, one might think that a responsibility gap arises here as well. This may fuel the feeling of a technological and economic determinism, in which impersonal technological and economic forces govern society and ultimately our lives, so we may just sit on our sofa and, so to speak, enjoy the show as much as we can.

3.1.3. 'New' conception

A first response to this conflict is the observation that decades of philosophy and sociology of technology have shown that technology is not determined by mysterious inscrutable forces, although there may be path dependencies. Despite their complexity, machine-learning systems often serve the interests of the companies that

developed them quite well. So companies are to blame? Perhaps, but this is not the point. Instead, the goal should be to achieve a more acceptable distribution of control and responsibility in our current socio-technical landscape. Responsibility gaps may well arise in complex socio-technical systems like the social media companies, and they are a problem insofar as we know that if all stakeholders feel they have an excuse if something goes wrong, nobody has an incentive to avoid these situations. But, again, responsibility gaps are not a destiny. If they are clearly on the horizon, accompanied with a technological project, then there is a forward-looking responsibility to prevent their emergence or mitigate their effects. There are various societal stakeholders who could be addressed with such responsibility, such as designers, software developers, engineers and regulators.

"Meaningful human control" (a term invented in the political debate on autonomous weapon systems) on the contrary requires more than the traditional conception of control. It requires that our interaction with the technology is designed to give us a fair capacity and opportunity to have our deepest values and interests reflected in the behaviour of the technology. And that nobody is in the position to control us, by selling us the illusion of an easy, superficial, and of course fake, "control".

A promising framework for meaningful human control is developed in Santoni de Sio and Van den Hoven (2018). This new conception of control is supposed to replace the old, traditional conception in contexts where artificial autonomous systems are causally, but not morally, responsible for outcomes. What distinguishes this account from older conceptions of control is (a) that for meaningful human control it is necessary nor sufficient for an agent to be able to causally intervene in a system, (b) that it is sensitive to the epistemic conditions necessary to have the kind of control to render an agent responsible, (c) that it applies to the entire "socio-technical system" (and not just to intelligent artefacts) and (d) that it is meant to provide general design guidelines to achieve the required kind of control.

Santoni de Sio and Van den Hoven present two necessary conditions; a tracking and a tracing condition. The tracking condition requires a socio-technical system to be responsive to the relevant normative reasons to act. The tracing condition requires that one or more human agents are present in the system design history or use context who appreciate the capabilities of the system and their own responsibility for the system's behaviour. Because the tracking condition requires a system to be responsive to the reasons of the relevant agents, it does not require that an agent causally influences the system's actions. So if properly implemented, responsibility gaps can be

avoided. This conception also avoids the other conflict with our moral judgements. If an agent is merely 'in the loop', that is, is able to causally intervene, then this does not entail that she has meaningful human control over a system. The reason for this is that meaningful human control requires that a system is responsive to the agent's reasons. And, if one is able to causally intervene but doesn't have the knowledge required to intervene appropriately, the system will still not be responsive to the agent's reason.

3.2. Critical thinking

3.2.1. 'Old' conception

Critical thinking is a form of goal-oriented thinking (Hitchcock 2018), a process meant to arrive at a practical decision about "what to believe or do" (Ennis 1962). There are two ways of understanding critical thinking: as a descriptive and as a normative term. The descriptive notion of CT captures the kind of thinking which is not routine and contrasts it to more automatic modes of reasoning such as a logical deduction, the application of a rule to a class, etc. However, this is not what interests us here, as we will focus on the normative conception of CT since we will show that this conception is put under strain by technological change. The normative conception of CT became popular in the latter half of the 20th century, in the aftermath of the Second World War, when educational researchers and philosophers asked themselves: what can be done so that these kinds of dictatorships will not happen again in the future? It was claimed that if the majority of the population would be endowed with critical thinking skills, then anti-democratic regimes would be much harder to instate since critical citizens would see right through attempts of manipulation, propaganda and populism which are the signature moves of authoritarian regimes (Stanley 2015). In this political context, education for critical thinking was aimed at instilling the skills and virtues that allowed students to spot bullshit (Frankfurt 2009) and manipulation in mass-media texts or in political speeches. The value of CT was to support the exercise of democratic virtues by the citizens by keeping them informed and helping them avoid the pitfalls of anti-democratic discourses. Normative CT has thus been identified with thinking in the service of democratic goals, or thinking that leads to an informed citizenry by helping individuals take informed actions and making informed choices.

The distinguishing feature of critical thinking from other kinds of thinking is its difficulty: critical thinking is not a routine cognitive operation, such as recalling

information or applying a routine procedure for solving a problem. Rather it is deliberate and entails some cognitive effort, it "only occurs when the reasoning, interpretation or evaluation is challenging and non-routine" (Fisher 2019, 29–30) for the epistemic agent doing this thinking. Because of this effortfulness of CT, epistemic agents need to decide for themselves under what circumstances it is worthwhile to launch into a process of CT. The 'classical' concept of CT stresses the autonomy of the thinker, since CT is about thinking for oneself and refusing to delegate one's epistemic authority to authorities. From this perspective, CT could be dangerous (one can only think about conspiracy theorists who also 'think for themselves' and systematically distrust public sources of epistemic authority such as scientific institutions) if taken too far, and this is safeguarded by ensuring that CT is described not only as intellectual autonomy, but as a cluster of epistemic virtues such as "curiosity, open-mindedness, attentiveness, intellectual carefulness, intellectual courage, intellectual rigour, and intellectual honesty" (Baehr, 2013, p. 248). Since any of these intellectual virtues are laudable having in themselves, the distinctiveness of CT lies in the claim that having these particular virtues will determine the epistemic agent to launch into a relentless pursuit of the truth of the issue at stake, regardless of the personal costs in terms of effort and social costs (such as becoming an outcast because one asks the uncomfortable questions) and this will lead, ultimately, to achieving of the epistemic goals of a democracy. We should notice that the way classical CT is described is in terms of properties of the individual epistemic agent (virtues, skills, dispositions) and assuming that one can launch in a process of CT regardless of how hostile an environment is.

3.2.2. Conflict with Moral Judgements

When social media platforms gained popularity, certain behaviours emerged as deeply problematic: regular users engaged in sharing misinformation, started taking sides in polarising debates while ignoring the nuances and demonising the opposing side, and some users even self-radicalised after watching increasingly aggressive videos suggested by a platform's algorithm (see Alfano et al. 2018). It is reasonable to believe that some of these behaviours were taking place before social media (see gossip spreading as a model for misinformation sharing), yet the new platforms of social networking made these behaviours highly visible and also more toxic. One's intellectual vices of laziness and carelessness in assessing information suddenly had rippling consequences since any user could share misinformation or polarising opinions and it became impossible to predict who would see this misinformation and

be affected by it. Social media platforms were depicted as a threat to democracy (Sunstein 2018) because of how easy it became to manipulate users and sway them simply by truncating the information that reached them.

To tackle the epistemic threats emerging from social media platforms, media scholars advanced the idea that critical thinking should be promoted for users by stepping up formal education or by nudging them to think critically while being online (Blair 2019). The conception used here was the old conception of critical thinking as a cluster of individual character traits of the epistemic agents. However, this old conception did not solve the problems posed by the new informational environment that was social media. For one thing, social media platforms are an ambiguous epistemic environment where context is easily collapsed (Marwick & boyd 2011); this means that users cannot reasonably decide when they should be thinking critically about a post, and when they can simply ignore it. The epistemic ambiguity of social media means that a statement posted by someone on social media can be genuinely informational for some users, while for others it counts as noise, in an unpredictable way.

While the old conception of critical thinking was about an individual's features (a set of cognitive skills) that one carried throughout one's life, as if one owned the critical thinking skills, social media platforms challenged this assumption and showed instead that the environment in which we act as epistemic agents matters just as much as our pre-existing skills and dispositions. With social media platforms oriented towards personalisation of user-experience, we become trapped in a filter bubble (Pariser 2011) whereby we only see things that we agree with and we are not confronted with opposite opinions. This filter bubble is invisibly created by algorithms which deliver what they estimate we may want to see, with the purpose of making us more engaged on the platform. Social media platforms achieve an addictive effect by signalling to users that they are right in every aspect: the music they like, the activities they choose and, when they care about such things, their political views are shown as the best in the world. In this context of continuous affirmation of the self, it is very hard to deliberately put in the effort to think for oneself, to look for counter-evidence or to even become aware that one has cognitive biases. When users are immersed in social media environments where information is overwhelming them through its sheer quantity while also highly personalised, enacting the classical dispositions of critical thinking such as self-restraint, curiosity about other points of view, and intellectual humility becomes increasingly difficult. Social media thus enacted a conflict between our epistemic values needed for safeguarding democracy and the conception of CT

as an effortful individual process. Solving the conflict would entail us commiting to the idea that all social media users need to be critical about all information they see on social media, since we cannot predict which piece of information would be toxic for democracy. This is not feasible in practice and it entails trivialising the concept of CT which was designed in the first place to be a mode of thinking which was not to be deployed in everyday circumstances.

3.2.3. 'New' conception

Recent work in cognitive sciences has made popular the notions of nudging and boosting (Thaler and Sunstein 2008; Gigrenzer 2018). The idea behind nudging and boosting is that we do not need to reflect ourselves all the time in order to make rational choices if we can delegate to our environment some of these choices. Nudging is about designing environments which promote our values without us having to make these choices deliberately (think about a cafeteria line where the healthy foods are placed at eye level and easier to reach) while boosting is about creating learning environments where users learn to take the most rational decisions for them in an effortless manner (thus relying less on deliberate reflection and more on intuitions). Nudging and boosting show that we need to take some of the pressure off the individual users and look at the design of our cognitive environments. Social media platforms appear as cognitive environments tailored toward entertainment and confirmation of pre-existing biases, and as such will not incentivise critical thinkers to apply their skills (also see Williams 2018 on distraction by design and Voinea et al. 2020 on the cognitive detrimental effects). This does not mean that critical thinking is impossible online, but that it becomes an uphill battle that most of us do not know we are facing when we open our social media platform of choice. Furthermore, this difficulty of engaging in cognitive thinking cannot be attributed to user weakness of will or lack of education, but instead highlights the role that the cognitive environment plays in exercising our capacities for critical thinking, namely the key role played by dispositions.

Our old conception of CT appears as incapable of facing the novelty of social media as epistemic environments and needs to be revised. The new conception of critical thinking needs to accommodate the understanding that the environment plays a role in how well we think - critically or not - and this needs to be part and parcel of the new concept of critical thinker. The new conception of critical thinking designates the unity of thinker and the cognitive environment, taken together, by acknowledging that there are no free floating dispositions to be critical, these are triggered by friendly

environments or stifled by hostile ones. Such critical-friendly environments could be designed from the beginning following some design principles such as increasing user friction and diversifying the user's informational diet with information sources, and generally avoiding to personalise the user's experience. Thus, the critical online environment would be targeted not towards maximising engagement or entertainment, but towards fostering reflection and self-knowledge in an ecosystem of human-technology interactions. There are already experiments in design showing how nudging towards online critical thinking can be achieved by changing the environment. In these experiments users were primed to themselves if a certain piece of news was misinformation (Lutzke et al. 2019) and, after a few iterations, the users took this new habit with them onto other platforms. The new conception of critical thinking recognises that the thinker is embedded in an environment which needs to be designed for criticality in a deliberate manner, either through nudging or boosting.

3.3. Freedom

3.3.1. 'Old' conception

Freedom is a fundamental concern for most normative political theories. A crucial point of reference for the philosophical debates on social and political freedom is Isaiah Berlin's distinction between a negative and a positive interpretation of freedom. The negative conception refers to the actual absence of relevant interferences or constraints on one's actions, while the positive conception concerns self-realization and self-determination (Berlin 1969).

More traditional political theories, such as liberalism and libertarianism, endorse a negative definition of freedom. These theories concern themselves primarily with the external sphere of action of individuals and claim that individuals should not be unduly interfered with by the State or other actors or bodies. Under this traditional conception of freedom, one's freedom is restricted when there is interference, understood as an intentional and actual intervention by other people that restricts the number and quality of the set of options or choices before agents, or as Berlin put it, that affects what doors and how many doors are open to the agents (Berlin 1969: xlviii).

3.3.2. Conflict with Moral Judgements

Currently, we see increased concerns regarding the power of Big Tech amongst several groups in society. Scholars have increasingly devoted their attention to the effects of big data on democracy (e.g., Zuboff 2019; Nemitz 2018; Macnish & Galliott 2020), even wondering whether democracy will survive this new trend (e.g., Helbing et al. 2019). In addition, governmental bodies are becoming more active in controlling Big Tech (Ovide 2021). This increased attention suggests that traditional freedom as non-interference can be conceptually and morally deficient if confronted with the kind of power and influences that are exerted via new emerging technologies, such as machine learning or Big Data. Such technologies pose new risks that have led scholars to talk about hypernudges, which are Big Data nudges ("big nudging") that can shape people's behaviours and their choice context in a more efficacious, targeted and pervasive way through the extraction and collection of their data (Yeung 2017).

Consider the Facebook-Cambridge Analytica scandal, where millions of people's data was unknowingly gathered and used for political profiling. The public uproar that resulted from this scandal was partly due to privacy violations and a feeling that millions of users were manipulated. The liberal conception of freedom defended by 'pure negative theorists' often implies an exclusive emphasis on interferences conceived as options-removals, or any conducts or dispositions of some other persons that prevent an action, rendering such action impossible to perform (Carter 1999; Kramer 2003; Carter and Kramer 2008). However, our moral judgements about this scandal and its public consequences are not adequately exhausted by such an account of freedom. Indeed, no removals of options or conducts that prevented individuals' actions occurred prima facie. Nonetheless, such a scandal has shown a peculiar restriction on freedom that stands in need of normative justification and appraisal. There are different cases of interferences or constraints in social life - coercive or otherwise, such as manipulative ones, physical or psychological, actual or not, and so on - and the stakes of conceptually individuating and normatively justifying them are high.

In addition, filter bubbles decrease the possibility for individuals to be exposed to debates outside their own preference (Pariser 2011). Mill (2003, specifically ch. 2) already argued for the importance of being exposed to other points of view outside your personal realm. Conflicts of opinions enhance democracy, but mostly are important to exercise one's freedom and autonomy. The traditional liberal conception does not adequately account for this since less coercive restrictions are not immediately considered freedom-restrictive, yet with the increased phenomenon of

filter bubbles in combination with Mill's idea of 'conflict' exposure, this conception no longer seems to hold. However, the effects of new technological influences are not merely visible on an individual level; democratic harm is equally done on a collective level (Macnish & Galliott 2020). The conceptual insufficiency of freedom as non-interference in this case is arguably motivated by the fact that such conception tends to neglect interpersonal relations and social standing and to place its predominant focus on the set of options and choices an individual enjoys.

Furthermore, what we see is that there is no satisfying possibility to check the power of big tech companies. Scholars such as political scientist Francis Fukuyama draw attention to this intuitive concern (Fukuyama, Richman & Goel 2021). Regarding social media platforms, these authors point out that the real issue relates to the question of who is in charge. Twitter may suppress and fact-check Trump (now even having banned Trump), but who is suppressed depends ultimately on the person in charge, less so on the justice of the suppression. Fukuyama et al. (2021) compare the power of these companies with a loaded gun on a table: right now, nobody picks it up to shoot, but we may wonder to what extent it is safe to leave it there. The authors correctly refer to the necessity of checks and balances within a liberal democracy. The gun example illustrates that currently, these companies can exercise their power without being controlled by a body who oversees their power and decisions.

This lack of checks and balances suggests a deeper concern related to interference and non-interference: is there robust non-interference? Consider again the gun example: even if these companies do not interfere with you, the fact that they can if they wish implies an unequal political relation between the companies and society in general. Indeed, the huge amount of data these companies have on a person makes it possible to single them out if they wish: you only have 'nothing to hide' if you are not explicitly on their radar. Since there are currently no adequate checks and balances mechanisms, these companies have the possibility to exercise their power unaccountably. This unaccountable exercise of power implies that freedom does not depend on interference or non-interference, as traditional liberals believe, but perhaps more so on whether there is an insurance for non-interference, in other words, whether the interference is in fact robust.

Robust non-interference depends on controlling entities. An adequate checks and balances mechanism requires not just providing control after the event; it includes providing checks on these new influences beforehand. The vast increase in technology companies simply exceeded legal regulations and public scrutiny. Only recently do

these companies face increased pressures from the law, governments and the public that suggest a change in relation between these companies and society.

3.3.3. 'New' conception

The intuition regarding the problems with potential interference matches a recent scholarly discussion on freedom which has been motivated by a desire to overcome the distinction between a negative and a positive interpretation of the concept. In particular, among those theorists there are some that have identified and promoted a view of freedom as "republican" freedom. Contemporary neo-republican political theorists such as Philip Pettit or Quentin Skinner have attempted to go beyond negative freedom, understood as the absence of interference, and have individuated the conceptual core of freedom in "non-domination" (Pettit 1997; Skinner 1998; 2002). Freedom as non-domination is a status, namely the enjoyment of a position that guarantees that no interference from arbitrary kinds of power is exerted (Pettit 2011; 2012). A status like that of a slave makes him susceptible to being interfered with by a master, independently of any actual interference from the latter. Therefore, a politically worthy society is the one that maximizes in its institutions and mechanisms such a conception of freedom according to neo-republicans.

Interestingly, this neo-republican framework is precisely what Fukuyama et al. seem to describe. Indeed, it is not the actual interference that worries the neo-republican, it is the potential of interference when there are no mechanisms available to hold the power accountable. Although not expressed in these terms, Fukuyama et al. describe the concept of domination regarding tech companies and society. Their possibility to interfere without having to face consequences for their actions defines domination as understood by neo-republicans: being subjected to a superior and unaccountable power. This debate sparked in the literature by neo-republicans and the situation as we encounter in today's reality suggest that the old conception of freedom as non-interference does not provide sufficient explanation why we should worry about the power of these companies. An existing conception of freedom such as the one endorsed by neo-republicans can be a better candidate to frame and advance the account of freedom in the context of new emerging technologies. Indeed, it shows that political and social freedom is not about the absence of actual interference or about the doors that are open to individuals, but rather it requires that no doorkeeper has the power to close or conceal a door without a cost (Pettit 2011: 709). In terms of our moral intuition about facts, this means that the main concern

with these companies is that they have the power to interfere with their users without being held (adequately) accountable (Maas 2020).

4. Discussion

As we have seen, many philosophers of technology engage in some form of morally motivated conceptual engineering. To provide a structure to these different projects we distinguish four lessons-learned for doing conceptual work in the philosophy of technology.

4.1. Approaching conceptual engineering

There currently is no consensus on the methodology of conceptual engineering. Matti Eklund, for example, calls the question what the proper methodology for conceptual engineering is one of the "big questions [that] remain[s] entirely unresolved" (2015, p. 382). In section 2 we indicated that there are two main approaches to conceptual engineering: a metaphysical and a pragmatic approach. Proponents of the metaphysical approach hold that we should use those concepts that fit best with metaphysical reality: i.e., those concepts that carve nature at its joints. One of the advantages of this approach is that it fits well with our intuitions about scientific and other empirical concepts.

On the pragmatic approach, when deciding which conception of a concept to employ, we should first determine what function, or purpose, this concept should perform in the context that we are discussing. Once we have determined what the function is, the best conception is that conception that fulfills this function best. Sometimes the function that a concept ought to perform is the function that it has at the moment. Suppose for example that one had in one's society a conception of marriage that precludes same-sex couples. Moreover, suppose, as is plausible, that the function of marriage is to afford a special legal and social status to a range of close relationships (Cappelen 2018). In these circumstances we can come to see a conception of marriage with includes same-sex couples as better than the old one.

From the examples that have been discussed in the previous section we see that the pragmatic approach is favoured. The conceptions that are the target of the engineering work all relate, or ought to relate, to a specifically moral concept. For example, when discussing which conception of Control to use, we started out by pointing out that we are interested in a specific function of that concept: i.e. its relation to responsibility. The question which conception is best, was also addressed using this

heuristic. To answer this question, we looked at an example in the context; a person behind the wheel of a self-driving car that has adaptive cruise control switched on. Given some additional details, we can now truly say that the driver has causal control over the car, but lacks meaningful human control. So we cannot say that one description is true while the other isn't. When deciding which conception is best in the context of autonomous intelligent systems, we based our decision on the question which conception relates best to appropriate responsibility attribution. The reason for this, is that this is the conception that fulfils the context-determined function best.

The same can be said for *Freedom* and *Critical Thinking*. When we assess which conception of *Critical Thinking* is best, we do think with a specific goal in mind: making sure that the kinds of dictatorships that we have seen in the past will not happen again, hence that democratic values are supported by informed citizens. Given this role, it was argued that the new conception, which includes critical environments as a part of the new concept, was able to fulfil the normative constraints that the old one could not. In section 3.3 about *Freedom*, the function that was highlighted was to indicate a desirable relation to those that are powerful. It was argued that in the context of influential social media companies, the neo-republican conception of freedom does a better job indicating which relation is desirable than the negative conception of freedom. Whether the conception is indeed better in fulfilling a specific function is a matter of normative debate. And this is of course exactly what the pragmatic approach predicts (Thomasson 2020, 451-5).

4.2. Are we changing the topic?

Some philosophers are critical of the practice of conceptual engineering in general, because they think that concept revision is always just changing the topic. The risk is that "[r]evisionary projects are […] providing answers to questions that weren't being asked" (Haslanger 2012, 225). Indeed, sometimes conceptual engineering really does change the topic. An example of this approach is the introduction of the concept of "online harassment" that questions our traditional understandings of moral wrongs, intentionality and evil (Cocking & Van den Hoven 2018). A different example is Miranda Fricker's work on "epistemic injustice" (2007). Fricker recognized that we lacked the conceptual tools to represent a certain kind of injustice: the injustice which consists in a wrong done to someone specifically in their capacity as a knower. Introducing this new concept made it possible to discuss different instances of this injustice under one heading. But we do not always need to introduce a new concept. Sometimes, we remain on the same topic (i.e. concept) and propose a new answer to

an old question. These scenarios generate a new challenge: how do we distinguish between scenarios where we change the topic and scenarios where we do not?

Let us look at one of the use-cases above to answer this question. Proponents of Meaningful Human Control are also sometimes accused of changing the topic. What their critics maintain is that they are no longer talking about control, but about something else. The underlying accusation is that what the MHC proponents discuss is not relevant to the original discussion. And indeed, if we are primarily interested in our current conception of Control, then an analysis of meaningful human control might not be relevant. A better way to understand the argument of proponents of MHC is twofold. First, they argue that in the current context we should be interested in those conceptions that are suitably related to attributions of responsibility. And secondly, they argue that with regards to autonomous intelligent systems, MHC performs this function best. Remember that we stated in the introduction that the question of what counts as remaining on topic depends on context-sensitive similarity relations. The first argument is supposed to fix the similarity relation, and thereby the set of conceptions that are "on topic". Given this set of conceptions, the second argument now indicates which of these conceptions is best.

4.3. Do we aim for a purpose- or context-specific revision, or is the revision supposed to be global?

In section 2 we have presented two views on conceptions that are sometimes defended in the literature: conception relativism and conception imperialism. The conception relativist holds that no coherent conception is better than another one and the conception imperialist maintains that only one conception of a concept is correct in all contexts. We see that in all projects we discussed, conception relativism was implicitly rejected. A conception can be completely coherent and still be deficient in different ways. All examples made clear that conceptions can be morally deficient in specific contexts. We have, for example, shown that traditional common-sense and engineering conceptions of control should be problematized in the light of developments in digital technologies.

These examples also show, however, that conception imperialism is at least sometimes rejected. If conception imperialism is correct, then revisions are always global. That is, if a conception is best in one context then it is ipso facto the best in all contexts. MEANINGFUL HUMAN CONTROL is, for example, a conception of CONTROL that is explicitly supposed to be local. The tracking and the tracing conditions refer to design-histories and socio-technical systems, and are therefore not suited for many

other contexts (e.g., when discussing the question whether a human has control over her own bodily movements). A conception imperialist who wants to defend this conception of meaningful human control would therefore have to deny that 'regular' control and 'meaningful human control' deal with the same topic. But as we have seen above, this is highly implausible given the purposes that we have.

From the limited number of cases that we present in this paper, there is only one conception that can arguably be said to be a global conceptual revision: **F**REEDOM. The neo-republican conception of freedom has been applied in other contexts as well[26], and a case can be made that this conception is the only correct one in all contexts. What is important, however, is that one is not committed to this position. It is even possible, and not even farfetched, to reject Skinner's and Pettit's proposal for the neo-republican conception of freedom in the domains that they are interested in, and hold that this conception of freedom is superior in the context of digital technologies. It is important to note that maintaining that the neo-republican conception of freedom is appropriate for all contexts in which *Freedom* is used, does not commit one to conception imperialism. The claim that for a specific concept there is only one appropriate conception is compatible with the claim that there are other concepts that have different appropriate conceptions in different contexts.

4.4. Is there already a candidate conception available, or should we construct a new conception?

The fourth issue that is relevant is whether the conception that is proposed is new, or if a conception that is used in a different context can be used for the new context as well. In that case we might consider a variety of contexts, e.g. disciplinary context, application context, or historical context. John Rawls, for example, argued for justice as fairness in the context of liberal societies. The aim of his revision was narrow, i.e., only supposed to apply to a specific context. However, this doesn't rule out the possibility of applying this conception of justice appropriately in a completely different context. As we have seen in section 3.3, for example, the neo-republican conception of freedom can be used in contexts that were not envisioned when this conception was introduced. On other occasions the conception we want to propose is an entirely new one (e.g., the original introduction of justice as fairness). The fourth issue is therefore: *is there already a candidate conception available elsewhere, or should we construct a new conception?*

[26] E.g., example medical care (O 'Shea 2018), workplace (Anderson 2017), immigration (Costa 2015).

As section 3.3 has shown, reviving older or less popular theories help us address some of the moral judgements we have with the new technological influences and society. Originally, freedom from domination has been used in particular to address institutional arrangements, like the relation between the State and her citizens. This conception of *Freedom*, however, may be used in new settings, such as the relation between companies and their users.

In *Figure 1*, we illustrate this question in an (oversimplified) diagram: one concept (freedom) includes several conceptions (freedom as non-interference, and freedom as non-domination) that correspond to a particular context. Where initially the conception 'freedom as non-interference' would relate to the context 'social media', section 3.3 illustrates how this traditional conception conflicts with our moral judgements. The question then is whether we require a new conception of freedom or an already existing alternative. As argued in 3.3, the already existing alternative 'freedom as non-domination' is a good fit in this context and hence there is no need for developing a new conception of freedom to accommodate our moral judgement. The red cross in *Figure 1* reflects our conflict and the green arrow our moral judgement. For the concept of freedom in the context of social media, the answer to the question whether we should construct (1) a new conception or (2) is there already a candidate conception available elsewhere therefore concludes: there is already a good alternative available, namely freedom as non-domination.

We see that by expanding this already existing conception to a new realm - namely, ethics and philosophy of technology, we can better formulate what needs to be done to address concerns with these new technologies. For instance, for neo-republicans it is important that citizens have the opportunity to *contest* governmental decisions, as this provides a checks and balances system (Pettit 1997). With Facebook-Cambridge Analytica, people were not even aware that their data was being used for political profiling, which makes contestation a rather difficult thing to do.

Not only does the new conception of freedom as non-domination meet some moral judgements with the power of emerging technologies - specifically in the field of Big Data - but it also provides a way to formulate policy and legal regulations that are necessary to ensure users' freedom and autonomy. Freedom from domination hence proves to be a promising candidate for managing issues arising in the Digital Age.

The use-case of 'meaningful human control' illustrates a similar point. On a purely causal conception of 'control', someone in the driver seat of a semi-autonomous vehicle who is physically capable of intervening in the car's operation is thereby morally responsible. Moreover, on this conception, fully autonomous vehicles are

morally problematic because they seem to introduce responsibility gaps. To remedy this problem, Santoni de Sio and Van den Hoven (2018) have adopted a conception of control from a different context: guidance control (Fischer & Ravizza 1998). They subsequently modified the conception to make it suitable in the context of autonomous artificial agents. So this is another example in which an existing conception was introduced in a novel context.

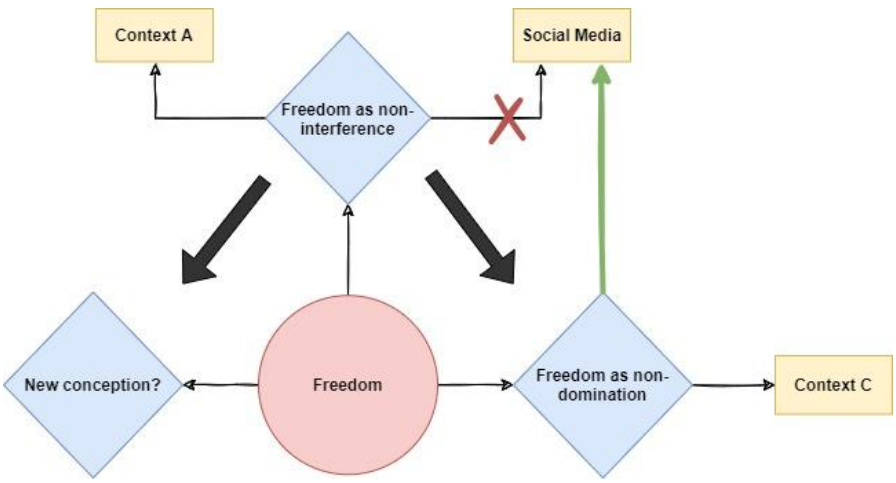


Figure 1: Schematic illustration of question 4 applied to the concept of freedom in the context of Social Media.

5. Conclusion

In this paper we have presented three use-cases that can be interpreted as concrete examples of conceptual engineering. In all three cases, the "old", current conception was shown to be morally defective. Certain conceptions fit better with our moral judgements than other conceptions, and, as we have seen, this moral defect is an important, albeit defeasible, reason to engage in conceptual engineering. We have argued that this can be understood as an instance of the pragmatic approach to conceptual engineering. The moral defect is a reason for conceptual change, because it is part of the function of that concept to contribute to that specific moral value.

We showed that the prime reason for opting for a different conception of *Control* is that the new conception has a better relation to responsibility attributions. For the concept *Critical Thinking*, we showed that the concept ought to promote the support of democratic values by encouraging citizens to take responsibility for evaluating information. Consequently, we argued that the new conception fulfills this function

better than the old one. For *Freedom*, we argued that the neo-republican perspective better captures morally problematic power relations than the more traditional freedom as non-interference.

We have also shown that the moral adequacy of our conceptions is context-dependent. Conceptions that are morally adequate in existing contexts can be shown to have moral defects in new contexts. This is why the disruptive nature of new technologies functions as an important driver for work in the ethics of technology. When new socio-technological ecosystems are introduced, new contexts in which 'old' conceptions are evaluated are introduced as well.

In the final section, we aimed to make explicit what answers engineering philosophers implicitly give to a number of open methodological questions concerning conceptual engineering. As such, this paper makes two contributions to the current literature. Firstly, it contributes to the literature on conceptual engineering by presenting three cases in which conceptual revisionary work has actually been done and has direct real world consequences. Secondly, we believe that it would be useful if future conceptual research were more integrated and explicitly interwoven with existing methodologies in the philosophy of technology (such as Values Sensitive Design and Design for Values). Our hope is that our findings are helpful for such a project.

Interlude 2

The previous chapter discussed the need for conceptual change in the context of freedom and online platforms due to their concentrated power. The same reasoning applies to AI systems used in core societal sectors. Specifically, I will argue that domination applies to the relation between those that shape a system (developers and deployers) and those affected by one (end-users). In order to make this claim, I need to provide a concrete conceptualization of how these power relations take shape in the context of core AI systems.

The main point of the following chapter, *Machine Learning and Power Relations,* is to argue that there is a power relation between those who shape an AI system and those affected by one. The chapter answers the following research question:

> **RQ2:** How should we understand the power dynamics underlying AI development and deployment between those who shape a system and those affected by it?

I conceptualize the power dynamics underlying AI development and deployment following Castelfranchi's framework of power-dependence relations. According to this framework, agent A has power over agent B when B depends on A to achieve their goal. In order to situate the normative relevance of these power dynamics, I have framed this chapter in the context of republican theory. According to this theory, uncontrolled power relations constitute the moral complaint of domination. Given the purpose of the paper is to provide a conceptualization of the power relations underlying AI development and deployment, the main focus lies on this conceptualization. Consequently, the link between domination and AI systems used in core societal sectors remains relatively underdeveloped. In Chapter 6, I provide a more substantive argument for the claim of 'digital domination'. For the purpose of the next chapter, however, the provided analysis suffices.

Chapter 2: Machine Learning and Power Relations[27]

Abstract There has been an increased focus within the AI ethics literature on questions of power, reflected in the ideal of accountability supported by many Responsible AI guidelines. While this recent debate points towards the power asymmetry between those who shape AI systems and those affected by them, the literature lacks normative grounding and misses conceptual clarity on how these power dynamics take shape. In this paper, I develop a workable conceptualization of said power dynamics according to Cristiano Castelfranchi's conceptual framework of power and argue that end-users depend on a system's developers and users because end-users rely on these systems to satisfy their goals, constituting a power asymmetry between developers, users and end-users. I ground my analysis in the neo-republican moral wrong of domination, drawing attention to legitimacy concerns of the power-dependence relation following from the current lack of accountability mechanisms. I illustrate my claims on the basis of a risk-prediction machine learning system, and propose institutional (external auditing) and project-specific solutions (increase contestability through design-for-values approaches) to mitigate domination.

Key words Responsible AI, machine learning, power relations, domination, AI design, design-for-values

1. Introduction

It is now well-established within the AI ethics literature that consequences of AI systems, particularly opaque machine learning (ML) systems, are not clearly separated from the people involved in the system's lifecycle. Human decisions influence the algorithm's training data, the chosen model, or feature weighing. One aspect of this influence relates directly to issues of power between those who shape a system and those affected by it, as reflected in the call to establish effective accountability mechanisms (e.g., Jobin, Ienca & Vayena 2019). In particular, there is an interest in who has–or should have–the decision-making authority regarding a system's development (e.g., Busuioc 2020; Coglianese & Lehr 2016; Crawford 2021; Kalluri 2020; Sloane & Moss 2019). The debate, hence, seems to invoke a moral intuition

[27] Published as Maas, J. (2023). Machine learning and power relations. *AI & SOCIETY*, 38(4), 1493-1500.

that there is something deeply problematic about how ML systems are currently developed and used within society.

Despite this intuition, there remains an inconsistency in the debate between the socio-economic importance of power and the level of conceptual clarity regarding what power is. Moreover, it remains unclear–even if we were to have a clear conception of power–how said power relations between people should be analysed from a normative perspective. Power relations entail exercises of power that inherently are normatively laden, implying that illegitimate power relations hinder responsible ML development. Thus understood, conceptualizing power relations is an underdeveloped part of AI ethics that we can–and should–ethically evaluate in order to identify potential pitfalls in current AI ethics initiatives that hinder responsible ML development (e.g., ethics washing through the use of ethics guidelines, see Hagendorff 2020).

In this paper, I investigate the power dynamics underlying the development and use of ML systems and argue that said power dynamics give rise to the moral wrong of domination. Domination, as understood by the neo-republican framework, occurs when one is subjected to a superior and unaccountable power (Pettit 1997). It constitutes a moral wrong as domination provides an obstacle to human flourishing, or what is necessary to lead a good life (Lovett 2010). The concept of domination fits well the debate on power within the AI ethics literature as it normatively and theoretically grounds the moral intuition that there is something problematic with the current power dynamics of ML ecosystems. My two main contributions with this paper are therefore (1) providing a workable conceptualization of said dynamics and (2) establishing normative grounds for familiar though relatively abstract issues of power and accountability of ML ecosystems.

My argument is as follows: first, the moral wrong of domination requires both superior and unaccountable power (Pettit 1997). Second, following the work of Cristiano Castelfranchi (2003), there is a power-dependence relation between those who shape a system (i.e., developers and users) and those affected by a system (i.e., end-users). This ultimately implies that those who shape a system wield *some* power. This power asymmetry reflects the superior power necessary to constitute domination. Third, we currently face a lack of accountability mechanisms in ML systems due to their opaque and learning characteristics, resulting in responsibility gaps (Matthias 2004). This constitutes (to some extent) an unaccountable power of the developers and users (*via* the ML system). Therefore, the power asymmetry of the developers and users in combination with the lack of accountability mechanisms constitutes the moral

wrong of domination, or at least gives rise to the potential of domination as current power dynamics are presented with the main ingredients necessary to constitute this moral wrong.

In the first sections, I lay out the building blocks for my argumentation. I discuss the concept of domination (section 2) and elaborate on the different actors (developers, users, end-users) involved in an ML system (section 3). In section 4, I discuss my core argument, i.e., that current power dynamics constitute a power asymmetry, and, consequently, that the lack of accountability mechanisms establishes the potential of domination. I end this paper with some recommendations at both institutional level (external auditing) and project-specific level (increase contestability through design-for-values approaches) on how to mitigate potential domination (section 5).

2. Domination

Domination, as understood by neo-republican theory, implies that one is subjected to a superior and unaccountable exercise of power (Pettit 1997). In other words, someone is dominated when they depend on another's unaccountable or arbitrary will, i.e., there are no effective accountability mechanisms in place to 'check' the power, obstructing the dominated agent's possibilities for redress when wronged or to contest the dominant agent's decision. This constitutes a moral wrong as it provides an obstacle to human flourishing, understood as to what extent an individual can flourish, and taken as a core value to realize[28] (Lovett 2010). Superior and unaccountable exercises of power hinder human flourishing as they establish insecure situations in which the subordinated agent is psychologically damaged because of a constant threat of abuse.

Indeed, as neo-republicans point out, a benevolent dictator remains a dictator, even in the absence of non-(harmful)-interference (Pettit 2011, 714). The fact that the dictator can choose to change his or her behavior towards the citizens implies that citizens subjected to the dictator are not secured from unlawful and potentially harmful interference. So, contrary to a dictator who has unaccountable power due to lack in effective accountability mechanisms, a democratic government, though

[28] Human flourishing constitutes the basis for several normative accounts (see Lovett 2010, 131, fn 6). We see aspects of this term incorporated by the European Parliament in the Charter of Fundamental Rights (2012) (e.g., in Title II 'Freedoms' and Title IV 'Solidarity'). I realize that these values may not be globally applicable due to contextual and cultural differences. For this paper, I endorse the European Union's key values, rooted in the value of democracy.

exercising power over its subjects, does have these mechanisms as its subjects control governmental power thanks to accountability mechanisms like public contestation and the separation of powers.

Though neo-republicanism often relates to states, a similar reasoning holds between two individuals (e.g., parent-child relation) or groups of individuals. To this extent, we see that an individual's ability, or 'power', to achieve their goal rests in their political relation with another agent (or agents, for instance, a child and multiple parents). This gives strength to neo-republican theory, as it crosses the boundary between the common distinction 'power-to' and 'power-over', where the former is often more understood in an individual's capacity to realize their goal and the latter often understood in an exercise of power between agents (Lovett 2010; Haugaard 2012).

This, however, is not to say that such power-over is necessarily problematic. Power-over becomes morally problematic in situations where the power-over unaccountably impedes an individual's power-to, thereby constituting the moral wrong of domination. Indeed, domination comes in degrees: it is constituted by the degree of the individual's dependency, the degree of the dominant agent's reach of power, and the degree of the arbitrariness of the exercise of power (i.e., opportunities to hold the dominant agent accountable for their actions) (Lovett 2010).

So, domination combines the idea of how an individual's power-to rests in their political relation with another with a lack in ability to hold the dominant agent accountable. Given the debate on power dynamics underlying ML ecosystems, domination, then, seems to fit well the moral intuition that is present in the AI ethics debate on power. Scholars mention the increase in power of those that have decision-making authority regarding the development and deployment of these systems, but criticize the public's lack in decision-making guidance or possibility to reverse a decision (e.g., Whittaker et al. 2018, 30). This resembles the idea of a dictator, benevolent (i.e., good decision-making) or not, in that the public is left with little control over the decision-making process. However, before making any claims related to potential domination, it is essential to identify which actors are involved and how to understand the power dynamics between these actors[29].

[29] Note that the neo-republican framework reflects a more Western and individualistic mindset. Although it is essential to recognize the limitations of a Western perspective, particularly given the limited focus on social relations (see e.g., Segun 2021), I endorse a neo-republican framework precisely for its link between individual and social power. Indeed, several neo-republicans argue that

3. Actors involved

I distinguish three main categories of actors in an ML ecosystem: developers, users, and end-users. The developers are the most relevant category regarding the influence on the system's behaviour resulting from design and deployment decisions, and so relate more directly to questions of power. I interpret the category 'developers' in a broad sense, including all those actors that are involved with the development of the software. With 'development' I refer to all input from the initial thought processes behind the system up to the moment the system is deployed[30]. Thus understood, developers include the management that is in charge of the business side of an algorithm, those that wield the "algorithm-specifying power" (Coglianese & Lehr 2016, 1216) including specifications related to value-judgements and determining acceptable error rates (Wieringa 2020, 3), and the programmers that code the algorithm. In addition, this category also includes stakeholders such as expert groups (e.g., doctors for medical AI). The 'user' category is more easily defined and relates to the actor that deploys the system (which can but need not be the same as the developing company). Finally, with 'end-user' I refer to the actor who is *directly* affected by the system. *Directly* affected means that the end-user needs to stand in a direct relation with the system itself, although of course the effects of a system can 'trickle-down' to other individuals[31]. In addition, the end-user must be the target of the algorithm.

To illustrate these different actors, consider an ML algorithm that is developed for a bank in order to determine whether applicants should receive a loan by analysing similarities of new applications with previous successful and unsuccessful ones (the 'loan-algorithm'). Here, the developers include management actors that are in charge of the business side of the algorithm and programmers that code the algorithm. The user is the bank that implements the system and applies it to its customers: the system's end-users. The management, programmers and users all play an essential role with regard to their relation with the end-user: the management provides the opportunity for the algorithm to be created in the first place, the programmers design the system,

---

domination is necessarily embedded within larger societal structures (e.g., Gädeke 2020), highlighting the socio-contextuality between the more and the less powerful within a given relation.

30 Although software development is usually an iterative process and hence deployment may inform the development again.

31 E.g., a DSS that predicts my credit score has a direct relation with me and an indirect relation with my family as it will likely affect them also if I get flagged as high-risk and miss out on an important loan, mortgage, or social benefit.

and the user employs (and interprets) the system which all ground the system's influence on end-users in the real world. Note that although an algorithm determining whether an applicant receives a loan directly affects the bank as well, the bank is not the target of the algorithm so does not conform to the end-user criteria.

Besides the roles of actors, we can distinguish between *levels* of actors, referring to the individual, group and organization level (Wieringa 2020). To illustrate, consider again the loan-algorithm, focusing only on the role of a 'programmer': on the individual level, we have one programmer developing the algorithm; on a group level, we have a team of programmers that together are responsible for the coding of the system; on the organization level, the programmers blend in with the company for which they work, i.e., the bank then forms the 'developing' actor.

There are, of course, many other roles of actors involved, which makes isolating one particular 'role' (e.g., 'programmer') impossible if not incorrect. For instance, credit-scoring algorithms often use open-source software that was not necessarily built by the programmers employed by the bank. The point, however, is to show that when discussing a particular role of an actor, for instance in the context of assigning responsibility, we must keep in mind that it matters for the discussion whether we talk about *one* individual, a group of individuals, or refer to the developing actor in general, since moral and legal responsibility are not necessarily equivalent. As these different roles and levels of actors confirm, the influence and corresponding power relations occurring during the development and use of an ML system are not traceable to one particular individual involved in the process (Mittelstadt, Allo, Taddeo, Wachter & Floridi 2016) [32].

Moreover, the involvement of each actor depends on the context and type of algorithm that is developed, so to isolate one role or level of actor who influences the system does not do justice to the broader societal structures in which the development and deployment of an ML system takes place. For instance, an algorithm used for public policies with a different developer and user arguably requires more consultation with stakeholders and the algorithm's user than an algorithm developed and used by the same company for its private ends, such as Facebook's recommendation systems. There is hence an interplay between the algorithm's development and deployment context and the actors' degree of involvement with the development and use, which determines the distribution of influence on the system amongst these actors involved.

[32] I thank an anonymous reviewer for pressing me on this point.

4. Machine Learning and Domination

So what is the connection between domination and the influence of developers and users on a system's behaviour? The moral wrong of domination urges us to critically reflect on any relation between actors involved in an ML ecosystem because a concrete moral concern is at stake: that is, one's potential for human flourishing. Yet a relation of influence does not necessarily constitute exercises of power, let alone *illegitimate* exercises of power. In the following two subsections I argue that there is in fact a potential for such illegitimate exercises of power.

4.1. Power-Dependency relation

First, I argue there is a power-dependence relation between the developers and users on the one hand and end-users on the other. For this I turn to the theoretical framework of Cristiano Castelfranchi, who shows how dependence relations turn into power relations. According to Castelfranchi, dependence is based on one agent's "*Power-of*" and another agent's *"Lack of 'Power-of'"* (2003, 216, original emphasis). With 'Power-of', Castelfranchi refers to both internal and external 'powers' that enable agent X to execute action A in order to achieve her desired goal G (Castelfranchi 2003, 213). So, when agent X does not have the ability (power) of doing A to get G, she lacks either skill, resource, or opportunity (Castelfranchi 2003, 214). When agent Y does have this power of producing A to fulfil G, X depends on Y doing A so X can achieve G. Dependence can hence be defined as "X needs Y's action or resource to realize [Goal]" (Castelfranchi 2003, 216).

Note that dependence relations go hand in hand with power relations (i.e., dependence and 'power-over' are intrinsically related). Indeed, where X needs Y's action to realize her Goal, this simultaneously implies that Y has a "capability (power) of letting X realize her [Goal]", resulting in Y's 'power-over' X (Castelfranchi 2003, 221; 2011). Castelfranchi's power-dependence relation is appealing as it discusses how one's individual power becomes someone else's power. This reflects one of the two main ingredients of domination, i.e., a dependency (and hence power) relation between two (groups of) agents. Thus Castelfranchi's framework bridges the gap between theory and practice as his description allows both for a conceptualization of current power dynamics of ML ecosystems and an ethical evaluation of potential wrong done to end-users.

There are other models that discuss power relations in multi-actor systems (Singh 2014; Kafali, Ajmeri & Singh 2019). For instance, the models of Singh (2014) and

Kafali et al. (2019) are based on the interplay between social factors, technical factors, and 'norms' that form the heart of a socio-technical system. These norms can be understood as power relations as well. While these models could similarly provide a conceptualization of the power dynamics underlying an ML system, particularly emphasizing the socio-technical elements of said system, they less explicitly bring the individual actor to the foreground and are less concerned with the step from *individual* power to *relational* power. For this reason, Castelfranchi's framework is more suitable for the purpose of this paper.

So how does Castelfranchi's framework relate to ML systems? Here, I argue that the influence of developers and users on an ML system produces a dependence asymmetry between those who develop and use the system and the end-users. Given that (1) the developers and users of systems have an influence on the system's behaviour, and (2) the system has an effect on the end-users, the end-users depend to some extent on the developers and users to design and deploy the system in such a way that it meets the end-users needs, upholds their rights, and respects democratic values like privacy, freedom, and autonomy. This dependence then, following Castelfranchi, automatically entails that the developers and users have some 'power-over' the system's end-users. To illustrate this dependence, 'power over' and their relation to the influence of developers and users, consider again the loan-algorithm mentioned previously.

The loan-algorithm is part of decision-support systems (DSS), which are increasingly used as predictive tools in numerous fields to indicate a level of some risk (e.g., health risk, fraud risk, recidivism risk). End-users stand in relation with a DSS when it makes a risk-profile of them. In the case of the loan-algorithm, the risk-profile is based on the applicant's credit score. In determining whether the applicant should receive a loan he or she is profiled by a DSS. The end-user is hence necessarily dependent on the DSS–and the human involvement that accompanies the DSS–in order to receive the loan. More formally: end-user [Agent X] lacks the power-of attributing approval [Action A] to receive a loan [Goal G], whereas the bank does have this power to attribute approval (by using DSS). In this sense, the end-user depends on the DSS. Yet since ML systems are socio-technical and constituted by social factors, the dependence of end-users on the DSS indirectly corresponds to a dependence on the DSS developers and users, constituting a power-dependence relation between the developers, users, and end-users *via* the DSS.

An implicit claim in this power-dependence relation is that developers and users wield power over end-users (*via* the system). This is rather strong and arguably an

objectionable claim: there are so many actors involved during the development of an ML system that any individual influence is negligible, let alone that it could count as an exercise of power. Yet, as the different roles and levels of actors illustrate, we should not isolate one particular individual. The point is that when all actors are put together there is in fact an exercise of power. Indeed, to this extent, ML systems 'shift power' towards the developers and users (Kalluri 2020). The power-dependence relation is hence not so much meant to discuss the power of one individual developer or user in relation to one individual end-user, it is rather to show the power dichotomy, necessarily constituted by the ML system[33], between those who shape the system and those affected by it.

4.2. ML systems and their lack in accountability

Second, this power dichotomy is interesting for an ethical evaluation. If such power is exercised in an unaccountable manner, there is a serious potential for the moral wrong of domination. And this, I argue, is precisely the case with ML systems. ML systems are notorious for their opaque and learning characteristic. Their learning characteristic weakens the causal relation between the design process and the system's behaviour, which creates so-called responsibility and accountability gaps where no individual can be reasonably held responsible for the system's behaviour (Matthias 2004; Santoni de Sio & Mecacci 2021).

Moreover, the (current) opacity of the system enforces these gaps as it provides technical limitations to system interpretability (Lipton 2018). Although sometimes potentially discriminatory inferences are identified in ML systems that developers can either tend to (e.g., Google's classification of people as 'gorillas') or choose to abstain from using the system (e.g., Amazon's sexist recruitment tool), the model's opacity makes such identification difficult and not always successful. This is problematic as (1) identifying causal relations within the data is necessary to judge the moral and epistemic reliability of a system, and (2) identification of causal relations between the developers' input and the system's behaviour is necessary to assign moral responsibility and accountability, which is in turn essential for establishing effective accountability mechanisms. To this extent, machine learning systems, due both to

[33] This softens the claim that the 'shapers' wield power, as any exercise of power similarly depends on the ML system. However, the claim is stronger than arguing that the power rests solely in ML systems (see also Neyland & Möllers 2017).

their learning characteristic and their opacity, reduce the room for accountability (see also Diakopoulos 2015; Busuioc 2020; Wieringa 2020).

Consider again the loan-algorithm, which bases its recommendation for new applications on statistical similarity. If most applications containing a particular postal code did not receive a loan, the system learns to reject new applications with that same postal code. This implies that new applicants are judged on *other* people's applications, rather than being individually assessed. This need not be an issue, yet bias in training data can lead to discriminatory outputs. Moreover, the algorithm may use new applications as input data, thereby establishing a biased reinforcement loop.

Now, whether it is fair to judge someone based on statistics arguably depends on one's choice of normative framework. For neo-republicans, such treatment might be acceptable *as long* as there are effective mechanisms in place that allow the end-user of the system to hold the relevant actor accountable. Yet, since it might not always be clear on which grounds a system produces its output and whether these grounds are morally–and legally–justified (Hildebrandt 2021), holding the responsible actor to account is not always easy. We are therefore confronted with a lack of effective accountability mechanisms due to the opaque and learning characteristics of the ML system.

So, combining the power-dependence relation with the lack in accountability mechanisms, we see the ethical dimension underlying the power-dependence relation of ML systems following the moral wrong of domination. Those who shape the system stand in a power-dependence relation with those affected by it, constituting a power asymmetry via the ML system. And the fact that it is not always clear who to hold accountable and on what grounds induces unaccountable exercises of power to which the system's end-users are then subjected. This, therefore, creates the potential for domination.

I explicitly state *potential* for domination. Domination, as mentioned before, comes in degrees. Ultimately, it depends on whether an end-user has the possibility to use a different system, how extensive the dominant agent's reach of power is, and to what extent there is *some* accountability possible. In the case of the loan-algorithm, for instance, the degree of domination would increase if there is only one bank available. If there are multiple options for the end-user to turn to, there is less dependence on that particular bank. Moreover, if the person is seeking a loan merely to have some spare money on their account the effects of (not) granting a loan are arguably less significant than when a person requires a loan to support their family or buy a house. Finally, if the bank appoints one person to be responsible for all output of the system,

there is at least some (legal) accountability. Hence, these three factors contribute to the degree in domination, ultimately making domination a possibility and not necessarily an unavoidable consequence.

Nonetheless, any potential for domination is problematic. Indeed, to be increasingly dependent on such an unaccountable exercise of power is not just problematic when the system proves to be incorrect in its results, it is problematic more generally as it opens up the possibility for a moral wrong, limiting human flourishing by establishing a power dichotomy between the developers and users, on the one hand, and the end-users, on the other. We should therefore seriously consider the potential political asymmetry that the increased use of ML applications bring to society, where developers and users–in combination with the ML system itself–increasingly gain more power over a system's end-users due to inadequate accountability mechanisms.

To conclude this section, Castelfranchi's framework of power-dependence illustrates how different actors in a system stand in relation with each other; in particular, how we can understand the power dynamics between the developers, users, and end-users. In addition, the lack of accountability mechanisms in ML systems are sufficiently worrisome due to their opacity and arising responsibility gaps that current power dynamics establish the potential for the moral wrong of domination of the developers and users over the end-users *via* the system.

5. Moving forward

In order to mitigate potential domination between those who shape the system and those affected by it, there are two general ways forward (cf. Pettit 1997). Either we equalize the level of power amongst the actors, thereby removing the 'superior' power necessary for domination, or we increase effective (i.e., promoting non-domination) accountability mechanisms, thereby removing the 'unaccountable' power necessary for domination.

The first option requires an equal level of power amongst the developers, users, and end-users of an ML system. This, however, is an unrealistic ideal. It is simply not feasible to have everyone participate as a developer, a user, and an end-user, which would be necessary to equalize levels of power. Moreover, these power imbalances are in fact inevitable, as not everybody has the technical knowledge or ambition to be involved with ML systems as a developer or user.

This leaves us with the second option: developing effective accountability mechanisms. Such accountability mechanisms can be either on a broader, institutional level (e.g., legal regulation) or on a project-specific level (e.g., necessary accountability measures for a particular ML system). I briefly discuss these two options in turn.

5.1. Institutional accountability: ethical guidelines and legal regulation

Establishing algorithmic accountability at the institutional level has already received much attention in the literature, particularly in the form of ethical guidelines (for an overview see Jobin et al. 2019) and proposals for regulatory frameworks (e.g., the recently proposed Act for AI regulation by the European Commission). However, while numerous scholars have already honorably devoted their attention to improving algorithmic accountability (for an overview see Wieringa 2020), these initiatives do not always necessarily mitigate domination[34].

For instance, the more wide-spread initiatives like the development of ethical guidelines have been criticized either for purely being a "marketing strategy", leading to 'ethics washing', or for their implementation showing "no significant influence" on the decision-making process during the development of these systems (Hagendorff 2020, 113). Arguably, such soft regulatory initiatives are ineffective to ensure responsible ML development. In response to these 'soft' initiatives, we find calls in the European Commission's AI Act for auditing and internal control checks aimed to increase accountability. However, it is unclear what such auditing should look like, and therefore to what extent it might effectively increase accountability.

Moreover, we must question to what extent *internal* control checks will be sufficiently effective. Indeed, the potential 'ethics washing' illustrates that we should not always trust companies to do the right thing. A neo-republican solution therefore requires *external* control mechanisms as an effective check and balance mechanism, as only external mechanisms ultimately cross the power dichotomy between a system's developers and users and its end-users.

Some scholars do note the need for external checks, pointing out how external audit mechanisms lead to less discriminatory outputs (e.g., Rambachan, Kleinberg, Mullainathan & Ludwig 2020; Kleinberg, Ludwig, Mullainathan & Sunstein 2018; 2020). Although these scholars do not explicitly address morally problematic power

[34] I discuss these concerns as well in Maas (2022), in which I argue that AI ethics should incorporate the neo-republican ideal of freedom as non-domination.

relations, they do show promising results for how accountability mechanisms in line with mitigating the moral wrong of domination actually contribute to more just ML systems.

However, such legal regulation is morally not fully satisfactory, as there is a difference between moral and legal accountability (i.e., liability) that the development of legal regulation may overlook. While legal liability definitely is one–and still underdeveloped–way to hold someone accountable in case of wrongful output, moral accountability is a more difficult topic due to the responsibility gap in ML systems. And although legal liability is a first step in the right direction towards effective accountability mechanisms as it provides a means for end-users to enforce accountability, thereby shifting the power from the 'shaping' side to the 'affected' side[35], we ultimately want such mechanisms to be fair as well, that is, to find the intricate combination of legal and moral accountability. Therefore, we need a second and complementary way to mitigate dominating tendencies.

5.2. Project-specific accountability: design-for-values approaches

One option is through so-called design-for-values approaches, such as value-sensitive design (VSD) (Friedman, Borning & Kahn 2002), participatory design (PD) (Simonsen & Robertson 2012) or Responsible Research and Innovation (RRI) (Owen, Macnaghten & Stilgoe 2020), and other democratic initiatives for technological innovation like participatory Technology Assessment (pTA) (Joss & Bellucci 2002). Although these approaches may require some adjustments due to the learning character of ML systems (Umbrello & Van de Poel 2021), they provide fruitful grounds for mitigating dominating tendencies because they aim to integrate stakeholder input during the system's entire lifecycle, including early planning stage and deployment stage.

Note that these approaches are not the same as equalizing levels of power as these approaches still distinguish clearly between the 'shaping' group and the 'affected' group. Instead, democratic design approaches like VSD or PD invite stakeholders to voice their concerns or preferences during the design process. This way, stakeholders have the opportunity to contest design and deployment decisions made by developers and users of a system during the lifecycle of the system. Especially in the context of ML systems, where the inherent opacity and learning characteristics of these systems provide inevitable technical limitations to *ex post* accountability mechanisms and

[35] I thank an anonymous reviewer for pointing this out.

increase the possibility for unintended biases, tending to potential ambiguous yet important design decisions during development of a system positively contributes to accountability by increasing moments for contestability–and hence control–for the affected group. For instance, end-users can be a greater part of testing, identifying earlier on potential problems (e.g., ensuring a diverse group to test the algorithm to avoid problematic consequences such as Google classifying people as gorillas). This way, moral accountability also increases as it is easier to pinpoint morally contestable decisions at a specific moment during the development process. Democratic design approaches hence match the neo-republican ideal for democracy, as they allow some form of public control.

That said, these approaches also have their drawbacks. For instance, VSD is often criticized for its vagueness regarding stakeholder inclusion (Davis & Nathan 2015). Yet clear decision-making processes, which include *why* and *how* developers choose their stakeholders, weigh different values, and to what extent stakeholders have the ability to contest developers' decisions, are essential to neo-republican theory and to realizing the ideal of non-domination.

6 .Concluding remarks

In this paper, I have attempted to provide a deeper analysis regarding the social relation between an ML system's developers and users and the system's end-users by firstly providing a workable conceptualization of the power dynamics underlying the development and use of an ML system. Here, I tried to show that there is some form of dependence of an ML system's end-user on the system's developers and users, with dependence understood in the sense that one agent requires another agent to perform a particular action. Following Castelfranchi's framework, this dependence simultaneously contributes to the developers and users' 'power-over' the end-users. Secondly, I have evaluated the moral concern of the combination of a power asymmetry and a lack of effective accountability mechanisms, grounded in the example of a risk-scoring DSS, in light of the neo-republican concept of domination, and discussed how this concept of domination can contribute to developing effective and fair accountability mechanisms on both institutional and project-specific levels. Though the ideal of non-domination provides fruitful grounds to establish effective accountability mechanisms, the solutions I have presented are still in their early stages and require extensive further research.

Interlude 3

The primary purpose of the previous chapter was to argue that there is a power relation between the 'shapers' (i.e., the developers and deployers) and the 'affected' of an AI system (i.e., the end-users). Drawing on this conclusion—and making use of the conceptual engineering methodology—the next chapter explicitly questions the conceptual background that informs AI Ethics. My primary purpose in this chapter is to argue that the AI Ethics literature has focused strongly on the concept of freedom as non-interference, thereby neglecting the moral wrong of domination. The chapter addresses the following question:

> **RQ3:** How does the conceptual background of freedom that informs the AI Ethics literature steer the focus away from the power dynamics underlying AI development and deployment?

I answer this question by showing how much of the AI Ethics is centred on preventing harm. This focus risks a bias that overemphasizes the ethical *behaviour* of a system while underemphasizing addressing the underlying power dynamics. This causes more attention to go in developing AI systems that conform to certain ethical principles and guidelines, while less attention goes out to the question *how* these systems are developed in the first place. The 'how' question is relevant, as it matters whether relevant stakeholders have had the opportunity to contest and challenge design decisions. This in turn is one way to mitigate domination. I elaborate on these points in chapters 4 and 5.

There are three things to note in advance regarding the chapter. First, although in this chapter I do not explicitly discuss the AI Ethics literature in the context of Pasquale's first and second waves, the chapter primarily targets the first wave in AI Ethics (i.e., the wave concerned with improving the behaviour/output of AI systems). Second, my main interest in this PhD is with AI systems used in core societal sectors such as healthcare or public administration. This paper uses the Cambridge-Analytica (CA) as a case study. The case study is exemplary to illustrate my concern with the focus of freedom as non-interference, and my argument can be extended to healthcare, public administration, and other core sectors. Finally, I focus strongly on Mill's harm principle. As mentioned in the introduction, this principle is heavily contested. To clarify my position, my interpretation of harm is a form of interference in such a way that I experience negative consequences (mental or physical). How I

perceive the harm principle is thus less concerned with how an action came about, and more with the consequences of that action.

Chapter 3: A Neo-Republican Critique of AI Ethics[36]

Abstract The AI Ethics literature, aimed to responsibly develop AI systems, widely agrees on the fact that society is in dire need for effective accountability mechanisms with regards to AI systems. Particularly, machine learning (ML) systems cause reason for concern due to their opaque and self-learning characteristics. Nevertheless, what such accountability mechanisms should look like remains either largely unspecified (e.g., 'stakeholder input') or ineffective (e.g., 'ethical guidelines'). In this paper, I argue that the difficulty to formulate and develop effective accountability mechanisms lies partly in the predominant focus on Mill's harm's principle, rooted in the conception of freedom as non-interference. A strong focus on harm overcasts other moral wrongs, such as potentially problematic power dynamics between those who shape the system and those affected by it. I propose that the neo-republican conception of freedom as non-domination provides a suitable framework to inform responsible ML development. Domination, understood by neo-republicans, is a moral wrong as it undermines the potential for human flourishing. In order to mitigate domination, neo-republicans plead for accountability mechanisms that minimize arbitrary relations of power. Neo-republicanism should hence inform responsible ML development as it provides substantive and concrete grounds when accountability mechanisms are effective (i.e. when they are non-dominating).

Keywords AI ethics, machine learning, accountability, Mill's harm principle, freedom as non-interference, domination, neo-republicanism

1. Introduction

The call to establish effective accountability mechanisms is a topic of major concern for AI systems. The majority of ethical AI guidelines conclude that ethical, responsible, or trustworthy AI requires 'accountability' or 'responsibility', though what precisely is meant with these terms remains to a great extent unclear (Jobin, Ienca & Vayena 2019, 10). The Trustworthy AI Guidelines established by the European Commission (EC) (2019), for instance, call for 'audits', but provide little concrete design recommendations. Though the desirability of accountability mechanisms is widespread, it is unclear what such mechanisms should look like. Particularly machine learning (ML) systems prove to be a challenge due to their lack

[36] Published as Maas, J. (2022). A Neo-Republican Critique of AI Ethics. *Journal of Responsible Technology, 9*, 100022.

in interpretability (Burrell 2016; Lipton 2018) and arising responsibility gaps (Matthias 2004; Santoni de Sio & Mecacci 2021). So far, the lack of accountability for ML systems remains a topic of concern.

In this paper, I argue that the difficulty with formulating effective accountability mechanisms is partly rooted in a limited focus of such guidelines and other initiatives, like the recent EC's proposal for an AI regulatory framework (2021). These initiatives predominantly focus on preventing harm, which risks to overlook other starting points to formulate accountability mechanisms, such as whether the power dynamics surrounding the development and use of ML systems are legitimate. Here, power dynamics are understood as the relation between the decision-making authorities who shape the systems and those affected by the systems without such authority. This implies that responsible ML development should not just aim to minimize or prevent potential harm, but also focus on *how* the interference constituting this harm is embedded within broader societal and institutional relations underlying the development and use of a ML system.

Based on current power dynamics between those who shape a ML system and those affected by a system, I propose that neo-republican theory provides a suitable framework to inform responsible ML development. Neo-republican theory concerns itself with whether interference is done in a legitimate manner (Pettit 1997). That is, the power that enables the interference must be controlled and held accountable, as uncontrolled or unaccountable interference constitutes the moral wrong of *domination*. Domination is a moral wrong as the dominated agent lives in constant threat of potential harmful interference, which ultimately diminishes human flourishing (Lovett 2010). The moral wrong of domination illustrates that AI ethics should not just focus on preventing harm, but should be concerned with who has the power to inflict this harm in the first place. This shifts the focus from harm through interference to a focus on power dynamics regarding ML systems.

My aim for this paper is not to develop a proper account for what non-dominating accountability mechanisms must look like in the context of ML systems, nor is my aim to develop a full account of how the power dynamics underlying the development and use of ML systems are shaped. Rather, with the neo-republican perspective, I wish to provide an alternative starting point which allows concrete guidance with regards to the establishment of effective accountability mechanisms for non-transparent and self-learning AI systems.

In order to substantiate my argument, I first discuss how the AI ethics literature is limited in its narrow view on what elements (i.e. harm) constitute a moral wrong.

Second, I discuss the benefits of a neo-republican perspective. Third, building on previous work (blinded for peer review) I relate the neo-republican framework back to ML systems. I ground these claims in the concrete example of the Facebook-Cambridge Analytica scandal. I conclude with some criticisms on current initiatives within the AI ethics (e.g., guidelines, limitations of regulatory frameworks) following from a neo-republican framework.

2. Responsible ML development: liberal thought as main source

While we should only applaud and support the growing attention towards responsible ML[37] development, it still is a nascent field facing room for improvement. In this paper, I argue that some of this room relates to the power dynamics associated with a ML system. Although current initiatives in the field implicitly hint at those power dynamics by referring to the lack in accountability mechanisms, the main focus remains quite restricted on preventing or minimizing harm caused by ML systems, rooted in the conception of freedom as non-interference.

Freedom as non-interference, also known as negative freedom (Berlin 1969), implies the *absence* of interference by others in your choices, regardless your preference in choice. A particular application of this view is Mill's harm principle, according to which one is free to do anything they wish unless they harm someone else. Harm, here, is understood as an action that injures rightful interests of others (Mill 1998, 83). Following this principle, if there is a risk of harm, it is morally permissible to regulate such potential harmful behaviour.

Berlin (1969) distinguishes negative freedom from positive freedom. Whereas negative freedom focuses on the absence of interference by others, positive freedom is more concerned with whether one is in control of their own choices, and hence considered positive precisely because of the 'presence' of this self-mastery or self-determination.

The negative conception of freedom is often associated with liberal democracies (Carter 2003), which commonly prevail in Western countries. Since AI ethics is still a field predominantly explored by scholars, companies, or other institutional organizations of Western countries (Hagendorff 2020), it should not come as a surprise that this conception then also prevails in such initiatives. For instance, in the Trustworthy AI Guidelines formulated by the European Commission (EC) (2019), one

[37] For the purpose of this paper, I solely discuss ML systems, although some of my following arguments apply to rule-based systems as well.

of the four principles is the 'prevention of harm'. This principle entails the "protection of human dignity as well as mental and physical integrity" and hence strongly resembles Mill's harm principle (European Commission 2019, 12).

This overall focus on freedom as non-interference is supported by the findings of an analysis of existing guidelines by Jobin et al. (2019). Following their analysis, the authors provide the five most mentioned principles out of 84 guidelines: transparency, justice and fairness, non-maleficence, responsibility, and privacy, which all aim to either minimize or prevent harm. Transparency is seen "as a way to minimize harm" (Jobin et al. 2019, 8), justice should prevent bias, non-maleficence must ensure that AI never causes "unintentional harm" (Jobin et al. 2019, 9), responsibility is concerned with "potential harm" (Jobin et al 2019, 10), and privacy should secure us from data breaches, which similarly is concerned with harm through interference.

Some attention does go out to other themes more related to positive freedom, often expressed in terms of autonomy. For instance, Mittelstadt et al. (2016) point out that algorithms can undermine autonomy, especially in the case of personalisation algorithms that compute which content the end-user would like to see. In addition, Floridi and Cowls argue that an increase in automated decision-making restricts one's positive freedom as it diminishes the "flourishing of human autonomy" (2019, 7, emphasis left out). Yet both these worries require *actual* interference, as these worries to one's autonomy only become visible when we are shown personalized content or are subjected to AI decision-making, and hence even here we see a nod towards the harm principle.

Exceptions are those guidelines that point explicitly towards the value of self-determination, such as the principle of respect for human autonomy of the EC's Trustworthy AI guidelines, or, perhaps not coincidentally also a European initiative, the principle of democracy as formulated by the European Group on Ethics (EGE) (2018) which views self-determination as a human right. Nevertheless these exceptions, most principles for responsible ML development draw from the liberal tradition, aiming to prevent harm through interference.

This focus on freedom as non-interference, or the harm principle in particular, is problematic as it does not account for other fundamental concerns, in particular that someone does not necessarily need to interfere with and harm another in order to do them morally wrong. I understand this moral wrong as being in a vulnerable position to abuse. Such a vulnerable position implies that one's set of choice options depends on someone else. Although dependence in itself is not necessarily problematic – and arguably essential to social life – what makes one vulnerable is to have no means to

redress, or hold accountable, those on whom you depend. In other words, one is vulnerable to abuse when there is *unaccountable* interference.

To this extent, the AI ethics literature will substantially improve by expanding its focus from preventing or minimizing harm to questions related to power, such as who has the ultimate authority to do harm. Here, we can think of questions like who decides whether these systems get developed in the first place, what decisions are made during the design process, and who are ultimately affected by these systems?

Such questions related to power increasingly gain recognition. For instance, Kate Crawford (2021) has recently published in her book *Atlas of AI* the production of an AI system in which she discusses power relations by emphasizing socio-technical aspects of an AI system (e.g., the lithium mining necessary for AI computation). Other scholars have pointed out that system design necessarily involves value-judgements (e.g., Busuioc 2020), raising questions of who should have the authority regarding such judgements (e.g., Kalluri 2020; Sloane & Moss 2019; Yudkowszky 2004). Yet these accounts either do not clarify precisely how the power dynamics between the different actors involved are shaped or do not provide substantial normative background why such power dynamics are problematic.

In this paper, I build on previous work in which I develop a full account of how these power dynamics between different actors are shaped (blinded for peer review) to provide a substantial normative claim why AI ethics should expand its focus to questions of power. My attempt so far has been to point out limitations of AI ethics, namely that there is currently an imbalance between a focus to prevent or mitigate harm and other moral wrongs related more to issues of power. In section 5, I return to how a preference for the liberal tradition hinders responsible ML development by analysing limitations in current initiatives (i.e. guidelines, regulatory frameworks). But before I do so, let me elaborate on the normative implications of being *vulnerable* to a harm.

3. Domination

The neo-republican concept 'domination' captures well the moral wrong of unaccountable interference. One is dominated when one is subjected to a superior and arbitrary or unaccountable power (Pettit 1997). Neo-republicans argue that what makes a person free is whether this freedom is robust, that is, you are secured from interference by others *even if* they wish to interfere. So, neo-republicans look more at the political relations between two agents, rather than judging one's freedom to the

extent whether they are or are not experiencing some instance of interference. Non-domination hence implies not having to depend on someone's arbitrary will.

To this extent, domination has also been defined more concretely in terms of dependence, with domination conceived as "a condition experienced by persons or groups to the extent that they are dependent on a social relationship in which some other person or group wields arbitrary power over them" (Lovett 2010, 2). Note that according to this definition, three requirements must be met: (1) there must be some kind of dependence; (2) there must be some power asymmetry; and (3) the power must–to some extent–be wielded arbitrarily.

First, dependence can be defined as exit costs (Lovett 2010, 38). Exit costs imply the subjective experience of what it means to lose the position you are in. For instance, in patriarchal societies, wives depend for a great deal on their husbands. If they would leave their husband, they enter a position that they may judge less desirable than being dependent on the husband (i.e. living life as a spinster). For some women, these exit costs of living life as a spinster do not outweigh a life dependent on their husband. Other women, however, could judge these exit costs worthwhile in light of the alternative (i.e. depending on the husband). These different groups may have lived under equally intrusive relationships, yet for one the exit costs were acceptable and for the other they were not. One's dependency is hence contingent on the subjective interpretation of the value of the exit costs as determined by the dominated agent and not through some objective measure.

Second, in order to discuss what is meant with the condition of power asymmetry, I must first discuss the concept of power. Though this is a much contested concept, for the purpose of this paper I will endorse the account put forth by Lovett (2010). Lovett distinguishes between different levels of power. One level follows a Hobbesian account of power, more related to an individual's potential courses of actions. The second level relates more directly to the concept of domination, according to which "one person or group has *power over* another if the former has the ability to change what the latter would otherwise prefer to do" (Lovett 2010, 75, original emphasis). Note that one's power over another directly relates to the individual level of the latter's ability to choose or prefer a potential course of action, and the exercise of this power is done either through more direct exercises of power such as coercion or more indirect measures like persuasion or manipulation to alter the subordinate agent's preferences and/or actions (Lovett 2010, 76-77).

Following this idea of power over, we can speak of a power asymmetry between two agents if one has more power over another than the latter over the first. For

instance, a security guard in a museum can throw me out of the museum if I touch a painting or sculpture, despite me wanting to continue my tour of the museum. This, then, counts as an instance in which the guard and I are in a power asymmetry.

Third, and most importantly, there is a difference between legitimate exercises of power and illegitimate exercises of power, where the former does not constitute domination and the latter does. Whether an exercise of power is legitimate or not depends on the manner how power is exercised. That is, when power is exercised arbitrarily, power is illegitimate. An arbitrary exercise of power implies that the power is not "externally constrained by effective rules, procedures, or goals that are common knowledge to all persons or groups concerned" (Lovett 2010, 96). To this extent, an agent can exercise their power without having to face any costs for his or her deed. For instance, it is arguably a legitimate action for a security guard to throw me out of the museum because I touched a painting. Indeed, I know this is wrong to do. However, a guard throwing me out without a reason seems illegitimate, as I did nothing wrong. In this situation, I could complain to the manager of the museum to rectify this illegitimate exercise of power. If the manager holds the guard accountable, then the initial illegitimacy of the action is rectified. If, however, the manager does not rectify the guard's exercise of power, we may speak of *arbitrary* power as the guard exercises their power without facing sanctions.

Finally, note that these three conditions are not on/off binaries. The subjective value of one's exit costs are not one or zero, but should be seen in a range. Similarly, the power an agent has depends on their ability to successfully alter the course of action of the subordinate agent, which can vary from the latter changing their course of action only slightly to tremendously. In addition, arbitrary power depends on the number of constraints in place and the effectiveness of such constraints, hence also providing a degree in arbitrariness. The fact that these three conditions come in degrees implies that domination itself also comes in degrees (Lovett 2010; see also Pettit 2005).

That said, regardless its degree, domination always constitutes a moral wrong. Domination should be regarded as a moral wrong as it provides an obstacle to human flourishing (Lovett 2010). Lovett (2010) discusses three ways in which a person is restricted in their flourishing. One, a dominant agent can coercively demand particular services or goods from the subordinate agent, which disallows the subordinate agent to fully develop themselves and explore their opportunities in life. Two, living under arbitrary rule implies living under constant threat of abuse. Such a threat constitutes insecurities, which ultimately frustrates one's autonomy by

hindering their development of arranging their life plans. Three, one's dependence on another's arbitrary will entails a second-class social and political standing, as not you but someone else has authority over your life. This, then, provides an obstacle to human flourishing as such inequality may reduce one's self-respect. Given the impact on human flourishing, expressed in material or psychological coercion, increased insecurities and accompanying reduced autonomy, we as society have a duty to always try to mitigate domination.

So, what is desirable about the neo-republican theory in the context of ML systems? In order to answer this, I must first contrast the conception of freedom as non-domination with that of freedom as non-interference.

Freedom as non-domination is both broader and more narrow than freedom as non-interference. It is broader in the sense that one's freedom may be restricted even when this person is free from actual interference. It is more narrow in the sense that even when someone is interfered with, she is still free. So, whereas freedom as non-interference merely focuses on the *breadth* of one's choice (are all options available?), neo-republicans believe that we should focus on the *depth* of one's choice (who has the ultimate authority and decision power of this choice?). As long as I am in authority of my own choice – albeit indirectly, e.g. through public control in cases of government-society relations – I am a free person. Domination hence does not just focus on the actual interference, but includes the threat of illegitimate interference.

A paradigmatic neo-republican example that illustrates the contrast in moral wrong in this threat is the master-slave relation. As a slave, you are subjected to the will of your master. The master controls your moves, decides whether you eat or sleep, and may choose to beat you if desired. Surely, this is a morally problematic situation, both for neo-republicans and traditional liberals. A key difference between these two conceptions of freedom, however, is whether a benign and non-interfering master, who allows the slave even to participate in societal life, would restrict the slave's freedom. According to negative freedom, in this situation it is not directly obvious how in this non-interfering instance provides a direct restriction on one's freedom understood as non-interference (Pettit 1997). Yet neo-republicans emphasize that even a benign master – despite his 'goodwill' – remains a master. *Your* freedom, life, and choices ultimately depend upon the master's wishes and demands. Slavery is bad not just because of the potential negative consequences that the slave must endure, but because the master uses his power in an illegitimate manner, leaving the slave vulnerable to abuse.

Mill initially appears more strict on this matter, strictly opposing slavery precisely because a slave would be under the control of a master (Moloney 2011, 10). The possibility for someone to be harmed is problematic, and this potential harm ultimately justifies measures to prevent such harm. To this extent, Mill would also object non-interfering instances of slavery as it opens up potential harm for the slave. Given this, it appears unclear why a neo-republican perspective is fundamentally different.

The difference lies in the reason *why* non-interference must be preserved. For Mill, it is the possibility to engage in harmful injuries, and hence the focus is on harm. For neo-republicans, the ultimate complaint lies in the fact that those subjected to a master do not stand in an equal political relation. So in the context of slavery, for neo-republicans the problem is not so much the potential harm the slave might endure, which is a reason for Mill's harm principle to dismiss slavery, but more that the slave experiences a second-class status as a human being. As Pettit (1997, 47, see fn 7) notes: "they [liberals] may have been concerned with security in the sense of wanting to reduce involuntary risk (…), but not in the sense of wanting to reduce exposure to the power of another". Thus, where more liberal arguments against slavery are rooted in preventing harm, neo-republican arguments are grounded in the political relation between the master and slave.

In sum, neo-republican theory provides a suitable normative framework to inform responsible ML development as it captures well the moral wrong of being in a vulnerable power relation in its conception of freedom as non-domination. If domination is indeed a wrong to avoid, responsible ML development should concern itself not solely with the prevention of harm through interference, but also with the prevention of *unaccountable* interference, rooted in the power dynamics between the agents involved. This requires addressing the question, however, to what extent we can speak of unaccountable interference with regards to a ML system and its actors.

4. Power dynamics underlying a ML system

As already mentioned, the aim of this paper is not to provide an elaborative account of a conceptualization of power dynamics. Rather, the aim is to argue that neo-republican theory provides essential insights that should inform initiatives for responsible ML development. That said, I need to make some sense of the underlying power dynamics in order to provide substance to my claim. So, in the following

paragraphs I briefly discuss how the idea of domination applies to the power dynamics underlying the development and use of a ML system.

Elsewhere, I have argued that there is a power-dependence relation between developers, users, and end-users of a ML system (blinded for peer review). The power-dependence relation is formulated following the work of Cristiano Castelfranchi (2003; 2011), where you depend on Agent X if you lack the ability to realize your goal, yet Agent X does have this ability. This dependence, according to Castelfranchi (2003) automatically entails a 'Power-over' situation, and hence a power relation between the two agents involved. So in the case of a ML system, this implies that whenever end-users rely on a system to realize their goal, they depend on the system. Yet given that value-judgements necessarily inform the development of a system, a ML system is never purely objective and technical but necessarily advances into the social domain (blinded for peer review). This socio-technicality implies that a system's end-user does not solely depend on its technical capacity, but also on the social and political decisions made during the development and deployment phases. End-users that depend on a system therefore depend on those who shape the system as well[38].

But how does this power-dependence dynamic relate to the three conditions (dependence, power, and arbitrariness) of domination? The first condition, dependence, at first sight seems obviously defined within the power-dependence relation. Nevertheless, note that dependence condition for domination was understood in terms of exit costs. The end-user's dependence hence is rooted to the degree that the end-user wants to make use of the system. For instance, a person relying on social media platforms for social communication arguably has greater exit costs than someone who does not care for the service. This, therefore, relates more to the Hobbesian level of power, where dependence on a system is rooted in one's ability to *need* to use the system or not in order to satisfy their desired course of action.

Second, in order to constitute domination, the system and its shapers must exercise power over the end-user either directly through coercion or more indirectly through persuasion. In the case of Facebook's automated recommendation systems (ARS), these algorithms have the ability to influence people's preferences through the content shown on their newsfeed. There is hence some exercise of power at play. This is commonly the case with ML systems, although the actual influence of a system comes

[38] For instance, Facebook's end-users depend on an automated recommendation system (ARS) to organize their news feed, and Facebook develops and uses the system to which the end-users are subjected, resulting in a power-dependence relation between Facebook and their end-users via the ARS.

in degrees. Given the relation between a system and its developers and users, this exercise of power between the system and the end-user is indirectly linked to a relation between the developers and users on the one side and the end-users on the other.

Third, we must wonder to what extent ML systems allow for arbitrary exercises of power. If ML systems merely fulfil condition one and two, there is no particular reason why neo-republican theory is so beneficial to inform responsible ML development. However, ML systems increase the potential for arbitrary exercises of power (blinded for peer review). Non-arbitrary exercises of power require effective societal constraints, i.e. effective accountability mechanisms. Particularly in the case of ML systems, accountability mechanisms and procedures with regards to the developers and users are highly difficult to establish, both on a moral and legal basis, because of the system's opaque and self-learning characteristics.

For starters, the system's opacity makes it difficult to trace back the system's output (Lipton 2018). Indeed, even the developers themselves cannot always interpret the system's behaviour correctly. If it is unclear which developmental decisions were responsible for the system's output, it becomes especially difficult to ascribe accountability to someone[39]. Second, the system's self-learning characteristic creates

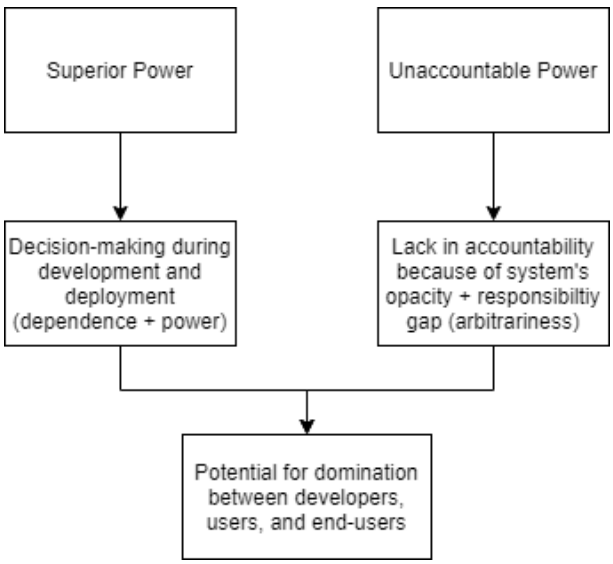


Figure 1: potential for domination of those who shape a system over those affected by the system

[39] I assume that a ML system itself cannot be meaningfully held accountable, and that accountability and responsibility will always (or at least as long as there is no 'General Artificial Intelligence') fall on humans.

a so-called 'responsibility gap', which implies that no individual can be held responsible and accountable for the system's output (Matthias 2004; Santoni de Sio & Mecacci 2021). Especially in situations where there are many different actors involved, accountability mechanisms are particularly difficult to develop. To this extent, the developers and users of a ML system do not yet experience effective and sufficient accountability mechanisms. This, then, leads to the potential for morally problematic power relations between those that shape a system and those affected by it, i.e. domination (see *figure 1)*.

So far, my argumentation has been relatively abstract. In order to ground it in the real world, I illustrate my argument on the basis of the Facebook-Cambridge Analytica scandal that shows that (1) a focus on interference is insufficient as (2) automated decision-making based on big data tactics reduces room for accountability.

5. Case study: Facebook-Cambridge Analytica scandal

In 2018, it was brought to light that the company Cambridge Analytica (CA) applied an algorithm based on Facebook user data to Facebook users in order to steer them into a particular political direction (Wylie 2019). The scandal, most often referred to in the context of the 2016 US presidential elections and the Brexit referendum, caused a public outcry related to people's privacy. Most Facebook users were not aware their data had been used for micro-targeting (i.e. individual profiling) to unknowingly steer individuals into a particular political direction. The idea behind this steering or 'nudging' (Thaler & Sunstein 2009) was rooted in big data analysis, where a machine learning algorithm identified people's political preferences, whether they were still debating which party to vote for, and if so, the algorithm would target these floating voters with biased advertisement (e.g., in the case of the US presidential elections, floating voters would be targeted with advertisements biased against the Democratic party and in favour of the Republican party).

The Facebook-Cambridge Analytica scandal is a prominent example in the field of AI ethics as it demonstrates well how ML systems harm democratic values through interference. The scandal has been greatly criticized for privacy violations and for distorting democratic communities through the privatization of people's political sphere, enabled through micro-targeting (Macnish & Galliot 2020, 4-5). The harm that was done with the interference surely are problematic and deserve the rightful attention. But besides the moral wrong of actual interference (e.g., privacy-infringements) there is a moral wrong of interference done in an arbitrary or

unaccountable manner which relates directly to underlying power dynamics[40]. This unaccountable interference was possible for three reasons.

First, Facebook arguably has a monopoly with regards to services offered by social media platforms (blinded for peer review). Due to network effects, there is a high collective burden to switch platform: the service only keeps its value if a significant number of people choose to switch. Facebook's monopoly implies that users, being limited in alternative social media platforms, depend more on the service Facebook offers as their exit costs ('no good alternative available') are arguably relatively high. Interpreting one way to hold Facebook accountable (i.e. public contestation) is by switching platforms, but this possibility for accountability is ineffective[41] precisely because of the fact that users are dependent on the service offered by Facebook and therefore have high exit costs. Other means, such as fines, are arguably similarly ineffective as it is unclear how this undermines the monopoly of Facebook, therefore upholding the dependency of Facebook users on Facebook. Second, Facebook interfered with its users without their consent and knowledge, which makes it simply impossible for Facebook users to hold the company accountable, e.g., through objecting that their data was used for profiling. Third, Facebook's ARS decrease room for accountability mechanisms as said systems further obfuscate current exercises of power (e.g., micro-targeting end-users). Indeed, the extensive use of big data tactics provide such influential companies with a strong tool to potentially abuse its users.

To illustrate how automation and big data tactics affect accountability, compare one's access to news via Facebook's automated recommendation system during the CA scandal and a political campaign before the internet. Political campaigns would target certain neighbourhoods more than others and go door by door in person (Macnish & Galliot 2020, 4). This appears similar to the individual advertising on Facebook, as both methods personally reach out to you. A major difference between these two situations is however that on Facebook you, as an individual, stand out, as opposed to the case before the internet, where the neighbourhood stands out.

This private vs. public realm is important as it affects one's ability to hold accountable the one responsible for the interference. For instance, if I were to disagree

---

40 As I mentioned earlier, AI ethics literature is limited because it does not consider other moral wrongs besides that resulting from actual interference. Unaccountable interference may beg the question, as it still is concerned with interference, yet the moral wrong of unaccountable interference puts more emphasis on *unaccountability* rather than the *interference*.

41 Note that Cambridge Analytica did go bankrupt after the scandal became public, which arguably is some form of holding the company to account and a great difference to the consequences Facebook faced.

with a campaign poster, I could protest its presence in the neighbourhood. Or, if I did not want to talk to someone at my door, I could request them to leave. Yet in the case of CA, people were *individually* profiled based on personal data. This moved the campaign from the public to the private realm (Macnish & Galliot 5), obscuring the fact that this micro-targeting was happening in the first place. This, in turn, made any control of the digital campaign (e.g., through contesting the advertisements) more difficult.

One might doubt the relevance of the lack of accountability mechanisms with regards to the CA scandal on the basis of two objections. First, one could object that there is no hard evidence that this micro-targeting did affect people's political preference. And if the interference did not have any affect, why should we care? This objection is in line with a consequentialist reasoning, where only the consequences of an action determine the rightfulness of the action. I do not find such reasoning convincing. For instance, when I shoot someone but miss, I will – and should – still suffer legal consequences for my action to shoot, even if it failed. Similarly, *even if* CA failed in its goal to nudge people towards a political preference, this does not mean the intentions behind the action should be disregarded. More importantly, similar to the fact that I should never have been able to pull the trigger, CA should never have been able to try nudging Facebook users into a political preference in the first place.

Second, someone could argue that Facebook users are still free *not* to use Facebook in order to avoid interference. While it certainly is true that people do not need to use Facebook, we must ask on which conditions this choice is based. This argument is two-fold. The first relates to the dependence relation between Facebook and its users based on their exit costs. Because of the network effects inherent to the service Facebook provides, Facebook users may simply not have an alternative to use. This leaves them with the choice either to use Facebook or to not use Facebook without a satisfying alternative. Though for some the exit costs of not using Facebook are acceptable, this need not be for all. For the latter group, their degree of dependence on Facebook is higher. Simply the claim that 'you do not need to use Facebook' hence proves to be an unsatisfactory solution.

To press this point further, imagine living in a male-dominated society in which sexual abuse, particularly sexual abuse of women, goes unpunished, resulting in women no longer leaving their houses out of fear for being harmed. Clearly, this is an undesirable situation for obvious reasons, but it is particularly morally problematic for women's freedom in choice. If women choose to stay home, they should not decide out of fear of being harmed but out of a preference to stay home. This requires *robust*

non-interference, something missing in their society. The same reasoning goes for deciding to use Facebook or not. This choice should depend *not* because you want to protect your own rights and avoid potential harm like privacy-infringement, but because you do not care for its service.

At first glance, the reasoning for the second objection seems to fit well with the negative conception of freedom, according to which one's set of choice options should not depend on one's actual choice. There is an important difference, however, namely that going out or using Facebook does not necessarily entail interference. The point is that whether you will be interfered with depends on someone else. This leaves one *vulnerable* to harm, but does not *necessarily* entail that the person will be harmed. This vulnerability, however, implies that my choices (i.e. leaving the house or not, using Facebook or not) become dependent on the arbitrary will of someone else (sexual abuse or not, non-consensual micro-targeting or not), and this hence constitutes the moral wrong of domination.

Thus, while both objections raise an interesting point, I do not find them convincing to dismiss my claim that there is a moral wrong with the CA scandal that goes beyond the harm caused by the actual interference.

With this example, I have tried to highlight that there is a morally problematic power relation between Facebook and its users which lies to some extent in the fact that ARS reduce the room for accountability. Nevertheless, as we also see with this example, domination depends on the interplay between the system and the company that developed the system (Facebook), as well as the company that used the system according to their own ends (Cambridge Analytica). However, domination is primarily between Facebook and Facebook users, as Facebook users ultimately depend on Facebook's service, as well as the fact that Facebook is the platform on which the exercise of power (privacy-infringements and manipulation of political preference) took place. In addition, the degree of domination depends on the degree of the exit costs of the Facebook users. For some, the costs of quitting Facebook were higher than for others, and hence they face a more intense dominating relation.

6. Towards Responsible ML Development[42]

Now, if we accept that domination is a moral wrong that is rooted in unaccountable power dynamics, it becomes clear why a predominant focus on preventing harm is

[42] In Maas (2023), I provide a more extended version of this argument.

unsatisfactory, as this focus does not necessarily tackle illegitimate power dynamics. Arguably, the liberal focus hinders current initiatives for responsible ML development from being effective. Indeed, both guidelines and proposed regulatory frameworks receive criticism either for ineffectiveness (Hagendorff 2020) or insufficiency (Ponce 2021). This criticism is predominantly rooted in the initiatives' lack of focus with regards to power dynamics. Developing effective accountability mechanisms, therefore, should aim to control the power of those who *can* do harm.

To illustrate some worries with current initiatives from a neo-republican perspective, consider how existing principles and guidelines – although essential to create awareness about potential risks amongst governments, AI developers, and society at large – have been criticized for their vague proposals to improve the situation as well as potential 'ethics washing' of companies' responsibilities (Hagendorff 2020). The concern goes that companies do not necessarily keep to these ethical guidelines and there are currently no mechanisms in place to ensure *and* enforce ethical behaviour (Resseguier & Rodrigues 2020).

The ineffectiveness of these guidelines have caused a shift from 'AI Ethics' to 'AI Regulation' (e.g., the recent proposal by the European Commission for an AI regulatory framework), to enforce a more top-down regulatory approach. However, these regulatory proposals also remain vague how regulation will effectively contribute to responsible ML development. According to the EC's proposal, so-called 'high-risk' AI applications require mandatory control checks, yet these checks are *internal* (except for biometric systems that do require third party checks), which raises some concerns. First, it is unclear how the EU will (effectively) monitor such internal control checks, and second, internal control checks can defeat the purpose. Just consider the Volkswagen sustainability scandal in 2015, or 'Dieselgate', where software mechanisms cheated emission tests (Jolly 2020). Internal control requires moral scrutiny, yet as Dieselgate shows, as well as the notion of 'ethics washing', there is no *guarantee* that these companies will always follow internal control checks or handle ethically. Too much focus on internal control mechanisms hence risks moving the debate from 'ethics washing' to 'regulation washing' as it remains unclear how the control is checked by the EC and whether companies will keep to this internal check-up.

Although AI guidelines and regulatory initiatives all implicitly point towards the lack in accountability mechanisms and despite the growing attention for the power dynamics regarding ML ecosystems, the main focus of these initiatives on the harm principle has diverted their attention away from developing workable

conceptualizations of these power dynamics. Yet developing guidelines and regulatory frameworks without such conceptualizations risks merely providing symptomatic treatment (e.g., technical fixes like algorithms that detect bias in data, see Gebru et al. 2018) whereas tending to the source of the problem provides a more robust and sustainable contribution to responsible AI development. Before we try to develop accountability mechanisms such as legal regulation, we should first establish means to ensure a legitimate exercise of power by those who the shape system on those affected by it.

Indeed, if we take a closer look at the current criticisms, we see that both guidelines and the EC's regulatory framework are enforced by the same party that exercises their power. Non-dominating ML systems require accountability throughout their entire software development cycle. This implies that decisions during each stage need not only be morally scrutinized, as the EC's regulatory framework proposes, but these decisions should be held to vigorous accountability procedures. In other words, the power with regards to AI decision-making procedures needs to be controlled in order to increase the room for accountability mechanisms. This is essential precisely because these systems decrease room for accountability themselves. Current initiatives are hence insufficient and ineffective as they do not address the power dichotomy between those who shape a system and those affected by it.

Thus, responsible ML development initiatives, such as principles, guidelines, and regulatory frameworks are limited in their effectiveness and moral robustness, since they are predominantly built on the ideal of freedom as non-interference and overlook to account for the moral wrong of domination. Although initiatives regarding responsible ML development are concerned with preventing harm, the initiatives spend little attention on why certain systems are developed in the first place and who granted the authority of these systems' development and use. This entails that responsible ML development should expand its normative horizon beyond Mill's harm principle towards designing systems that safeguard against arbitrary power relations.

7. Concluding remarks

In this paper, I have defended the argument that neo-republican theory should inform responsible ML development. Currently, responsible ML development initiatives focus too much on preserving the liberal tradition, in particular the aim to prevent harm through interference. This focus overlooks the problem of *unaccountable*

interference, which urges us to rethink of responsible ML development in terms of power rather than harm. Neo-republican theory provides a promising framework for responsible ML development thanks to its focus on power dynamics and its conception of freedom as non-domination.

My claim that responsible ML development should be informed by neo-republican theory has direct implications for developmental guidelines as well as regulatory frameworks. The criticisms related to the ineffectiveness and insufficiency of current initiatives gain more momentum with a neo-republican perspective. Besides the fact that ethics and regulation washing are problematic since they do not necessarily prevent harm, proposed solutions by current initiatives are insufficient as they maintain the power dichotomy between those who shape the system and those who are affected by it. Similarly, a neo-republican perspective implies that technical solutions are insufficient (though praiseworthy) efforts with regards to responsible ML development, precisely as they also do not address unaccountable power dynamics.

In order to mitigate potential dominating tendencies between developers, users, and end-users, effective solutions must cross the power dichotomy. This requires accountability mechanisms that go beyond ethical guidelines and internal control mechanisms. For instance, one might think of having public control during the development process to analyse the decisions made by developers. However, any concrete formulation of what such accountability mechanisms must look like requires further investigation, especially since the development of a ML is a highly complex process and hence analysing such a development from a neo-republican perspective risks frustrating other societal and economic values (e.g., public control may violate the intellectual property of a company[43]). Nevertheless, I hope to have shown in this paper an alternative approach on what elements are essential to consider when formulating accountability mechanisms to ensure responsible ML development.

[43] I thank an anonymous reviewer for pointing this out.

Interlude 4

Chapter 2 outlined the power dynamics underlying the development and deployment of AI systems. Chapter 3 discussed how AI Ethics builds on a negative conception of freedom, and therefore insufficiently considers other moral wrongs such as domination. In both chapters, I have primarily focused on direct power relations between the shapers and affected of an AI system. However, the broader societal context either enables or constrains such power relations, as this context provides certain agents with their position of power and with the ability to exercise their power in a specific way. The following two chapters elaborate on the need for incorporating the broader societal context. In the next chapter, co-authored with Jeroen van den Hoven, I provide the initial steps by clarifying the need to focus on the broader societal perspective. It answers the following question:

RQ4: Why should AI Ethics consider a broader societal perspective?

We argue that a broader societal perspective is essential to ensure that an AI system is developed in accordance with the best interests of the public. This is what Langdon Winner refers to as 'political ergonomics'—i.e., the fit between technology and society. As noted in the introduction (and I will expand on this in Chapter 6), non-domination requires power to be exercised in accordance with the best interests of those subject to that power. Although I do not make this claim explicit in the paper, a political ergonomic AI system—i.e., an AI system that tracks the best interests of society—is therefore aligned with the ideal of non-domination.

In order to ensure a political ergonomic AI system, we must avoid two things. First, we must avoid 'political naivete'. With political naivete, I have in mind developing a system with one specific ethical concern in mind that does not actually deal with broader issues in society. Think of a content-moderation algorithm that avoids echo chambers but which is deployed in a country that censors its citizens. Second, we argue for the need to embed AI systems in a politico-legal setting (or what we refer to as 'deontic provenance'). It must be clear how a system precisely came about, as this increases the possibility for citizen contestation, and therefore increases the likelihood a system will indeed be aligned with the public's best interest. Taking into consideration how a system is embedded in society (both by aiming to avoid political naivete and tracing a system's deontic provenance), it is thus more likely that the

system will in fact be political ergonomical (i.e., aligned with society's best interests) and therefore non-dominating.

Chapter 4: Opening the Black Box of AI, Only to Be Disappointed[44]

Abstract The common approach to AI Ethics is concerned with developing and deploying non-harmful AI systems by aiming to align the systems with commonly acknowledged ethical principles and values. However, this approach has been criticized for neglecting the broader societal context in which these systems and values are embedded. We argue that due to this neglect, the standard approach to AI Ethics remains normatively disappointing. For one, a narrow construal of AI Ethics risks some form of political naivete. Two, by leaving a larger normative context out of consideration it fails to show why we have moral reasons to act on the output of an AI system, everything considered. A normatively satisfying account of AI Ethics requires what Langdon Winner calls a 'political ergonomics' of AI in which larger politico-economic questions regarding society, politics, and power distribution are part and parcel of the lifecycle of an AI system. Integration of normative political theory and AI Ethics is therefore essential.

Keywords AI Ethics, politics of AI, deontic power, Langdon Winner, political ergonomics

1. Introduction

In his paper "Upon Opening the Black Box, Only to Find it Empty", Langdon Winner (1993) critiqued social constructivism. Social constructivism is a theoretical framework used in Science and Technology Studies (STS) to describe how a technological artefact has come into being in a social setting. This framework has provided a wealth of insights of how different stakeholders, perspectives, and interests are involved in the development and deployment of a particular artefact. It shows how artefacts have been shaped in and by social interactions, and at which inflection points in the history of the origin of technologies different turns in their development could have been taken.

While an important and valuable framework, social constructivism has its limitations. As Winner points out, the framework provides a historical and social science point of view that only *describes* the development of a particular technology, thereby neglecting the normative implications and assumptions underlying

---

[44] Maas, J. & Van den Hoven, J. (forthcoming). Opening the Black Box of AI, Only to Be Disappointed. In *Computer Ethics Across Disciplines - Applying Deborah Johnson's Philosophy to Algorithmic Accountability and AI*. Noorman, M. & Verdicchio, M. (Eds.). Springer Handbook.

technological innovations. Winner argues that, while it is surely helpful to open the black box of technology and understand how technological artefacts have come about, if we neglect the moral dimension of their development, important political and power aspects in their genealogy remain hidden.

Winner's observation regarding social constructivism contains an interesting lesson for our thinking about AI Ethics. AI Ethics aims to align AI systems with commonly acknowledged ethical principles and values, with Explainable AI (XAI) being one of the holy grails of the AI era. Ideally, XAI will realize other values such as accountability and safety as it allows for a better understanding of the system. Although AI Ethics clearly addresses moral dimensions and is therefore not as normatively empty as Winner considers social constructivism to be, numerous scholars have voiced concerns about the current state of AI Ethics (Mittelstadt 2019; Crawford 2021; Birhane et al. 2022; Gebru 2022; Selbst et al. 2022; Van Maanen 2022). In particular, much critique centres on AI Ethics' neglect of the broader societal and political context.

A vivid example of the worries of a narrow AI Ethics is portrayed by Keyes et al. (2019, 3). They discuss the disturbing algorithmic software that is supposed to manage overpopulation and food shortages by turning lonesome elderly into nutritious food such as "hash browns (Grandmash™)" and "bananas (Nanas™)," a process the authors refer to as "mulching". The authors apply the FAccT (Fairness, Accountable, and Transparent) framework in order to see how we can develop and deploy this system more responsibly. This satire points out how a narrow construal of AI Ethics can sometimes fail to address highly relevant questions. While most initiatives in AI Ethics may perhaps not leave us completely empty-handed, they often do leave AI Ethics in a normatively disappointing state.

This disappointing state can be characterized in the following two ways. First, in its current form AI Ethics is often politically naïve. We distinguish between a simple version and a complex version of political naivete. Its simple naivete quite straightforwardly neglects the broader societal perspective, such as when we praise volunteers and soup kitchens for the poor, without addressing rampant inequality in society. Complex naivete can take the form of limitations of conceptualizations of moral values. Common mistakes include a failure to take into account a moral dimension of a concept, such as when we work with a statistical notions of fairness, e.g., fairness as equalised odds, without taking alternative conceptions of fairness into consideration.

Another example is to ignore the wealth of fundamentally different conceptions of the value one is considering. Thinking about digital democracy requires consideration of democratic conceptions which each foreground different normative key concepts such as representation, information, contestation, deliberation, or participation. These different conceptions each point to different technological 'democratic' innovations and technological development trajectories. Simple and complex political naivete then risk developing ineffective methodologies and tools to achieve ethical AI.

Second, a narrow construal of AI Ethics is disappointing as its main focus is on 'what' the system does, i.e., whether it is aligned with ethical values and principles, whereas it neglects 'how' the system has been brought into society. Drawing on work by John Searle (1995), this 'how' question can be seen to be important as systems have a particular deontic status and are embedded in a normative context and in a network of deontic relations. The AI Ethics literature is less concerned with what provides these systems their status, and therefore what normative grounds we have ultimately to base our actions on the output of AI systems. In order to fully appreciate the moral reasons we may have for doing so, we need a clear picture of what we call the *deontic provenance*—i.e., the normative pedigree of AI's status in society—of the relevant AI we are confronted with. Mapping the deontic provenance of AI brings in view the normative dimension of how an AI system is embedded in a larger politico-legal context, such as the rule of law, human rights, and democracy.

To improve on both points, the naivete and the lack of a clear picture of the deontic provenance of AI, we need better alignment between society and technological innovations, or what Winner (1987) calls "political ergonomics" in the analysis and design of Technology. In Section 2 and 3 we discuss in more detail our two concerns with a narrow construal of AI. In Section 4 we propose that a political ergonomics along the lines suggested by Winner is much needed to open the black box of AI in a normatively satisfying manner, and in Section 5 we conclude.

2. Political Naivete in AI Ethics

The AI Ethics literature discusses moral implications of AI systems. It has spawned numerous ethical guidelines that have been developed to highlight important values and principles that systems should meet (e.g., transparency, autonomy-preserving, privacy-safeguarding, see Jobin et al. 2019; Hagendorff 2020; Floridi & Cowls 2022). These values and principles now feed into proposals for regulatory frameworks like

the AI Act currently under development by the European Commission (2021).[45] It also includes the immense collection of studies concerned with technical and practical attempts to design AI systems better aligned with values and principles that are often presented at AI Ethics conferences such as FAccT and AIES. Think of such more technical and practical attempts proposals to increase explainability (Gunning et al. 2019) and fairness (Hirota et al. 2022). These initiatives broadly aim to generate Responsible AI systems from a principlist standpoint, which according to the European Union's plans for AI governance and regulation imply technical robustness, compliance with extant law, and important moral values and human rights.[46]

However, as several scholars have remarked, the dominant approach to AI Ethics rooted in principlist ideals often remains narrow, shallow, and compartmentalized.[47] Deborah Johnson (2021) points out that striving for XAI by simply suggesting technological solutions does not capture the political, economic, and social complexity of a digital society. This results in the inability to achieve and realize values like accountability in practice. Others have argued that the literature fails to come to terms with the fact that a normatively satisfying approach to ethical AI requires institutional and structural change (Gebru 2022), and must position the development and deployment of AI—including the accompanying values—in its relevant societal context. Without a contextual awareness, we run the risk of falling into what Selbst et al. (2019) refer to as 'abstraction traps' (e.g., inaccurately applying a system designed for one context to a different one, or neglecting how new technologies affect and change societal behaviour and values).

Birhane et al. (2022) have empirically investigated these worries and claims. The authors annotated a total of 526 papers part of important AI Ethics conferences (FAccT and AIES) from 2018 to 2021 along the three following lines. First, they identified their abstract/concrete engagement with ethics (including concepts like bias, fairness, etc.) applied to a specific context. Second, they assessed to what extent

---

45 The link to the EU's plan to regulation shows how a particular conception of doing AI Ethics (i.e., a principlist one) feeds into and shapes regulatory measures.

46 Of course there are numerous approaches to AI Ethics, and some more aligned with what we believe a normatively satisfying account of AI Ethics would look like. Our criticism is specifically directed towards more principlist approaches to AI Ethics. In Section 4, we further discuss different conceptions of AI Ethics.

47 Such principlist ideals are characterized by the need to conform to certain principles such as explainability, fairness, or accountability, and are represented well by e.g., Floridi & Cowls (2022). The satire by Keyes et al. (2019) mentioned in the introduction highlights neatly the discomfort reasoning from isolated principles.

a FAccT or AIES paper mentions the disparate impact of AI. And third, in case of strong, medium, or weak disparate impact mentions, the authors assessed to what extent these disparate impacts were made explicit. Although more towards 2021 the authors noticed that the AI Ethics research increasingly connects to broader issues, their overall findings support worries and claims that the dominant AI Ethics literature as presented at FAccT and AIES remains rather abstract and conceptually ambiguous on ethical values and principles, and indeed insufficiently acknowledges societal context.

The current approach to AI Ethics, then, "[shrinks] ethics into one-dimensional problems," focusing strongly on ambiguous values and principles that get translated into technological design solutions (Gerdes 2022a, 2). In the mulching satire mentioned in the introduction, for instance, the authors propose to increase accountability by giving potential elderly-to-be-mulched a ten-second window to object the algorithm's decision (Keyes et al. 2019). 'Accountability' is then reduced to a ten-second window, and does not include the political, economic, and social complexity that Johnson emphasizes.

The example of installing a ten-second window shows how a narrow and compartmentalized construal of AI Ethics is problematic as a one-dimensional AI Ethics results in a shoppable 'checklist ethics' (Bolte et al. 2022) that can easily be repurposed and cherry-picked by more powerful agents in society. One issue with this is that it leaves AI Ethics in its current state prone to 'ethics washing' (Van Maanen 2022). However, it furthermore risks being politically naïve, which we can understand in at least two ways.

First, we can be naïve simply by overlooking the broader societal context in which the AI system is developed and deployed. For example, the whole point of developing a solution for the filter bubble and content moderation issue on social media platforms would be undermined by the system being deployed in a context in which the state practices internet censorship. Similarly, we can wonder what 'democratic design' meaningfully contributes if a system will be deployed in a dictatorship. Consider efforts in sustainability. Clearly, an airline company will not make much headway on sustainability by asking passengers to bring their own re-usable coffee mugs. Such 'micro-focuses' steer towards a solutionist ideal—one where technology can fix all issues (Morozov 2014), if only they are designed ethically. However, these are examples of non-solutions for much larger and more complex political concerns. Studies of professional codes of conduct and micro-ethics issues will thus not get us far if they do not address macro questions as well.

Second, there are also more complex versions of being politically naïve related to the ambiguity of principles and values. For instance, we may focus on preferences as economists do, but forget to ask how our preferences are manipulated, falsified, and adapted. These then fall short of serving as reliable indicators of what people have reason to value. This would for example ignore Steven Lukes' (2021[1974]) third dimension of power—i.e., the power to shape preferences, desires, and will formation—that he distinguishes from his first and second dimensions of power to affect decisions and the power to set agenda's, respectively. Without explicit attention to this third dimension—which seems particularly pertinent to critique the role of social media in Big Nudging and computational propaganda and cognitive warfare—one could responsibly and carefully take stakeholder's views into account without considering how these views have been shaped by the dominant socio-technological infrastructures.

Another form of complex naivete occurs when conflicting conceptions of one concept are insufficiently considered. Maas (2022), for instance, has argued that AI Ethics remains too narrowly focused on Berlin's negative conception of freedom that is aligned with Mill's harm principle (i.e., you should be free to act as long as you do not cause harm to anyone). However, when we take on a different perspective on 'harm', one can be morally wronged even when one is not interfered with actively. Such moral wrong—known as 'domination' in contemporary political philosophy of a neo-republican type—relates to a different conception of freedom than the more traditional freedom as non-interference commonly associated with liberal democracies. A change in what precisely constitutes harm matters for how we go about designing systems. We must hence be weary of complex naivete, given that how we conceptualize specific values and principles influences how we design and develop technology (Veluwenkamp & Van den Hoven 2023).

Both forms of political naivete draw our attention away from the fact that law, ethics, and AI are all already situated in a particular complex socio-political, economical, and institutional setting that may undermine or support mechanisms to protect and further democratic ideals. Indeed, technology is situated in a particular societal context, being shaped by and shaping society (Winner 1980; Bijker 1995; Johnson 1997). Given this interplay between technology and society, democratic values are at stake when we fail to reflect critically on the complexity and dynamics of an (increasingly globalized) society. Consequently, such failure may potentially result in an inability to translate ethical insights into meaningful pragmatic methodologies and tools (Gebru 2023; Carnegie Council 2021).

Worries about compartmentalization in applied ethics have already been voiced in other areas such as sustainability development and climate justice. Lehtonen (2004, 212), for instance, argues that we need a "coevolutionary framework" to resolve difficult trade-offs in sustainability development, and Simon Caney advocates for the "method of integration" over the "method of isolation" (Brandstedt 2014, 73). As Caney argues, whereas isolationism detaches the debate around climate change from other normative considerations such as poverty, the method of integration advocates a holistic approach that includes other concerns besides climate justice. Climate change can, for instance, result in an increase in climate refugees, so we may need to re-evaluate our migration rights (Brandstedt 2014, 74). Both Lehtonen and Caney recognize that different spheres of life inevitably influence each other, and contend that viewing conflicting normative concerns in isolation therefore makes a methodological error that will never truly lead to normatively satisfying solutions.

AI Ethics currently misses such a co-evolutionary framework and integrationist approach. As is increasingly shown by scholars in the field, AI Ethics initiatives including the guidelines, regulatory proposals, and the FAccT and AIES literature have remained on a level where they do not provide us the bird's-eye perspective necessary to address the larger political, social, economic, and legal concerns. And so, we can be normatively disappointed in the current state of the art in AI Ethics due to both its simple and complex political naivete.

3. The Missing Debate on *Deontic Provenance*

There is a second reason why the neglect of the broader societal picture gives us reasons to be normatively disappointed. AI systems are inevitably intertwined with our social lives and institutions (Gabriel 2022). They complement or replace human agents, producing a particular output we act upon. In other words, drawing on Searle's (1995) theory on the construction of social reality, we are constructing the social world with AI. Whereas AI Ethics is seriously concerned with 'what' makes AI systems ethical, there is less of a focus on 'how' the system has come to be part of our social world. However, this 'how' is fundamental. In Searle's terms, constructing our social world necessarily grants these systems an assigned and accepted 'deontic status', which, as we claim, co-determines the moral and practical reasons we may have to act on its authority. As will become clear in this section, in order to normatively assess our reasons to act in the company of AI systems, we thus need to know the *deontic provenance* of the relevant digital entity.

This requires representing the entity's normative pedigree, which cannot only be done via opening the black box of the system by means of internalist perspicuous representations such as XAI provides. It is also a matter of understanding of how the black box was delivered on our doorstep, where the system got its status and authority from, which is an externalist perspicuous representation. Before we look at the quality of the content of the box, let's inspect the bill of lading. We do not mean to suggest that the proposed broadening of the scope replaces the work that is done in the field of XAI and should be abandoned. Yet to anticipate the next section, we just suggest that work on an internalist representation needs to be supplemented along the lines suggested by Winner's political ergonomics, which meaningfully situates a technology in a normative analysis with a broader scope.

Searle (1995) points out that certain entities (objects, individuals, or groups of individuals) in society have a particular moral standing because the function of their status comes with certain normative or 'deontic powers', i.e., they have certain rights, duties, and obligations that are directly linked to their particular status function. Think, for instance, of the exclusive right the President of the United States has to order the use of nuclear weapons. Or, a scrap of paper with a distinctive ink pattern printed on it that counts as money. A status function with deontic powers, or 'deontic status', is imposed on the piece of paper, and from there on the paper figures in a network of deontic relations of rights and permissions. There is a collective acceptance, recognition, and acknowledgement of the deontic powers of relevant actors and institutions, e.g., the national bank, to confer such status functions and to turn scraps of paper into legal tender.

In his work on status functions and deontic powers, Searle leaves aside the complex societal background that supports the conferral of this status or status function imposition. However, this provides a limited perspective on the *normative* construction of our social world, in particular our socio-technical world, as a status function with deontic powers has some form of authority that gives us reasons to act on it. After all, when we act on a twenty dollar bill, we act on the fact that it counts as *money*, not on the fact that it has written 'twenty dollars' on it. Without its status function, we would have no reason to give back change, for instance. The conferral of deontic status is thus relevant to normatively assess our relation to the status function.

To make this more concrete, consider a railway timetable. We can study a piece of paper with rows and columns of train numbers, departure and arrival times and stations, etc. In and of itself such a piece of paper is practically useless and lacking deontic authority. However, if the timetable was issued by the proper institutional

party (e.g., British Rail) with the official institutional powers associated with it, the timetable now suddenly gets its deontic authority from the fact that it is the *official* timetable.

Clearly, there is a difference between 'just' a piece of paper with columns and rows vs. the 'official' timetable, even if they would have the exact same information on it. If we would find both papers on the ground, we would base our actions on the official timetable because it is more evidently linked to other status functions, such as the British Rail. Indeed, *because* it is the official timetable, we have moral and practical reasons to act on it. Practical, because it is a more reliable indicator of the British Rail's timetable. Moral, because if we were to share information with others, we ought to share the information presented by the *official* timetable. There are thus certain entities of which we need to know their *deontic provenance*—i.e., the pedigree of the entity's deontic status—in order for us to determine their authority and reliability.[48]

We can understand deontic provenance along the lines of the inheritance principle. Formally put, the inheritance principle implies that "if A inherits ϕ then there is some Δ that is a source of A's ϕ-ness (i.e. A inherits ϕ from Δ and no entity among Δ inherits ϕ)" (Trogdon 2018, 5). Taking money as an example for ϕ, Schaffer (2016, 95, footnote and emphasis omitted) writes that

(…) a grounded entity inherits its reality from its grounds, and where there is an inheritance there must be a source. One cannot be rich merely by having a limitless supply of debtors, each borrowing from the one before. There must actually be a source of money somewhere.

In the context of Searle's work on the construction of the social world, this means that there must be money that is accepted as legal tender, there must be a contract and normative expectations that borrowers pay back their loans, etc. Indeed, money does not come from nowhere, it requires a particular normative context and is embedded in a complex network of deontic relations. Deontic provenance is then to make explicit how status function A inherited its deontic status ϕ from which source of 'higher' deontic status Δ.

Just as in a piece of paper with rows and columns there is nothing intrinsically authoritative, there is nothing in the output of an AI system—or in the technology producing it for that matter—*by itself* that can explain why we would have moral and practical reasons to act upon it. To consult it multiple times, or to consult a slightly modified version of it will also not help. It will have to get its deontic status from

[48] Of course, there are other reasons to act on an entity's inherent properties. A comb cannot be used as a chair, for instance, whereas a tree stump may.

elsewhere in order for us to have moral reasons to act on its output. Indeed, it needs a source of higher deontic status.[49] It is one thing to look at an AI system in the legal or health care system and ascertain that it has certain desirable properties of transparency and fairness, but we also need to know whether it has been approved to play a particular role in decision making in these contexts. Without such a status function imposition no amount of assurance concerning its moral properties justifies its utilization.

We find this additional normative dimension has received insufficient attention in the AI Ethics literature. Of course, some work points in this direction. For instance, scholars like Binns (2018) and Dobbe et al. (2021) have voiced their worries about the social power of AI developers,[50] and have suggested ways to ground this power by building on political ideals like public reason and deliberation. Approaches that aim to ground developers' power clearly hint that something is 'off' with development and deployment of AI. By more clearly delineating the deontic provenance of AI systems, we may well better understand precisely *what* is off. This should then also shed light on whether the approaches such as those suggested by Binns and Dobbe et al. in fact resolve the issue.

In addition to other regulatory measures such as the GDPR, the AI Act more directly aims to sketch out the deontic provenance of an AI system. The Act asks AI developers to provide extensive record keeping and documentation, so that we can "assess the compliance of the AI system with the relevant requirements" (European Commission 2021, 30). The idea, here, is that if the system complies with the relevant requirements mandated by the AI Act, we can trace back its deontic provenance via compliance with the Act to the European Union, an established political institution. In other words, documentation is essential to show that the system is backed up by a

---

49 Some clarifications are in order. First, we refer in particular to 'high-sensitive' AI systems, such as those deployed in legal, medical, financial, and public governance sectors. Second, as suggested in footnote 2, there are other reasons to act on a system's output. A high accuracy, for instance, could give us moral and practical reasons to act on a system's output. Accuracy, however, focuses on how the system itself behaves, yet not on how it was developed and deployed. A system's deontic status, then, gives us a distinct set of moral and practical reasons to act. Third, where a deontic status gives us certain reasons, deontic provenance allows us to assess whether we have these reasons. We thank Herman Veluwenkamp for discussion on these points.

50 With social power of AI developers, we refer to the fact that developers make certain design choices that influence the system's behaviour, which in turn has an effect on the system's end-users. There is therefore an indirect dependency relation between the system's end-user on the developers (via the AI system) in the sense that end-users to some extent depend on the developers how the system will behave. We consider this dependency relation a social power relation. See Maas (2023).

politico-legal institution, where an AI system inherited its deontic status from the AI Act, which in turn inherited its status from the European Union.[51]

In theory, the AI Act can thus be an effective mechanism to outline the deontic provenance of AI systems. In practice, however, we see the AI Act is criticized for being too heavily influenced by industry (Perrigo 2023; Perarnaud 2023; Schyns 2023; Galvagna 2023) and for being negligent in oversight regarding the fact that AI developers can decide which risks are acceptable (Burri 2022; Laux et al. 2024). These criticisms point to a further direction regarding deontic provenance, namely whether the AI Act in fact provides a suitable normative source for AI systems from which to inherit their deontic status.

Indeed, some important yet underdeveloped questions in the literature in this line of research are (1) what does it mean for AI systems to have a normatively substantiated and well-founded deontic provenance, and (2) are those requirements currently met? Although these questions are beyond the scope of this chapter, we are encouraged that taking into consideration the broader societal context can in fact shed light on these issues. For now, we merely wish to point out that AI Ethics is too strongly fixated on 'what' makes a system ethical, thereby neglecting the 'how' the system's deontic status was assigned, accepted, and preserved.

4. Towards a normatively satisfying account of AI Ethics

As we have seen, we have at least two reasons to be normatively disappointed in AI Ethics as typically practiced today for its neglect of the broader societal context in which AI systems are necessarily situated. First, it risks being politically naive as it may overlook important questions at the heart of democratic societies, which ultimately

[51] This raises at least three further questions. It requires us to ask, first, where we begin with the source of deontic power; second, whether deontic provenance can be traced back to a single source; and third, whether 'deontic provenance' is simply conflated with 'lawfulness'. We believe deontic provenance need not have one source. For example, there are also national-level regulatory measures in place that provide some politico-legal embedding of AI. The AI Act would be an additional one. The source, however, would be distinct: one on a national basis, one on a EU basis. In addition, deontic provenance need not be embedded by legislation. Informal political representatives that are not necessarily 'elected' by means of formal selection procedures but rather by bottom-up audience uptake (think of Dr. Martin Luther King in the context of American segregation or Greta Thunberg for climate change) can provide a source of deontic provenance as well. The link between source and legislation is something to be further explored for a full account of deontic provenance, which is unfortunately beyond the scope of this chapter. We thank Merel Noorman for pressing us on these issues.

may lead to developing ineffective methodologies and tools. Such fixes could be pointless simply because macro-issues are left unaddressed, or because they focus on the wrong conception or leave out important normative dimensions of a conception. Second, it centres too much on how a system behaves, rather than on the politico-legal background that gives us moral reasons to act on the system's output. The question then is how to make AI Ethics normatively satisfying.

Winner's (1987) "political ergonomics" provides a promising vantage point for a method of integration combining a plurality of perspectives. With *political ergonomics*, Winner refers to the 'fit' between technology and society. Each artefact applies to different social relations between citizens and institutions, has a different setting in the physical environment (and therefore a distinct effect on its surroundings), and so each technology raises different technical problems. Such a triadic view of political, spatial, and technical is essential in order to correctly assess how well technology and society fit together.

Winner (1987, 26) applied his approach in the eighties to the development of the card computer, a small computer the size of a credit card that made certain activities and transactions such as shopping or banking move from a physical realm to a digital one. Studies, however, warned against how such digitalized practices "pose a threat to privacy and freedom" and create a world of electronic dossiers in which "persons have no control over the kinds of information routinely collected about them and no control over how such information is used" (Winner 1987, 27).

One response by David Chaum in the 1980s to these concerns was an alternative design of the card computer which was supposed to preserve people's privacy, freedom, and autonomy. Under this alternative approach, privacy is respected by means of "digital pseudonyms" (Chaums 1985 as cited in Winner 1987, 29), where individuals preserve identifying data such as their names or home address. This method of pseudonyms would allow banks and shops to combat fraud, while the individuals would remain in control over their own personal information.[52]

At first glance, Chaum's card computer seems perfect for preserving individual freedom and privacy, and therefore contributes to democracy overall. However, Winner shows how Chaum's laudable intentions for his alternative version of the card computer lack normative theoretical grounding. As Winner (1987, 38) points out, Chaum "assumes that privacy is crucial to freedom and that freedom is essential to democracy", but he remains ambiguous on the distinction between public and private,

[52] See for further technical elaboration Chaum 1985 and Winner 1987.

and what conception of democracy he favours. Indeed, do we want a more Rousseauvian form of direct democracy that favours participation, or rather one more aligned with a republican type where one's civil liberty stands central. Moreover, Winner asks why we need freedom and individual privacy as Chaum's card computer would facilitate, and how they would further a healthy and strong democratic society. As the reader is convincingly shown, it most certainly matters that these questions are answered before the actual development of such a technological artefact. Designing for democracy, indeed, requires a sufficiently fleshed out understanding of democracy to be feasible and desirable.

Similarly, Johnson (1997) implicitly applies such a political ergonomics approach when she questioned the oft-assumed democratic character of the Internet. She shows that "being a democratic technology" is an ideal that is easily misinterpreted when we more broadly analyse how and which values are embedded in the technology. Again, ambiguity and lack of conceptual clarity risk idealizing a technology while missing a more substantial and fundamental impact on society. As Johnson (1997, 24-25) argues, for the internet to be a democratic technology we need to have a perspective on the historical roots of its development, the institutions, practices, and people who maintain the technology. Moreover, we need to know about the materiality of the technology, and the cultural meaning of the values that are embedded. For each of these understandings, our perspective may change regarding the question of whether the technology is democratic. And so, only by considering the larger picture can we prepare a judgment on design requirements.

As we see, both Winner and Johnson recommend combining political theory, history, economics, philosophy of technology, and science and technology studies in order to assess the artefact's political ergonomics, i.e., the fit between either the card computer and society or the internet and society. They approach the technology from a plurality of perspectives, and show that depending on how one approaches a technology, also the approach to its design, development, deployment, and regulation will change. They therefore urge us to rethink the artefact itself completely.

Implicitly, the approach taken by Winner and Johnson draws attention to larger societal questions as well as to the provenance of a technological artefact. Part of the solution to the political naivete and the lack of attention to deontic provenance of AI, then, is a 'political ergonomics' of AI that charts the deontic provenance of socio-technical systems. This can provide us with a co-evolutionary framework to adequately assess the fitting interface between AI and institutions that reflects the politico-legal background in which AI is necessarily situated. Such an approach must

consider the complex interplay between contextual background, the shaping of AI, and in turn the shaping of society.

For this, integration with different fields like normative political theory, STS, media studies, and political economy are essential. They provide the conceptual tools necessary to navigate large and complex societal issues by zooming out and incorporating questions related to historical background, social power, authority, and legitimacy essential to situate AI systems in their political context. An emphasis on AI's situatedness addresses concerns related to political naivete and helps trace their deontic provenance.

To some extent, we see such integration already. However, specifically the integration of normative political theory has fallen short. For a long while digital ethics has disregarded political theory as a fruitful source of normative guidance, and political theory has omitted digital ethics as a topic worthy of serious reflection (Van den Hoven 2017[1994]; Gabriel 2022). Questions related to injustices and violations of human rights to scholars concerned with ideological critique of the digital era were therefore often left to political economy (e.g., Zuboff 2019) and media studies (e.g., Rushkoff 2010). Yet like digital ethics, media studies and political economy perspectives miss the conceptual tools of normative political—most relevantly authority and legitimacy—to adequately capture some of these core problems of AI systems. Consequently, despite some broader integration with other fields, questions related to authority and legitimacy have not received the attention they deserve (Lazar 2023).

Fortunately, political theory increasingly shows more interest in the digital ethics side. To name a few, Van den Hoven (2017[1994]) has built on Rawls' account of justice to inform our understanding of justice in the information society, and Oosterlaken and Van den Hoven (2011) have used Sen and Nussbaum's capability approach to analyse how digital technologies impact our capabilities. Sharon (2021) has applied Michael Walzer's spheres of justice to online platforms, Habermas (2022) has extended his own theory of structural change of the public sphere to the impact of social media and online platforms, and Susskind (2022) has used the neo-republican conception of freedom as non-domination to argue for a 'digital republic'. Lazar (2023) is developing a theory of communicative justice, and we now even have an introduction to political philosophy of AI (Coeckelbergh 2022) and a political theory of the digital age (Risse 2023).

In addition to these more normative political theory related-approaches, other perspectives from critical, feminist, and decolonization studies can be helpful for a

political ergonomics of AI (see e.g., Waelen 2022; Adams 2021; Mhlambi & Tiribelli 2023). These theories question the power asymmetry of society at large, thereby necessarily taking on a broader perspective.

Finally, we can think of more practically-oriented approaches. In the policy-domain, for instance, we see a shift in perspective as well when it comes down to AI. The future of work in particular is severely being challenged by the increasing automation thanks to AI. The European group on Ethics (EGE) (2018) has suggested that these changes may require a radical shift in how we organize our society, and advocates a redesign that puts solidarity centre stage. Such a redesign would require us to think about, for example, a universal basic income or about "contributive justice", which suggests that we owe to each other that each can make an appreciated contribution to society.

Other more practically-oriented approaches concern our approach to the design of technologies. Here we also see how such practical approaches build on democratic ideals. Think of the work done on design justice by Costanza-Chock (2018), DiSalvo's (2022) book *Design as Democratic Inquiry*, and Forestall's (2022) work on building community in digital environments by focusing on designing for democracy. These works urge us to look beyond merely meeting certain ethical principles, but to design artefacts with and for a community. This is then different from the example mentioned in the introduction of the algorithm aimed to identify to-be-mulched elderly. Rather than designing this system according to fair, accountable, or transparent principles, this system simply would not have been designed if we reasonably consider the communities wishes, needs, and desires.

As Rawls' famously pointed out in his Theory of Justice, we can distinguish between concepts and conceptions of justice. What we all agree on is that AI Ethics has an important role to play, yet there are important differences in ways of how to conceptualize AI Ethics. In this chapter, we have been sceptical of narrow conceptions of AI Ethics and have argued for a broad conception. Broad conceptions question the politico-economic background that informs AI development, deployment, and even ethical assessment. They aim to address political naivete by placing the technology in its larger societal context, and by doing so they contribute to the elucidating picture of AI's deontic provenance. These approaches seem to be aware of the political framework AI systems are embedded in, and precisely *question* this framework as well where necessary. Broad and narrow conceptions need not to exclude each other, but as we saw in Section 2, broader conceptions of AI Ethics are underrepresented in the FAccT and AIES literature.

Although the increase in attention in recent years regarding a more comprehensive approach to ethical analysis is promising, more work can and should be done. We must focus on developing systems from a comprehensive and societal broad approach that situates ethical issues in a larger political-economy, *cum* ideological assumptions. This may serve to keep us from political blunders and naivete such as fixing filter bubbles in countries with state censorship, arguing that democratic design is the solution in a country run by a dictator, or insisting on privacy-by-design in countries without checks and balances and democratic control over the use of personal data by intelligence agencies. Ideally, more comprehensive meta-level approaches to AI systems will avoid such unfortunate non-solutions and reduce potential for ethics washing. Approaches that fall under broad conceptions of AI Ethics move beyond a focus on morally acceptable *systems*, and include a morally acceptable shaping of society in a broader sense. Only with a broad conception of AI Ethics, then, can we escape the normative disappointing framework when we open the Black Box of AI.

5. Concluding remarks

In conclusion, what we find upon the efforts of opening the black box of AI as AI Ethics currently does, while not normatively empty, is disappointingly meagre in a normatively relevant sense. It fails to take into account how much the quality of ethical debates is determined by their being embedded in a larger political, economic, and legal system. Just like we need a certain distance to be able to identify what impressionist paintings like those of Monet or Van Gogh depict, we need to approach ethical AI from a certain distance in order to get an adequate impression what moral and political concerns are truly at stake. Otherwise, we risk simple and complex political naivete and lack a clear picture of AI's deontic provenance, without which it is unclear what our reasons are to respect the output of the system. To strive for moral progress should ultimately be the goal of any innovative endeavour, including those in the field of AI. We therefore now turn to our fellow digital ethicists to not (just) ask whether a system's behaviour is 'ethical', but also ask how it is normatively constructed, how it ended up—and with which deontic credentials—at the time and place when and where its output is acted upon and makes a difference in practice.

Interlude 5

The previous chapter argued that in order to design a system that is well-aligned with the public's best interests (i.e., a political ergonomic AI), it is key to consider the broader societal context in which an AI system is developed and deployed. When we can trace a system's deontic provenance (i.e., the historical lineage of an AI system development and deployment), it is clear to citizens who designed the system and who were involved in the decision-making processes. Such 'process transparency' (Zerilli 2020) in theory facilitates public contestation, which in turn improves public empowerment as contestation provides the means to shape technology according to one's interests. However, meaningful contestation is only possible under the appropriate democratic and participatory conditions. As I argue in the next chapter, initiatives in AI Ethics that aim to empower AI stakeholders, however, lack such appropriate participatory conditions, precisely because of limitations of the broader societal context. The chapter, co-authored with Aarón Moreno Inglés, addresses the following question:

> **RQ5:** How can we move beyond current initiatives in AI Ethics in order to successfully address the power relations between the shapers and affected underlying AI development and deployment?

We provide two suggestions how to move beyond current initiatives in AI Ethics. First, public empowerment requires a shared distribution of decision-making power. As we show in the chapter, empirical research on participatory AI initiatives shows, however, that stakeholders still depend predominantly on the developers and deployers to be included in decision-making procedures with regard to AI development and deployment. Regulation and legislation such as the AI Act do not mandate a form of stakeholder inclusion, implying that whether stakeholders are included in the decision-making process rests on the developers and deployers. We propose that regulation must ensure that AI developers and deployers are *forced* to track the best interests of society as a way to control the power of AI developers and deployers.

Second, decision-making power rests within the ownership model of a specific company or technology. In current societies, this entails that ownership rests in capital holders and less so in stakeholders contributing to a particular system. Drawing on the Marxian concepts of relations and means of production, we propose that truly

empowering participatory AI must address forms of ownership in a way that democratizes the production chain of AI development.

Chapter 5: Beyond Participatory AI[53]

Abstract The 'participatory turn' in AI design has received much attention in the literature. In this paper, we provide various arguments and proposals to move the discussion of participatory AI beyond its current state and towards stakeholder empowerment. The participatory AI literature points to Arnstein's understanding of 'citizen power' as the right approach to participation. Although we agree with this general idea, we argue that there is a lack of depth in analyzing the legal, economic, and political arrangements required for a genuine redistribution of power to prioritize AI stakeholders. We highlight two domains that the current discourse on participatory AI needs to address more. These are (1) the legal-institutional background that could provide 'participation teeth' for stakeholder empowerment and (2) the political economy of AI production that fosters such power asymmetries between AI developers and other stakeholders. We conclude by offering ways forward to explore alternative legal arrangements and ownership models for participatory AI.

Keywords Participatory AI, power asymmetries, stakeholder power, AI ownership

1. Introduction

In the past few years, the literature on AI Ethics has seen an increased interest in democratizing AI, often through a participatory approach to AI design. Different forms of participatory mechanisms have emerged to align AI systems with stakeholder values more effectively and to empower stakeholders by allowing them to resist and contest the diverse socio-technical risks and injustices posed by AI-powered technologies. These risks have already been conceptualized and discussed, including questions of racism and bias (Akter 2021; Gebru 2020; Noble 2018), privacy breaches (Bartneck et al. 2021), and mass surveillance trends (Feldstein 2019), amongst other issues. Multiple scholars have engaged with the 'participatory turn' in AI to address these risks (Delgado et al. 2023; Birhane et al. 2022; Donia and Shaw 2021; Costanza-Chock 2020; Denton et al. 2020; Cammaerts and Mansell 2020; Rahwan 2018). These scholars build on participatory design schools such as the Scandinavian tradition and Value-Sensitive Design that have placed democratic participation in the design process at the centre of the discussion as the key to addressing technological risks and stakeholder empowerment.

---

[53] Maas, J. & Moreno Inglés, A. (2024). Beyond Participatory AI. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society (AIES '24), October 21-23, 2024, San Jose, CA, USA. ACM, New York, NY, USA.*

Participatory AI perspectives have also faced certain criticisms. Sloane et al. (2022) denounce participation-washing and "exploitative and extractive forms of community involvement" in participatory design practice for machine learning systems. Birhane et al. (2022, 3) warn of the risk of reproducing colonial dynamics through the legitimation of unfair power structures "under the veneer of participation", in the same way how former colonizers "co-opt[ed] rules and authority structures, [...] turning participation in governance as a form of colonial power." Himmelreich (2022) objects to participatory AI design on the assumption that calls for democratizing AI are based on weak normative grounds and therefore do not justify the costs associated with participation. The general line of reasoning these critics object to is that (1) democracy (i.e., citizen empowerment) involves citizen participation, and (2) therefore democratic AI (i.e., stakeholder empowerment) requires any form of stakeholder participation in AI.

Our aim in this paper is three-fold. First, we argue that although the general participatory AI literature points to stakeholder empowerment as the right way of participation, the analysis of the legal, economic, and political arrangements proves too superficial to establish a redistribution of power that genuinely prioritizes stakeholders. Second, we want to justify calls for participatory AI despite the costs. We do so by objecting to Himmelreich's argument that participatory AI rests on normative weak grounds. Third, we highlight two domains the current discourse on participatory AI has insufficiently explored. These are (1) the legal-institutional background that could provide what we call 'participation teeth' for stakeholder empowerment and (2) the political economy of AI production that fosters such power asymmetries between AI developers and other stakeholders.

Based on our analysis, we agree with the critics that participatory AI is, as the literature currently stands, flawed. As numerous scholars in political theory have acknowledged, citizen participation is not a sufficient proxy for citizen empowerment (Young 1990; Cohen 2009). Democracy requires citizen freedom and autonomy, which requires effective control over democratic processes (Pettit 2012). Moreover, such freedom and autonomy are entangled with the political economy (Gould 1990; Sen 2001). We conclude, however, that we should not be 'against' participatory AI, as Himmelreich claims, but we should move beyond the current participatory AI discourse. To address the legal-institutional background, novel regulatory processes and external auditing could help to mitigate unjust power asymmetries. To address the political economy of AI production, we argue in favour of exploring alternative ownership models of the AI design and production processes.

The rest of the paper is structured as follows. Section 2 provides a brief overview of the state of the discussion around participatory AI. In Section 3, we narrow in on one specific aspect of Himmelreich's critique of participatory AI and explain that, from a socio-technical perspective, there are good reasons to value participatory AI despite the costs it brings. In Section 4, we present our two criticisms that our characterization of participatory AI fails to address. Section 5 discusses how we can move beyond participatory AI to address our concerns through better regulation and analyzing alternative AI ownership models.

2. Participatory AI State of the Discussion: Intrinsic Goals, Co-optation, and "Citizen Power"

Participatory AI is a much-debated topic, partly because researchers have conceptualized it in various ways. The overarching discourse suggests that participatory AI implies the involvement of affected stakeholders during the development and deployment process of AI systems. In this section, we provide an outlook on the state of the discussion.

First of all, the work by Delgado et al. (2023) serves as the stepping stone to understanding the diversity among theories of participatory AI. To assess the different perspectives, they propose a distinction between four "dimensions of participation", these are the goals ("why is participation needed"), the scope ("what is on the table"), the participants ("who is involved"), and the methods ("what form does participation take") (Delgado et al. 2023, 3-4). These four dimensions indicate the different grounds for consensus and disagreement within the participatory AI literature. They also do an important job of highlighting the multiplicity of schools of thought involved in participatory AI research, including "user-centered design, service design, participatory design, co-design, and value-sensitive design, as well as participatory action research, participatory democracy (including deliberation theory), social choice theory, and mechanism design" (Delgado et al. 2023, 2).

From our perspective, the best dimension to synthesize the state of the discussion in participatory AI literature lies in the study of goals pursued by proponents of participatory AI design. We can distinguish between two trends in participatory AI, depending on the goal of the participatory mechanisms. We define these as instrumental vs. intrinsic goals. Participatory AI projects that pursue an instrumental goal do so to achieve more useful, better-functioning AI systems. We could think of a system that reflects outcomes that better represent stakeholders' values (Muller and Liao 2017; Gerdes 2022b) or outcomes that increase "profit or legitimacy for the

designers" (Delgado et al. 2023, 3-4) through participatory mechanisms. Although we believe the instrumental aim is justified, in this paper, we are primarily interested in the intrinsic value of participatory AI.

For proponents of participatory AI pursuing intrinsic goals, "participation is important because inclusion of the people who may be impacted by the technology [...] is simply the right thing to do" (Delgado et al. 2023, 4). In this sense, stakeholder empowerment entails citizen empowerment. The intrinsic value of participation in AI design lies in realizing certain political values, such as equality and freedom. Here, participation aims to empower stakeholders by including them in the design process (cf. Greenbaum 1993; Schuler and Namioka 1993). The need for empowerment is two-fold. First, AI systems affect our lived experiences in multiple ways. Therefore, people subjected to these systems should have a say in supporting self-determination and autonomy. This conception of participatory design represents calls to democratize technology more generally (Sclove 1995). In addition, participatory design can address structural power asymmetries between those that develop and deploy these systems and those affected (see e.g., Birhane et al. 2022; Sloane et al. 2022; Zimmermann, Di Rosa, and Kim 2020; Kalluri 2020; Costanza-Chock 2020; Gould 2019).

One of the main risks that the proponents of participatory AI for its intrinsic value have faced is institutional co-optation. Under the flags of democracy and participation, state actors and corporations aim to benefit from a general legitimation of their investments in AI technologies. Birhane et al. (2022, 6) state that corporate actors can act under the header of participation but use "the results of participatory efforts towards corporate benefits." As a result, these actors can "capitalize on such efforts and build products that maximize profits, with little benefit to communities remains open." Such potential misuse leads to participation-washing, where the ideal of empowerment masks private gain. As the authors elaborate, this form of participation-washing fits into a broader trajectory of using participation to legitimize corporate practices stemming from the Western colonial legacy (Birhane et al. 2022, 6).

To avoid co-optation and foster the pursuit of participatory AI, Sherry Arnstein's (1969) ladder of participation has become very effective. The ladder metaphorically models citizens' involvement in decision-making processes into different ladder rungs, each representing a degree of participation. At the lowest rung (non-participation), citizens have no influence or voice in decision-making. They are mere spectators, educated or manipulated into supporting the decisions but excluded from the actual

design process. As the rungs are climbed, a level of tokenism is reached, in which participation serves a symbolic purpose. Citizens may be consulted, but their inputs hold little weight—just like receiving a token gesture with no genuine impact. At the highest level, Arnstein explores what she calls 'citizen power', representing different stages of meaningful participation. Citizens can collaborate with authorities, sharing the agenda-setting and final decision-making power; they can have substantial influence by having institutional authorities delegate their decision-making power to them, or, at the highest rung of the ladder, they could actively shape decisions by controlling the totality of the decision-making process. The higher rungs of the ladder represent what Young (1990) has described as a just and healthy democratic order.

In the context of AI, an understanding of 'citizen power' is more nuanced. Since AI systems imply transnational networks and data streams, we do not refer to citizens of a particular region or state but to various stakeholders affected by the deployment of AI systems. Which stakeholders are affected depends on the context of deployment and the reach of the AI system. AI systems deployed in narrow contexts, such as the medical field, will demand less for achieving stakeholder empowerment than AI systems deployed in broader contexts, such as public governance, where all citizens of a society are affected, which in turn demands less than AI systems with a global reach such as ChatGPT.

Arnstein's ladder of participation has seen support in the participatory AI literature (e.g., Birhane et al. 2022), and we agree with the general idea that a genuine commitment to empowerment requires providing stakeholders with actual decision-making power. However, although we can point to Arnstein's 'citizen power' as the right way of participation, we argue that the participatory AI discourse lacks depth in the analysis of the legal, social, and political arrangements to achieve a real redistribution of power that prioritizes stakeholders. Avoiding participation-washing and moving towards real control would require, for instance, the possibility for stakeholders to refuse the development of the AI system in the first place, which is often not part of the options stakeholders can have when asked to participate in AI design (Delgado et al. 2023). Challenging the way AI systems are produced and deployed requires challenging the legal apparatus and the economic monopolies that uphold power asymmetries between the developers of AI systems and the rest of the stakeholders in the first place, a point of discussion that is not often present in the participatory AI literature.

The key takeaway of this section is that the intrinsic value of participatory AI is only realized at the highest levels of Arnstein's ladder, through citizen power (which

participatory AI translates to stakeholder power). However, this empowerment process requires rethinking the institutional arrangements that foster unequal power relations between AI developers and other stakeholders. Like any democratic endeavour, such genuine commitment to stakeholder empowerment is resource-intensive and costly. In the next section, we argue that participatory AI design is not unjustified despite these costs.

3. Justifying the Costs of Participatory AI

One of the most prominent criticisms of participatory AI design is found in the paper 'Against "Democratizing AI"' by Johannes Himmelreich (2022). Unlike the title suggests, Himmelreich does not oppose 'democratic AI'. He endorses a deliberative approach to AI governance and argues against claims that more participation is necessarily better. We agree with many of his objections, such as the moral desirability and practical feasibility of calls to participatory AI where everyone should be involved (cf. Zimmermann, Di Rosa, and Kim 2020). However, while we do not argue that more participation is necessarily better, we do hold that some participation is better (and necessary) for stakeholder empowerment, provided it is done in the right way (i.e., according to Arnstein's 'citizen power'). We support a more representative version of participatory AI design, where not everybody is involved, but representative groups are. We thus still categorize ourselves as falling within the group that endorses participatory AI, and we wish to provide some pushback to some of Himmelreich's claims to justify why some participation is necessary for stakeholder empowerment.

Himmelreich rejects the calls for participatory AI design on different grounds, supporting the notion that democracy as an idea "is not well-equipped to afford a proper response to problems of injustice and oppression" (Himmelreich 2022, 1342). One of his arguments is that calls for participatory AI have weak grounds and goes as follows. The idea of participatory AI is that it serves as a general legitimation mechanism for using AI. Legitimation via participation, however, requires justification due to the high costs of participatory democracy. Amongst others, these costs include the fact that participation is generally economically expensive and time-consuming and may lead to worse epistemic results. One way to justify these costs is when there is a presence of coercive power, pervasive impacts, and involvement in cooperative systems. However, Himmelreich argues that in the case of AI these conditions for democratic legitimization do not hold. AI systems do not create new forms of coercion and are only coercive in already coercive sectors (e.g., public

governance, judicial); do not have pervasive impacts (only the underlying societal structures do); and do not extend or create involvement in cooperative systems. Participatory AI, Himmelreich concludes, thus rests on weak normative grounds and does not justify the costs of participation.

While we agree with Himmelreich that participation costs are high, we believe these costs are worthwhile for climbing the ladder of participation. We specifically reject his argument based on coercion and pervasive impacts. First, we disagree that AI does not coerce. Himmelreich discusses coercion in a way where an authority can force you to do something (e.g., a state forcing you to pay taxes). AI systems coerce differently. They do not coerce us into doing a certain action or meeting a specific goal; rather, they coerce us in that we—citizens in modern, digital societies—are forced to be subjected to these systems. AI systems have become so fundamentally entangled with our society that we cannot function as we do today without these systems. Core societal sectors such as medical, financial, and judicial have increasingly adapted to and incorporated AI systems. If we focus specifically on one AI application, we can agree with Himmelreich that 'no new coercion' is created. However, zooming out to society at large, our modern societies are structured so that we cannot opt out of AI anymore, as these systems have become deeply engrained in core societal sectors. Taking on this broader perspective, we can say that AI coerces citizens in modern digital societies in a different, broader sense.

This broader understanding of coercion is found along the lines of Gabriel (2022), who argues that AI forms part of the background structure of society. It is also what Sclove implies with his assumption that "technologies profoundly affect and partly constitute those circumstances" (Himmelreich 2022, 1334). Himmelreich (2022, fn. 17) explicitly rejects Sclove's assumption as a reason to democratize technology. However, this rejection misinterprets the meta-analysis Sclove provides on technological influence. Sclove (1995, 16-17) writes:

Technologies function politically and culturally as social structures by coercing physical compliance; prompting subconscious compliance; constituting systems of social relations; establishing opportunities and constraints for action and self-realization; affecting nonusers; shaping communication; psychological development; and culture generally; and constituting the much of the world within which our lives unfold.

Rejecting the assumption that technologies profoundly affect and partly constitute people's lives disregards the fact that we do not live in a social world but in a socio-technical world. Although AI 'coercion' in our interpretation of Gabriel and Sclove

differs from the common understanding of coercion, Himmelreich's claim that AI does not coerce thus does not hold from this broader perspective.

Second, and relatedly, given the interconnectedness between society and technology, we cannot attribute the pervasive impact only to societal structures. In doing so, Himmelreich treats AI as a computational artefact and assumes we can isolate the technological artefact in question. However, as many authors have already argued, we cannot separate the computational artefact of AI from its social context, including its developers and deployers (Johnson and Verdicchio 2024; Noorman and Swierstra 2023; Kudina and Verbeek 2019). For instance, such a socio-technical understanding of AI recognizes that values become necessarily embedded within a system during design and development, embedding and reinforcing these values in society. For instance, defining a 'fair' system matters for how it affects specific groups. Given that these systems are inevitably part of society, to dismiss participatory AI based on the resource intensity of democracy and the lack of a well-grounded foundation overlooks these socio-technical aspects. Thus, we find that calls to democratize AI design through participation are not as unfounded or weakly motivated as Himmelreich suggests.

Although we are sceptical of some of Himmelreich's points to reject participatory AI, we do support many of his claims that much of the debate remains on a superficial level. Aligned with his claim that "[t]he call to "democratize AI" should not emphasize only participation" (Himmelreich 2022, 1344), we contend significant concerns remain with the current discourse on participatory AI that ultimately hinder the possibility of citizen power à la Arnstein. First, it primarily focuses on what stakeholders can contribute and how that empowers them. Not only should we ask what a stakeholder can contribute, but also on whose terms the stakeholder can contribute. Second, the debate overlooks how political-economic structures uphold relations of power between developers and stakeholders. The current discourse then remains limited as it overlooks the broader institutional and economic structures that inform AI development. In the following section, we address these concerns.

4. Moving beyond Participatory AI: Two Arguments

In this section, we present two relevant arguments towards participatory AI. While some criticism of participatory AI has been closely or loosely related to our criticisms, we focus on two points missing from the debate. First, we argue that the participatory AI discourse should consider the arbitrariness of the power relations between

developers and those affected. Second, we argue that participatory AI should consider a political-economic perspective incorporating power relations beyond the developers and end-users.

4.1. The Current Participatory AI Discourse Insufficiently Considers Arbitrary Power Relations between Developers and Affected Stakeholders

Our first argument concerns the power dichotomy between, on the one side, the developers and deployers of an AI system and, on the other side, the people affected by the system. We follow Arnstein in the call that participation–and hence participatory design–does not necessarily empower the participating agent. However, rather than focusing on ineffective participation not empowering an affected stakeholder, we focus on what this lack of empowerment means for the participatory agents' relation with the developers and deployers of AI systems.

First, let us clarify the power dichotomy, which we can understand in at least two ways. Some scholars focus on the fact that the groups facing the negative consequences of AI systems are already marginalized groups. Moreover, these groups facing the negative consequences differ from those shaping these systems, which typically belong to better-off demographics (Eubanks 2018; Noble 2018; Benjamin 2019; Kalluri 2020). Therefore, this understanding of the power relation is specifically concerned with existing power asymmetries in society that reflect marginalized vs. non-marginalized groups (male/female, white/black, etc.).

Another way to understand the power dichotomy is that these systems exercise power over the people subjected to the system. For instance, decision-support systems affect whether someone is allocated social benefits, and recommender systems show what people see. In this way, these systems exercise social power over the people subjected to the systems. This power relation extends to the developers and deployers via the AI system because of the sociotechnical aspect of these systems, such as value-embedding during design (Maas 2023). Therefore, this understanding of the power relation is specifically concerned with the relation between the developing and deploying agents vs. those affected by the system. These people constitute the 'relevant stakeholders'. We are specifically interested in this second relation of power.

This second relation of power resembles governing power, where 'to govern' refers to implementing and enforcing the norms of the relevant institutions (Lazar 2024). Regarding such governing power, we must ask whether the power is exercised in a legitimate way and by the right authority. This raises the questions of (1) what makes

exercises of power legitimate and (2) what constitutes the right or 'proper' authority? (Lazar 2024) One approach to justify governing power is through direct democracy. Zimmermann, Di Rosa, and Kim 2020) provide clear support for this approach. They write that "[r]ather than allowing tech practitioners to navigate the ethics of AI by themselves, we the public should be included in decisions about whether and how AI will be deployed and to what ends." They thus specifically highlight the need for participatory approaches to AI design. However, although we believe there is value in public control, the current calls for participatory AI design are insufficient to effectively address the power dichotomy between the 'shapers' and the 'affected'.

Our reason is rooted in the arbitrariness of the power relation, or what is known in political philosophy as domination. Being subject to arbitrary power means being at the whim of someone else's will (Pettit 1997; Lovett 2010). Think of a benevolent dictator. Even though the dictator treats their subjects well, they still remain a dictator. In other words, subjects living in a dictatorship have no guarantee or control about how the dictator will treat them. The dictator may be benevolent now, but this does not provide the subjects with a guarantee that they will be treated well in the future. This lack of guarantee shows the moral wrong of domination: it leaves people vulnerable to abuse (Pettit 1997). Thus, to meaningfully address the power dichotomy between the 'shapers' and the 'affected', this power relation must be non-arbitrary.

Participatory AI design draws heavily on stakeholder-friendly approaches such as the historical Scandinavian tradition of Participatory Design (Robertson and Simonsen 2012) and Value-Sensitive Design (VSD) (Friedman et al. 2013). PD originated as a way to emancipate the people working with new technological innovations, and VSD was primarily a way to improve technological innovation. Both approaches aim to achieve this by providing stakeholders with a meaningful say by means of stakeholder inclusion during the design process. However, these approaches have a well-known challenge of accounting for power (Bratteteig and Wagner 2012; Friedman et al., 2021). Particularly in VSD, this challenge includes (1) which stakeholders are 'relevant' and should be invited, (2) how developers should include stakeholder input, and (3) how stakeholders can ensure their input is considered appropriately. These three questions raise highly complex questions, such as what defines a 'relevant' stakeholder, who gets to define what is 'relevant', and who should oversee the final decision. As the key challenge of accounting for power is prominent in the VSD literature, these questions have yet to be answered satisfactorily. Given that decisions regarding stakeholder inclusion rest on the shoulders of the designers themselves, VSD is thus prone to arbitrary power relations. When participatory AI

design draws on existing design approaches that have such arbitrary relations of power, participatory AI design inherits these limitations.

The problem with participatory AI design resembles a familiar issue with voluntary ethics guidelines. Voluntary commitments to ethics guidelines are notorious for being ineffective (Hagendorff 2020) and lacking "teeth" (Rességuier and Rodrigues 2020). Similarly, suppose I ask you to participate but can decide at any moment to exclude you or choose which of your input suggestions is sufficiently relevant to consider. In that case, the participation lacks "teeth" and may even become a form of participation-washing. It may look good on paper but does little to equalize power asymmetries between shapers and affected or hold accountable the superior power of the shapers of AI systems. Indeed, the distribution of power remains the same as without the participation. After all, what can people who participate do to ensure their feedback is considered? And what can people who were not invited to participate do to be invited? Just like voluntary ethics guidelines are insufficient, so too for 'direct' or 'participatory' AI design, where the power of participation lies in the hands of the developers. In other words, as long as the power of participation remains in the hands of the developers, this power is arbitrary.

While comparing AI developers to dictators may seem extreme to some, there is an important analogy to be made. A dictator treats subjects well at the dictator's whim, and developers have the people participate at the developers' whim if participation remains voluntary and up to the developers/company's decision. If one believes the relationship between a dictator and his subjects is morally objectionable, this would hold–at least in a less extreme sense–for the relationship between developers and the people affected if the people's say depends on the arbitrary will of the developers. While critics of participatory AI design worry about the potential to repeat colonialist tendencies (Birhane et al. 2022) or exploit stakeholders (Sloane et al. 2022), there seems to be little attention to these existing concerns of participatory design approaches that already face objections regarding the voluntariness or arbitrariness of stakeholder inclusion. Himmelreich (2022, 1344) states that we need to address the "shortcomings of existing democratic institutions." Integrating non-voluntary commitments to participation in institutional legal arrangements is one way of doing so.

In sum, the power relations between the developers and deployers on the one hand and the people affected on the other depend on institutional arrangements. We must seriously consider how these institutional arrangements are shaped to see on whose terms participation happens. We need participation teeth if we are to address arbitrary

power relations successfully. In Section 5.1, we discuss how we can address arbitrariness. But first, we discuss our second critique of current participatory AI discourse.

4.2. The Current Participatory AI Discourse Does Not Sufficiently Address the Political-Economic Mechanisms That Uphold Power Asymmetries

The first argument that has been introduced deals with the power dichotomy between, on the one hand, the developers and deployers of an AI system and, on the other hand, the people affected by the system. Our second argument is that the participatory AI discourse lacks a wider political-economic focus exploring the social structures that uphold these power asymmetries. Without a political-economic outlook addressing how such power is distributed and reproduced, we cannot effectively ensure stakeholder empowerment.

As we have explained in our first criticism of the participatory AI discourse, power relations do not exist in a vacuum; they depend on checks and control mechanisms, such as by means of legal institutional embedding. Similarly, these power relations are shaped and upheld by social norms and economic structures, such as markets, that systematically reproduce arbitrary power (Haugaard & Pettit 2017; Gädeke 2020). In this sense, AI developers hold power over AI users because they take up a specific role in organizing the "production, distribution, and consumption of resources," referred to as the political economy (Mosco 2009; Wasko 2004). Thus, this power is rooted in economic relations, and an in-depth analysis of such power relations in the AI development process calls for studying the political economy of AI. There have been growing interdisciplinary efforts to set out a political economy of AI within the discussions of digital capitalism (Srnicek 2018; Zuboff 2019; Dyer-Witheford et al. 2019; Fuchs 2019; Luitse and Denkena 2021). We argue that the participatory AI literature has been disconnected from these research lines and does not sufficiently consider a political-economic outlook, which is necessary to deal with power asymmetries and ensure stakeholder empowerment.

There are two Marxian concepts through which power asymmetries become explicit, which can help us explore the political economy of AI. First, there are relations of production, a term that responds to the different ways people are organized for the production of goods. Second, there are the means of production, which relates to the ownership and control of the productive assets of society, allowing the production of such goods (Marx 1859).

On the one hand, power becomes explicit in the relations of production, which entail "the dynamics of the labour market, particularly in the process of workers entering and exiting firms, and in the operation of those firms themselves" (Bryan 2020, 120). These relations of production include arbitrary power exerted in the workplace and workers' lack of freedom to change or quit their jobs. Søren Mau (2023) refers to this as the economic power of capital or 'mute compulsion', the social domination that regulates the conditions through which capital is reproduced.

In the context of artificial intelligence, the analysis of the relations of production should be twofold. First, it should critically look at the companies and state actors involved in the extraction of lithium and other minerals (necessary for AI systems to function) and the exploitation to which its workers have been subject, following extractivist and neo-colonial logics (Crawford 2021; Birhane et al. 2022). Concerns about dangerous manufacturing and resource-intensive logistics should also be considered here (Dauvergne 2022). There are doubts about how effectively workers can opt out of the production of AI technologies at different points of the production process. Second, it should reflect on the material conditions of workers training the AI models (Perrigo 2023), as well as data extraction concerns, especially in the context of large language models. Such a reflection on the material conditions of workers would also impact workers in other sectors, such as people dedicated to artistic work. While participatory AI proponents might consider workers affected by AI technologies that companies deploy, they often fail to amplify the scope to consider the conditions and desires of workers beyond those directly involved in producing AI technologies.

On the other hand, relating to the 'means of production', within a capitalist economic system, the different structural relations of domination are "rooted in the distribution of control over the means of production," which is "generally granted to the owners of productive assets" (Bryan 2020, 128). In the case of AI technologies, it becomes difficult to define the boundaries of what counts as the means of production because of the complex ecosystem of organizations in which it emerges. Data, for example, could be a productive asset, as well as the networks of underwater cables, GPUs, or other computing systems, depending on how it is framed. Different analyses have highlighted that the key means of production of AI corporations is the massive computing power necessary for it to work, as this is what allows these companies to drive competition (Srnicek 2018; Dyer-Witheford et al. 2019; Luitse and Denkena 2021). The ownership of this variety of assets is a determinant of understanding the power configuration in the AI design process.

In the context of AI, most of the decision-making power of the design and production of global AI is in the hands of a small corporate oligarchy. As Muldoon (2022) comments, the "once disruptive and dynamic start-ups have grown into powerful monopolies able to use their resources to lobby regulators and implement laws to suit their own needs." In terms of total investment in AI systems deployment and research, companies like Google, Amazon, and Microsoft are way at the forefront, overtaking venture capital firms. In 2023, the deals amount "to two-thirds of the $27bn raised by fledgling AI companies" (Hammond 2023). Rethinking the distribution of the means of production thus entails challenging the accumulation of power of big-tech corporations.

To sum up, our second argument on why the participatory AI discourse fails to address power asymmetries lies in its very limited analysis of the political economy of AI, i.e., the ways AI systems are produced, distributed, and used are a byproduct of the capitalist economic infrastructure. Power emerges from this very core: the relations of production and the ownership and control of the means of production.

5. Suggestions and Ways Forward

Based on our previous criticisms, we propose two preliminary ways to move beyond participatory AI. First, there is a need to realize 'participation teeth' that address concerns regarding voluntary commitments to stakeholder inclusion. Second, we explore alternative AI ownership models to address the political economy of AI better.

5.1. Realizing 'Participation Teeth'

To successfully climb the ladder of participation and effectively empower affected stakeholders of AI systems, we need to reduce the arbitrariness of the power relations between a system's developers and its stakeholders. Doing so requires addressing the fact that decisions regarding stakeholder inclusions are in the hands of the developers.

One way of addressing the shortcomings of existing democratic institutions is by means of regulation. The AI Act is the most advanced regulatory initiative for AI that has been proposed (European Parliament 2024). The AI Act is anticipated to have wide-reaching consequences, including those beyond AI systems that affect European citizens. The EU provides a large consumer-base for many AI systems, which will likely have at least a limited 'Brussels effect' (where non-EU players come to conform to the EU regulation despite not being part of the EU) (Siegmann and Anderljung 2022).

However, the AI Act needs to be more specific on stakeholder inclusion. For instance, it states that "[w]hen identifying the most appropriate risk management measures, the provider should document and explain the choices made and, when relevant, involve experts and external stakeholders" (recital 65), or "[w]here appropriate, to collect relevant information necessary to perform the impact assessment, deployers of high-risk AI system […] could involve relevant stakeholders" (recital 96). These statements leave much room for interpretation and voluntary inclusion to the provider and deployer of the AI system, potentially rendering stakeholder inclusion relatively meaningless. A potential solution to mitigating arbitrary power relations could be, for instance, to make it mandatory to include stakeholders in the design process.

There are many obstacles to incorporating participation into legislation in a way that forces developers of AI systems to include stakeholders. Should the law specify a specific number of stakeholders to include? Should it already indicate which stakeholder groups are relevant to include depending on which AI application? Both seem infeasible and undesirable. The context of development and deployment is highly relevant. Consider the medical field. Different countries, even different hospitals in the same country, operate under different assumptions regarding diagnosing, treatment, and more. How should a general AI Act address these differences? Moreover, it is an obstacle to controlling how stakeholder input can best be considered. Do the developers write a reflection report after their discussions with stakeholders? In the case of designing an AI used for diagnosing in healthcare, it seems unlikely we can identify a set of measures that fits different cultural and context-specific preferences. So, how can we go about forcing AI developers to include stakeholders?

First, it is important to recognize that it is unnecessary to state precisely which stakeholders must be invited and which not. What can be incorporated, however, is that an independent party (either a body of members or one individual) is mandatorily (i.e., by law) included in the design process with special training in stakeholder inclusion. This person can suggest which stakeholders are relevant to include and how many and serve as a control mechanism to ensure that the developers take stakeholder input into consideration.

Second, people must know whether they can be included and how their input will be included. Developers must disclose the information discussed and the input received from the stakeholders and show how they incorporated it. The involved stakeholders must have the opportunity to contest if they disagree about how the

developers considered their input. However, the most important thing at these stages is that this is independently checked (the independent party that supports the developing team with stakeholder inclusion could undertake this check).

5.2. Exploring Alternative AI Ownership Models

To successfully challenge the political-economic structures that reproduce power asymmetries, we need to rethink how the productive assets related to AI are distributed and owned. This is not an easy task, as it is difficult to determine the scope; where AI starts and finishes is not clear-cut. If we aim for a small scope, we are at risk that our participatory proposals will not be sufficient to address power asymmetries, and Arnstein's 'citizen power' will not be meaningfully achieved. If we aim for a global scope, we encounter that redesigning the whole AI production cycle would be tied to wider democratic endeavours and political processes. However, we can start by providing different proposals that highlight the value of rethinking ownership regarding AI.

The literature about platform cooperativism already provides a good starting point. Coined by Trebor Scholz (2016), this term encompasses proposals surrounding "new ownership models for the Internet" to establish democratic governance. It starts with multi-stakeholder and worker-owned cooperatives in digital labour platforms. In his work, Scholz calls for reproducing the technical capabilities of platforms that Big Tech corporations typically own towards ownership systems rooted in solidarity and the redistribution of benefits. In this sense, platform cooperativism "technological, cultural, political, and social changes" beyond the mere change of legal ownership structures, but a wider reframing of the meaning of innovation and efficiency "with an eye on benefitting all" (Scholz 2016, 14).

On similar grounds, there are other pluralistic approaches to the social ownership of AI, like James Muldoon's proposal of platform socialism. This term describes "multiple and overlapping associations" that promote collective ownership models of digital assets "from mutual societies to platform cooperatives, data trusts and international social networks" (Muldoon 2022). This "complex ecology of organizations" could only be supported by a radical rethink of the scale in which the public could meaningfully own them. Muldoon suggests four levels of subsidiarity to work with: local, regional, national, and global. Ownership of digital assets would then ideally be managed at the smallest possible level to achieve more direct democratic control. Although a clear roadmap about how to achieve platform socialism still needs

to be outlined, this could serve as a good solution towards the previously mentioned issue of scaling democratic governance.

Other ownership alternatives try to benefit from the affordances that certain technologies could provide. Such is the case of Decentralized Autonomous Organizations, "DAOs" (Spelliscy et al. 2024), which use blockchain systems to design technical solutions that improve cooperative governance. Amongst the features they work on achieving are an effort to establish effective voting systems for the stakeholders, increase member engagement and accountability, improve the predictability of compensations, and better organizational transparency, amongst others.

6. Conclusion

In this paper, we have provided various arguments and proposals to go beyond the current state of the discussion in participatory AI. First, we have argued that the general participatory AI literature needs to consider the broader institutional and socio-economic context as a necessary foundation to establish the right way of participation that leads to Arnstein's 'citizen power'. Second, we have argued that participatory AI has a normative justification, particularly when considering a socio-technical perspective on AI systems. Third, we have indicated that the current state of participatory AI discourse needs to address power asymmetries between developers and stakeholders more extensively. We have highlighted two domains that need to be explored more in the current discourse for participatory AI. These are (1) the legal-institutional background that could provide 'participation teeth' for stakeholder empowerment and (2) the political economy of AI production that fosters such power asymmetries. While we have provided initial ways forward to explore alternative legal arrangements and ownership models for participatory AI, these lines of research require further investigation.

Interlude 6

The previous two chapters expanded on the relevance for AI Ethics to consider the broader societal context as this context provides certain agents (i.e., shapers and affected) with their particular position of power and their ability to exercise their power in a specific way. Based on claims provided in chapters 1, 2, and 3, I have further suggested that freedom as non-domination is a relevant conception of freedom to inform responsible AI development. The conception of freedom explicitly draws attention to power dynamics underlying AI development and deployment. In doing so, it proves to be a more robust source of freedom than the more common conception of freedom as non-interference. Incorporating the value of non-domination into the design process of AI systems—or what I refer to as *design for non-domination*—thus proves suitable to safeguard freedom in the digital age. In order to understand precisely how this value can be successfully incorporated during the development and deployment of AI systems, however, I must first provide a coherent account of digital domination. The aim of the next chapter is to do precisely that. It answers the following question:

RQ6: How should we understand domination in the digital age?

I propose three different ways to understand how domination by means of digital technologies comes about. I argue in favour of the socio-economic perspective that connects to the 'lawless cyberspace' discussed in the context of Zuboff's surveillance capitalism in the introduction of this dissertation. The lawless cyberspace supports a certain freedom to innovate. The combination of novel technological innovations that lack concrete regulation (such as the digital realm) with a (neo-)liberal ideology that strongly supports technological innovations results in an uncontrollability and vulnerability for citizens of modern, digital societies (nicely captured by Facebook's slogan of 'move fast and break things'). Unlike Zuboff, however, I identify the root problem not in the *consequences* of the lawless space to innovate (which led to surveillance capitalism), but rather in the lawless space to innovate itself. Consequently, digital domination is a *symptom* of socio-economic structures.

Drawing on a structural account of domination, I distinguish between interpersonal domination (i.e., when subjected to an AI system), and systemic domination (the domination embedded within these socio-economic structures that give rise to a vulnerability). Thus—to anticipate one of the conclusions of this dissertation—designing for non-domination requires ensuring that new technological

innovations, such as in the field of AI, are forced to track the best interests of society. With this conclusion in mind, we can now make sense of digital domination.

Chapter 6: Making Sense of Digital Domination[54]

Abstract Scholars are increasingly concerned with domination in the digital sphere, particularly in the Gig Economy and Big Tech context. Although this work has been highly insightful, digital artefacts such as automated-decision systems used in core societal sectors should also be considered a source of domination. I propose three analyses to understand such digital domination based on an interactional, marginalized, and socio-economic perspective. From the interactional perspective, digital artefacts dominate due to arbitrary interference in one's basic liberties. I argue this perspective is insufficient for similar reasons raised in the paradigmatic 'mugger' case, namely, who is dominated and when. A potential solution is the marginalized perspective, according to which only the groups that systematically face negative consequences are dominated. I reject this perspective as a source of digital domination. Instead, I propose a perspective based on our current socio-economic market structures that leave innovators relatively unrestricted in developing and deploying automated-decision systems. While not digital in itself, such innovative domination only applies to markets that are relatively unrestricted, such as the digital sphere. On this view, digital domination thus is a symptom of our underlying socio-economic order.

Keywords digital domination, AI systems, republicanism, basic liberties, structural domination

1. Introduction

Digital technologies in general—and platforms such as Google, Facebook, and Uber in particular—are notorious for their concentrated and uncontrolled power. This has led to concerns about algorithmic or 'digital' domination as understood in the republican tradition (Aytaç 2022; Muldoon & Raekstad 2022; Susskind 2022; Muldoon 2023; Hoeksema 2023). According to this tradition, an agent is in a dominating position when they have the power to interfere in another agent's choices in an arbitrary or unchecked manner (Pettit 1997; Lovett 2010). As the debate currently stands, much of the literature has been primarily focused on how the economic, social, and political power of online platforms produce various relations of domination (Aytaç 2022; Muldoon & Raekstad 2023). What is lacking, however, is a sustained argument for why AI systems in core societal sectors (henceforth 'core AI

[54] An extended abstract of this paper has been accepted to the student track for the AIES'24 conference and is forthcoming in the AIES'24 proceedings.

systems') such as healthcare, finance, judicial, and public governance dominate. Especially given that they are used in decision-making processes to allocate various goods like social benefits and healthcare.

Susskind (2022) is one exception, who discusses digital domination in the context of such AI systems. However, Susskind moves between online platforms and core AI systems, which does not provide a coherent account of why these systems count as dominating. Moreover, in his positive proposal to mitigate digital domination, he focuses solely on social media platforms, and doing so offers little guidance on how we can mitigate digital domination beyond online platforms. The result is that it remains unclear whether it is coherent to hold that AI systems in core societal sectors count as digitally dominating at all, despite possessing the relevant hallmarks.

My overall aim in this paper is to make sense of domination by core AI systems through proposing three potential accounts of 'digital domination'. The first—what I call the *interactional perspective*—is grounded in an interactional account of domination (e.g., Pettit 1997; Laborde 2010). The basic idea is that people in digital societies are digitally dominated because we face arbitrary interference in our choices due to our lack of control over digital systems. This lack of control is due to a lack of transparency, public engagement, and, most importantly, regulation that fails to meet the standards of non-arbitrary power. Although I support the basic idea of this account, it raises several challenges related to who precisely is dominated by whom and when the domination begins. A structural account of domination helps to address these challenges.

This leads me to the second account, or what I refer to as the *marginalized perspective*. Drawing on Gädeke's general criticism of interactional accounts of domination, a structural account of domination raises the question of which structures dominate. Generally, only those people undermined in their choices are people from marginalized groups (Eubanks 2018; Noble 2018; Benjamin 2020). A seemingly obvious answer is that structural forms of societal oppression dominate people of marginalized groups (Young 1995). Such a marginalized perspective implies we cannot speak of digital domination—after all, there is nothing digital about such structural oppression. This analysis, however, fails to address the fact that non-marginalized groups face significant (potential) interferences, most recently demonstrated by ChatGPT in the educational and writing sectors.

To include the domination non-marginalized groups potentially face, I propose that we should make sense of digital domination according to a third perspective, or what I call the *socio-economic perspective.* On this view, the underlying structure of digital

domination is a socio-economic order that supports, particularly in the United States, a relatively unrestricted possibility for individuals or companies to innovate (Torrence & Von Hippel 2015). Without proper regulation in place, such as with digital artefacts, this leads to domination. On my proposed structural interpretation of digital domination, domination is structural in that the socio-economic order provides much individual freedom to innovate, yet the domination remains restricted to sectors with relatively little regulation. Digital domination, then, is a symptom of current market structures.

Making sense of digital domination in the case of digital artefacts is relevant as each conception requires a different approach to decrease the degree of domination. For the interactional account, it suffices to focus on making AI transparent, increasing public participation, and improving regulation. On the marginalized account, the main goal should be to address underlying social injustices. On my endorsed socio-economic perspective, the root cause of this digital domination can be found that innovation occurs beyond public control. I thus conclude that a robust solution for achieving a non-dominating digital age is to shift our focus to the broader socio-economic context in which AI is developed and deployed.

The rest of the paper is structured as follows. In Section 2, I elaborate on domination as arbitrary interference. Section 3 discusses digital domination based on an interactional account. Section 4 discusses digital domination understood on a structural account. In Section 5, I conclude.

2. Domination

Neo-republican theory holds domination as antithetical to freedom (Pettit 1997; Lovett 2010). What precisely constitutes domination differs amongst scholars. A general agreement is that someone is dominated when subject to a superior power that can arbitrarily interfere with their choices (e.g., Pettit 1997; Lovett 2010; Laborde 2010). However, some challenge this and argue that domination is necessarily constituted by societal structures (Gädeke 2020). Following Gädeke (2020), I distinguish these two as 'interactional' and 'structural' accounts of domination. This distinction is relevant to make sense of digital domination. I first discuss interactional, after which I differentiate this from structural domination.

2.1. Interactional domination

Interactional domination occurs when there is a potential for arbitrary interference in one's choices. I interfere in your choices when I have the power[55] to:

- remove options from your range of possible options;
- alter the desirability of an option (e.g., by imposing a penalty or threats of physical/emotional abuse);
- manipulate you into choosing a specific option (e.g., by nudging or hypnosis);
- provide you with the necessary means to realize your goal but choose not to.

These four points are a relatively common way to understand interference (Pettit 2009), so I will not expand on it. Note, though, that interference is necessarily social. With this, I mean that the 'I' doing the interference cannot take the form of natural limitations but is always an agent that stands in relation to another agent.[56] My inability to jump ten meters in the air is not a matter of the earth interfering with me, it is simply a physical impossibility (List & Valentini 2016).

With 'arbitrary' power, I follow a substantive conception that entails that non-arbitrary power is forced to track the best interests of the subordinate agent (Pettit 1997, 55).[57] This entails that the dominator can be held effectively to account when they are failing to track the best interests of the subordinate. Due to the inclusion of the best interests of the subordinate agent, it is insufficient if superior power is controlled by a third party. It must somehow be controlled by the subordinate agent, hence necessitating some form of democracy (Pettit 2012). We can identify at least three relevant conditions for such 'effective' accountability or controlled power. These are transparency, democratic involvement, and regulation.

First, some level of transparency is required to provide the relevant information to assess whether an agent is tracking someone's best interests. In addition, people should be able to be involved in the decision-making process, either through active participation (cf. Laborde & Maynor 2009) or contestation (Pettit 2012), to shape the understanding of best interests. Third, such democratic involvement should be

---

55 With power, I include both physical (i.e., taller, stronger) and social (i.e., better socio-economic status, hierarchical relation, etc.). For the rest of the paper, the social power is most relevant.

56 Agent, here, broadly construed: individual, group of individuals, corporations, state, or, as we shall see, an algorithmic system. Unlike a tree or mountain, we can conceive of an algorithmic system as an agent because it changes states depending on the agent with whom it interacts.

57 See Lovett 2012 for a clear distinction between procedural and substantive arbitrary power and his endorsement of a purely procedural conception.

regulated to avoid that public involvement still depends on the arbitrary will of the powerful agent. In the next section, I revisit these points in more depth.

Finally, I understand 'best interests' as basic liberties. It is impossible to avoid arbitrary interference in all matters of one's life. The arbitrary interference matters primarily for the set of choices relevant for a person to function in society as an equal. Pettit (2012) refers to these sets of choices as the basic liberties or as the necessary capabilities to develop oneself (Sen 1983; Nussbaum 2007; see Robeyns 2005 for an overview). Think of freedom of speech, association, movement, or competition.[58] If you cannot exercise these basic liberties without being dependent on someone else's arbitrary will, you are not in ultimate control of how to live your life. If I constantly had to ask my neighbour for permission to use the road, I would be uncertain whether I could travel that day. We see this becoming an issue if I need to travel, for instance, to school or work. The ability to exercise one's basic liberties or to act on the capabilities necessary to develop oneself is, hence, fundamental to one's flourishing as a human (cf. Lovett 2010, ch. 5).

2.2. Structural domination

Where interactional domination applies to any form of arbitrary interference in one's choices, there are some critics of this conception. In her paper *Does a Mugger Dominate*, Dorothea Gädeke (2020) argues that a mugger holding their victim at gunpoint does not necessarily dominate. That a mugger does not necessarily dominate is a controversial claim, given that it is generally considered to be a prime example of when someone is subject to the arbitrary will of another.

The mugger case raises two relevant challenges for interactional accounts of domination. First, it is unclear when the domination starts (as soon as the mugger walks into the park?), and second, who is dominated (everyone in the park?). To address these 'when' and 'who' challenges, Gädeke argues that only in the case where the mugger's actions will not have consequences (e.g., the victim will not go to the police or will not be taken seriously by the police), can we talk of domination. Compare a female victim to a male victim in a sexist patriarchy, where females are considered inferior. Gädeke claims that because of such structural inferiority, a mugger mugging a male victim is a mere opportunistic exercise of power, as opposed

[58] I take these liberties or capabilities as a given. However, as Pettit, Sen, and Nussbaum highlighted, these liberties and capabilities are culture-specific and depend on how a society functions.

to a mugger mugging a female victim, which represents fundamental structural injustices. Gädeke holds that only in the latter case can we speak of domination.[59]

Gädeke's conclusion highlights a relevant distinction between interactional and structural forms of domination. Where interactional domination prohibits my flourishing, structural domination necessarily concerns my moral and political standing with others (Gädeke 2020). Indeed, on structural accounts of domination, I am not seen as a 'voice worth hearing and an ear worth addressing' (Pettit 2002, 350). Resolving structural domination then also differs from resolving interactional domination. Closing off parks at midnight removes opportunities for mugging. This approach may reduce domination on an interactional account. However, on a structural account, it fails to address female domination, as women remain structurally oppressed.[60] Whereas preventing the possibility of arbitrary interference could theoretically reduce or even resolve interactional domination, structural domination thus necessarily involves grand institutional reforms that tackle people's moral and political standing in society.

In sum, domination refers to relations of power in which the dominant agent can arbitrarily interfere with the subordinate agent. For structural forms of domination, non-domination requires significant transformations in core societal structures, norms, and values. In contrast, we can aim to prevent specific circumstances that may facilitate domination for interactional accounts. The following two sections discuss domination by digital artefacts from an interactional and structural perspective.

---

[59] Those supporting interactional forms of domination do not deny structural forms, yet Gädeke's point is that interactional forms of domination do not exist. I focus strongly on Gädeke's account because I find her arguments against the mugger case persuasive and helpful in making sense of digital domination. Unlike Gädeke, however, I believe there is still value in interactional forms of domination that undermine people's basic liberties. Gädeke's account, at least as described in the Does a Mugger Dominate paper, excludes different degrees of domination. Even if one rejects my proposed socio-economic perspective in Section IV, they may agree with the interactional perspective. In such a rejection, only the degree of severity of the domination is lost, not the notion of domination altogether.

[60] Pettit does believe that non-domination requires institutional reforms. My point, to be clear, is that closing down the park during the night may reduce mugging male victims but would not affect the underlying societal structures that make female victims lack the 'antipower' (Pettit 1996) needed to seek redress.

3. Digital domination: an interactional perspective

The primary point of this paper is to make sense of domination by digital technologies, or digital domination. My take on digital domination differs from the other approaches in the literature. Specifically, whereas others have shown interest in digital domination with regards to Big Tech companies (Aytaç 2022; Susskind 2022; Hoeksema 2023) or the Gig Economy (Muldoon & Raekstad 2023), we see less focus on more digital technologies that increasingly influence our daily lives.[61] While I do not deny the relevance of the focus on online platforms, if concerns of domination apply to such every-day technologies, the digital domination complaint becomes more severe. While not everyone is an employee of a Gig Platform, or not everybody uses social media platforms, we can no longer function in society without these digital artefacts (Gabriel 2020). Indeed, we have no 'opt-out'. Digital domination then becomes a problem for all. Before getting into the digital domination, let me elaborate on what I mean by 'digital technologies' and how they may interfere with people's basic liberties.

3.1. Digital technology and basic liberties

First, I limit myself to (semi-)automated decision-making systems (ADS) trained on large datasets that (semi-)automatically support decisions in core societal sectors such as healthcare, judicial, or public benefits. Second, as numerous authors have shown, AI systems like ADS are necessarily socio-technical. They are developed by people with a specific background and deployed (and interpreted) by people in a particular context (Hildebrandt 2021; Dobbe et al. 2021; Crawford 2021; Noorman & Swierstra 2023; Johnson & Verdicchio 2024). Although I will refer to ADS as a seemingly stand-alone object throughout the text, I do so for simplicity. It is critical to remember that we cannot separate the system's output from this social context. Speaking of technology as an agent that interferes or dominates is thus used as a short-hand.

How do ADS interfere with us? Core societal sectors such as medical, financial, judicial, or public administration domains are increasingly algorithmically structured (Gabriel 2022). ADS used in these sectors affect citizens' abilities to pursue their goals,

[61] Susskind (2022) extensively discusses the digital technologies I have in mind. His work is limited for two reasons. First, he discusses these technologies in parallel with online platforms, which sometimes obscures which parts of his argument relate to the artefacts of concern in this paper and which apply to online platforms. Second, his solution to a non-dominating technological order focuses solely on online platforms.

wishes, and desires (Maas 2023). Suppose I have a desire to receive social benefits, but the algorithm explicitly denies me this possibility. In that case, the ADS prevented me from realizing my goal even though the algorithm, theoretically, could have done so. Without making a claim on whether this denial is justified, this is a form of interference.

Such interferences are often necessary for us to function as equals in society and flourish, exemplified when ADS produce discriminatory outputs. Discriminatory biases that affect the already marginalized groups in society are well-known. Consider the limitations of facial recognition technology. Then-student Joy Buolamwini noticed the need for more accurate identification for people (especially women) of colour when the system consistently failed to identify her (Buolamwini & Gebru 2018). Or, think of how a system that should predict health care requirements consistently underscored black patients, who hence did not receive the health care they needed (Obermeyer et al. 2019; Benjamin 2019). Such discriminatory biases extend into many other fields. The US border uses facial recognition technology for immigrants, which negatively affects black asylum seekers in the US who are unable to file for asylum as the system does not recognize them (del Bosque 2023). Moreover, if I do not receive the health care I need, further developments, such as finding a job, may not be an option for me because of my physical limitations. My point is that discriminatory bias may undermine your potential to realize your capabilities and act on your basic liberties.

There is thus a form of interference of digital technologies that affects people's basic liberties. Such interference need not be dominating when done in a non-arbitrary manner. However, one of the most pressing concerns in the AI Ethics literature is the lack of accountability or control over these systems (and their developers and deployers). The accountability gap is due to a lack of transparency, lack of public engagement, and a lack of regulation.

3.2. Arbitrary interference

The lack of transparency, or opacity concern, relates both to the technical features of the system itself as well as to the broader development behind the system. Regarding technical opacity, AI systems are notorious for their black-box character, where we know their input and output but not how the AI system precisely turned the input into the specific output (Burrell 2016; Lipton 2018). Regarding the latter, most algorithms are developed by private companies that do not necessarily have to share information about their workings. To address these concerns, we would need *technical transparency*

(making clear the inner workings of the algorithm) and *process transparency* (being informed about the development process, including procurement assessments) (Zerilli 2021, 23).

There are two limitations to such efforts to achieve transparent AI. First, technical transparency is difficult if not impossible due to technical limitations of a system (Durán & Formanek 2018). The idea is that 'transparent' or 'explainable' AI is simply not feasible, as we will never be able to explain the *exact* workings of a system, but enter into an infinite regress with regard to which elements of a system are explained. Such an infinite regress necessarily reduces a person's ability to contest the output of an AI system. The lack of technical transparency (and its potential impossibility) is therefore a limitation to control after we face interference in a choice and wish to contest the output (e.g., after the denial of healthcare). A more fundamental limitation, however, has to do with the fact that transparency in and of itself is insufficient to ensure non-arbitrary power. A dictator may very well be open about how he will impose his authority. However, if there is no possibility to influence, contest, or challenge, there is no way to ensure that a transparent dictator will track the best interests of his subjects. Thus even if a company is transparent in a 'process transparent' way, it will be insufficient if the public cannot challenge design choices and if decisions regarding the sharing of information remain in companies' hands.[62] Given that we face limitations on 'ex post' means for contestation and a need to include contestation already during the design process, it is not surprising that we see increasing calls for public engagement during the design and development of ADS. Several scholars argue that a form of direct or participatory design is necessary both to better align the system's output with societal values, norms, and desires, as well as to address the power relations between the developers and deployers on the one hand and the end-user on the other (see Delgado et al. 2023 for an overview). A democratic approach to ADS design, however, also comes with challenges.

For one, the resource intensity that comes with participatory democracy more generally also applies to AI systems (Himmelreich 2022). Second, and more relevant for my purposes, 'participation' does not address underlying power asymmetries in and of themselves (Birhane et al. 2022; Sloane et al. 2022). Again, suppose the dictator has all the decision-making power regarding whom to invite and include and how to include their input. In that case, participation does not necessarily provide a meaningful sense of accountability. After all, the dictator is not forced to track their

[62] Note that this poses a concern for systems that can be 'gamed' or include sensitive information.

best interests. The question is not just 'Should this system exist?' but also who decides whether this system exists, and, to follow Zuboff (2019), *who decides* who decides? Regulation is therefore critical for establishing a coherent normative framework to which AI systems must conform, including certain transparency and stakeholder inclusion requirements.

The AI Act recently passed by the European Parliament is arguably the most relevant and serious attempt at regulating AI (European Commission 2024). Without the AI Act, the normative choices would remain in the hands of the developers and deployers (Laux et al. 2024). Such absolute power is clearly at odds with republican ideals, and the general idea of the AI Act is, therefore, precisely what is needed. However, the Act still has its limitations.

First, the AI Act must be clearer on what stakeholder inclusion implies. For instance, it states that "[w]hen identifying the most appropriate risk management measures, the provider should document and explain the choices made and, when relevant, involve experts and external stakeholders" (European Commission 2024, 42a). Or, "[w]here appropriate, to collect relevant information necessary to perform the impact assessment, deployers of high-risk AI system (…) could involve relevant stakeholders" (European Commission 2024, 58g). These statements leave much room for interpretation and voluntary inclusion to the provider and deployer of the AI system. The Act's attempts to regulate stakeholder inclusion thus have no teeth, leaving references to stakeholder inclusion relatively meaningless.

Second, and more importantly, on a broader level, we can wonder whether the AI Act is a proper means to address power relations, as the industry largely shapes the Act. For instance, OpenAI (the company that developed and deployed ChatGPT) convinced the European Parliament to move ChatGPT off the 'high-risk' level, which requires additional oversight and imposes stricter requirements on development and deployment (Perrigo 2023). Moreover, the AI Act contains several references to technical standards regarding documentation, validation, information, annotation, and training. Over time, these standards will be decided upon and established by private standard development bodies that often use employees from the industry (Büthe & Mattli 2011; Perarnaud 2023). Such standard-setting by employees from the industry can be problematic for two reasons.

First, we may worry about a 'revolving door' because the people who must be regulated are partly regulating themselves (Susskind 2022). Both on procedural and substantive conceptions of arbitrary power, such a revolving door would insufficiently control power. Second, we can wonder about the expertise of industry partners in

translating human rights and interests into actionable technical standards representative of said rights and interests. Granted, industry is often involved in developing technical standards, yet the AI Act specifically focuses on complex human rights, a highly contested and complex topic due to the ambiguity of potential interpretations of the values and rights at stake. In addition, standard-setting often requires making trade-offs between different values and rights. Some level of understanding of the humanities, therefore, seems essential. However, because the European Standard Commission is necessarily "purely technical" (Galvagna 2023, 6), the private industry partners only require an academic degree or experience in the industrial sector, but not in the humanities such as law or policy (Galvagna 2023, 23).

We can thus question whether private partners have the expertise and legitimacy to translate complex and ambiguous values into technical standards and make trade-offs between different values and rights. Moreover, the industry setting the standards does not 'force' them to track the best interests.[63] While the Act includes many relevant points, this lack of expertise and legitimacy by experts in human rights thus causes reason to question the normative foundation of the Act itself.

We can conclude that the means to control the interference of ADS in people's basic liberties is insufficient. Transparency, public engagement, and regulation still leave room for improvement. Consequently, digital domination applies not only to more specific relations in the Gig Economy or those produced by Big Tech. It applies to many more digital artefacts that increasingly affect people's ability to act on their basic liberties. So far, ADS present a case similar to the interactional account of the mugger, where everyone who is mugged (here, subjected to the ADS) is dominated as they are subject to arbitrary interference.

If this is true, however, we are confronted with the same challenges that apply in the mugger case. First, when does the domination start? Second, who precisely is dominated? And third, to expand on the mugger case, who precisely dominates? These questions are particularly relevant to identifying who to include during what phase of the development process of the ADS, which in turn is relevant to shaping regulation and policy. As we shall see, an effort to answer these questions through a structural perspective, however, might undermine our idea of 'digital' domination.

---

[63] On a procedural conception of arbitrary power, this would arguably not matter. See Lovett (2012).

4. Digital Domination: two structural perspectives

It may seem highly contradictive to first argue how digital technologies may dominate, only to question the idea of *digital* domination in itself. Although I will conclude that the 'digital' in digital domination is not completely misguided, I will argue that the root source is a socio-economic order that upholds and maintains the possibility for innovations to be deployed into society relatively unrestrictedly. Digital technologies such as ADS are a symptom of a broken system. But let me not get ahead of myself and return to how the three questions invoke this broader, structural concern. I do so by first addressing a more obvious structural perspective, grounded in already existing power asymmetries between marginalized and non-marginalized groups. Then, I propose a structural perspective grounded in socio-economic structures.

4.1. The marginalized perspective

The first question ('when') is when domination starts. Take a system to allocate childcare benefits used in the Netherlands.[64] While I am personally not subjected to that system (given I do not have children), I might in the future (were I to decide to have children and continue living in the Netherlands). If domination is not just in actual interference but in the potential to interfere, when does this potential for interference begin? Am I already dominated just because I potentially might be in the future? This seems too broad (Richardson 2002). Nevertheless, to say I am not yet dominated but only when I am actually subjected to the system conflates the domination into an actual act of interference, losing the 'essence' of the evil of domination (cf. Gädeke 2020).

The second ('who') and third ('by whom') questions relate to who is dominated and who dominates. Most of the examples in the previous section affect primarily already marginalized groups, such as women or people of colour. Indeed, generally, scholars focus on how digital technologies exacerbate existing power asymmetries (Eubanks 2018; Noble 2018; Benjamin 2020). The same holds for the childcare benefits system used in the Netherlands. This system has become notorious because it was indirectly discriminating against particular ethnic groups.[65] Due to biased training data, it was likely to flag people with a non-Dutch nationality as a higher risk of

64 The actual system was used to assist decision-making in benefit allocation, and the process was not fully automated. For ease of argument, I have changed the example that the system automatically allocates childcare benefits.

65 https://www.mensenrechten.nl/actueel/nieuws/2023/10/2/toeslagen-oordelen

fraud.[66] This *unprecedented injustice*[67] thus primarily affected immigrants with an already marginalized background. Would I, then, as a native Dutch, really be dominated if I were to have been subjected to the system?

Gädeke's (2020) distinction between interpersonal and systemic forms of domination helps us to answer the 'when' and 'who' questions. Interpersonal is when there is an actual agent who interferes with you. In contrast, systemic domination is when someone experiences a continuing form of oppression, even when not directly subjected to an identifiable agent. If digital technologies only exacerbate existing structures of domination of marginalized groups, someone belonging to an oppressed or dominated group in society would be interpersonally dominated when subjected to the system. When not subjected to the system, they stop being interpersonally dominated but continue to be systemically dominated. The distinction between interpersonal and systemic domination answers the 'when' question. 'Who' is dominated are only those who are part of the dominated group. Myself, in the case of the childcare system, would thus not be dominated, including when I am subjected to the system.

The relevant question here is the domination 'by whom' question. Marginalized groups stand in an inferior power relation with regards to their oppressors,[68] and domination is constituted by underlying structures that reflect power asymmetries such as male/female or white/black people. Such underlying structures raise the question of whether we can speak of digital domination. Digital technology may worsen things but does not provide a new source of domination. To speak of 'digital' domination is thus misguided as existing forms of domination are exposed/ exacerbated by digital technology.

Although I am sympathetic to the idea that these systems exacerbate existing power asymmetries, the marginalized perspective excludes digital artefacts that can affect anyone, including those from a non-marginalized group. Consider, for instance, how the large language model ChatGPT has affected numerous professions, ranging from teachers in educational institutions to scriptwriters for films and series (Roose 2023; Anguiano 2023). By focusing solely on the marginalized groups affected by

---

66 https://www.autoriteitpersoonsgegevens.nl/uploads/imported/onderzoek_belastingdienst_kinderopvangtoeslag.pdf

67 https://www.tweedekamer.nl/sites/default/files/atoms/files/20201217_eindverslag_parlementaire_ondervragingscommissie_kinderopvangtoeslag.pdf

68 The explicit focus on structural power asymmetries exposes some underlying questions about Gädeke's account. Is a female victim only dominated when a male mugger mugs her? If not, this would undermine the strong emphasis on oppressor vs. oppressed. I leave this question to the side.

digital technologies, we ignore that, theoretically speaking, anyone can be interfered with at one point. Those who are not (yet) are in some regards lucky. Although the marginalized perspective may include domination, it is a form of non-digital structural domination. To better understand the root source of digital domination, I find another perspective more plausible: the socio-economic structure we find in liberal democracies.

4.2. The socio-economic perspective

ChatGPT is an interesting case as it highlights a vulnerability in society. Educational institutions are vulnerable, just like scriptwriters, journalists, and others. This vulnerability is not only related to people of marginalized groups but extends to society more broadly. I believe it is a vulnerability rooted in how we perceive our freedom to innovate. Especially in the United States, there is strong support for an individual's freedom, or even 'right',[69] to innovate (Torrance & Von Hippel 2015; Fisher III 2009). As Torrance and Von Hippel (2015, 802-803) write, "the burden of proving that innovative activities do violate specific, existing legal prohibitions, or unreasonably endanger or harm others, generally lies with those who oppose these innovative activities." Innovations are thus, generally speaking, in favour of the innovator.

However, specifically for new technological innovations such as ADS, such burden of proof is difficult if not impossible. First, the negative effects are often unknown, making it difficult to identify potential harms. It has, for instance, been suggested that Facebook contributed to the genocide in Myanmar due to the way the platform's recommendation systems structured its content at least facilitated the genocide (Mozur 2018).[70] While Zuckerberg's mission of connecting the world was not necessarily evil, in following his vision Facebook did move fast and 'break' Myanmar. Yet arguably at the start of Facebook no one would have foreseen this as a potential consequence.

Second, as is well-known, regulation comes after innovation, further complicating the burden of proof. For one, in certain situations there simply is no regulation (or liability scheme) in place that provides citizens with the legal tools to seek

---

69 Right, here, is meant to endorse human activity. See e.g., Kranzberg (1986, 548), who writes, "[i]nvention is the mother of necessity." Or consider Hatta (2020, 1), who describes innovation as a "staple of human existence" (Hatta 2020, 1).

70 See the open letter to Mark Zuckerberg by interest groups in Myanmar as cited in Zhang, Gowder and Rogers (2023).

accountability. We indeed see a new development of a liability directive for AI systems specifically that is meant to "ease the burden of proof for victims to establish damage caused by an AI system" (Madiega 2023, 1). Yet, because of AI's black-box character, it remains to be seen whether citizens can act on their right to compensation in case of harms that have occurred (BEUC 2022). Furthermore, the judicial system is, logically, framed around addressing harms that have actually occurred. However, often the harms of digital technologies only show at a collective, statistical level, making it difficult to pinpoint down whether an individual was actually discriminated against, or rather received just an unfortunate inaccurate output. This makes it difficult for liability legislation to defend victims of digital harms (Cohen 2019; De Zwart 2023).

The 'regulation after innovation' is problematic for another reason, as it leads to inherently "anarchic innovation" (Hussain 2023, 121). Regulatory schemes to maintain moral obligations in a society is the selling point of the functionalist view of liberal democracies (Hussain 2023). Yet it is fundamentally flawed when we consider brand new fields such as digital technologies. Surely, the AI Act will check development and deployment. But the AI Act also is very recent, and not even in effect yet. On the one hand, it was beneficial when ChatGPT was deployed that the AI Act was still in development, as this made adapting regulation for such more general purpose technologies easier. On the other hand, we probably should not be too surprised ChatGPT was released *precisely* because there was not yet regulation, and OpenAI was less restricted than would have been otherwise the case. Zuboff's question, '*Who decides* who decides,' is hence particularly interesting for innovations in new and unregulated fields. Given a lack of regulation to guide and control innovators, the decision is up to the innovator, not society. Consequently, in such new fields, there is little room to ensure the innovation is forced to track the best interests of the public (Bennett & Claassen 2022; Mayer 2018).[71]

What does this socio-economic perspective mean for digital domination? The structure does not necessarily have something to do with 'digital'. However, the fact that the 'digital' is a relatively new field makes it different from other markets where regulation already exists. Innovations in the healthcare sector or car industry are much more regulated, and to speak of domination in those sectors would be incorrect. There is, therefore, something most certainly specific about the digital sector that not only exacerbates domination (as with the marginalized view) but truly brings a new

[71] Bennett and Claassen (2022) discuss how entrepreneurial freedom has changed throughout the centuries.

line of domination into society. As I see it, digital domination is a symptom of our socio-economic order.

The core reason we now face digital domination is due to a relatively unrestricted approach to novel innovations where socio-economic structures favour innovators over society. In the context of digital technologies, society is structured in a way that has enabled digital technologies to be developed and deployed without proper oversight. The living proof is that of a college kid who was able to move fast and break things without any oversight and checks and balances, or the deployment of a large language model that inspired a letter signed by more than 30.000 leading AI experts to in a sense *beg* for a pause on AI development. To know who dominates, we only need to ask to whom this letter is addressed. The answer is "AI Labs."[72]

On this structural account of digital domination from a socio-economic perspective, we can answer the 'when', 'who', and 'by whom' questions as follows. 'Who' is dominated? Everyone in modern, digital societies. 'By whom' are we dominated? Potential successful innovators. And 'when' are we dominated? We are interpersonally dominated when subjected to new successful innovations that lack regulatory frameworks. When not subjected to such innovations, we remain systemically dominated as we can be subjected to the next innovation due to a relatively relaxed and unconstrained approach to innovation.

5. Concluding remarks

In this paper, I have attempted to make sense of digital domination in the context of digital artefacts that pervade our everyday life. An interactional perspective of digital domination is limited as it remains debatable who precisely is dominated, by whom, and when. Although a marginalized perspective provides an answer, it also moves the debate away from the 'digital' to already existing power asymmetries. While attention to these increasingly exposed power asymmetries is vital to address structural injustices, I have proposed that the most fruitful way to look at digital domination is to consider how socio-economic structures allow new innovations to be developed and deployed in a way that is not necessarily in the best interests of society. The AI Act does constrain this in some sense, but on a substantive conception of non-arbitrary power it does so insufficiently. Regardless whether one accepts my proposed

[72] https://futureoflife.org/open-letter/pause-giant-ai-experiments/

perspective, I hope to have illustrated the relevance of a debate on 'digital domination' that goes beyond the Gig Economy and Big Tech.

In society, we necessarily depend on numerous artefacts, corporations, people, or institutions to function as an equal in society. Society is becoming more digital with the day, affecting and shaping *how* we exercise and may act on our basic liberties. How to address the innovation challenge is a question beyond the scope of this paper, but what is clear is that we must ensure that innovation is done in a way that tracks society's best interests. Langdon Winner (1987) refers to this as *political ergonomics*: the fit between technology and society. In doing so, Winner moves the debate to what is genuinely beneficial and necessary *for society*, rather than what is desirable for innovators.

In the context of digital technologies, a socio-technical design approach that includes stakeholders is critical. Moreover, regulation must *force* such inclusion in a way that provides stakeholders not just with a voice, but an actual say. While this may reduce digital domination, addressing the root source requires a more fundamental re-assessment of current socio-economic structures. Ultimately, to address the root source we need a way to effectively ensure a political ergonomics for innovations more generally.

Conclusion:
Designing for Non-Domination

I began this dissertation by highlighting the general worries related to online platforms, such as surveillance capitalism or the uncontrolled power of Big Tech. Drawing on these general worries, scholars like Susskind, Muldoon, and Aytaç have claimed that domination is a concern in the context of online platforms. In this dissertation, I have argued that these concerns of domination extend beyond online platforms. I have claimed that AI systems deployed in core societal sectors lead to relations of domination between those who shape a system (the developers and deployers) and those affected by it (the end-users and society at large). In addition, I argued that digital domination poses a problem for all members of modern, digital societies, even when they are unlikely to be actively interfered with (i.e., being constrained or coerced in one's choices).

In order to successfully address digital domination, the AI Ethics literature must explicitly design AI systems with the moral concern of domination in mind, a concept underexplored in the context of AI Ethics. To fill this gap, I propose that responsible AI development requires a *design for non-domination* approach. This approach provides new practical and conceptual insights that can improve responsible AI development by addressing power asymmetries between the shapers and the affected. In what follows, I dissect what it means to design for non-domination, reflect on several assumptions I have made in the dissertation, and propose a fruitful route forward based on my account of digital domination.

To start off, in Chapter 1, I argued that technological advancements can be a driver for morally motivated conceptual engineering. In liberal democracies, citizens are expected to be in control over their lives, to have the possibility to realize their self-governance. They are expected to be free. Throughout this thesis, I have argued that the development and deployment of AI systems in its current form inhibits citizens such a form of control, undermining their freedom. To safeguard freedom in the digital age, it is thus essential to explicitly focus on enhancing citizen control. Designing for *freedom,* here, is too vague for citizens to successfully regain their freedom. It risks that development and deployment of AI systems remains on the terms of the developers and deployers rather than on the people's terms.

Reconceptualizing freedom in more concrete terms of *republican freedom* provides a suitable approach to enhance citizen control over their lives. Rather than designing for *freedom*, thus, we must specifically design for *non-domination*.

Designing for non-domination requires both application-specific (e.g., stakeholder inclusion) and institutional measures (e.g., a regulation that empowers stakeholders to be included). Application-specific measures focus on what the development of a particular AI system must look like. I have in mind the direct power relation between the shapers and the affected of a system. In Chapter 2, *Machine Learning and Power Relations*, I provided an account of these power relations based on Castelfranchi's power-dependency relation. This power relation between shapers and affected exists because the goals and desires of end-users depend on the output and behaviour of an AI system, which in turn depends on the design decisions by the shapers. AI systems thus entrench an indirect power relation between the system's shapers and the affected.

In Chapter 3, *A Neo-Republican Critique of AI Ethics*, I showed how the conceptual background of AI Ethics is informed by a conception of freedom as non-interference. This suggests that AI Ethics is biased to focusing on mitigating harmful outputs rather than addressing these power relations underlying AI development and deployment. In both Chapter 2 and 3, I suggested incorporating design-for-values methodologies such as value-sensitive design or participatory design as an effective way to hold accountable the superior power of the developers and deployers of AI systems. Stakeholder-inclusive design approaches empower stakeholders. On the application-specific level, therefore, domination can be addressed by stakeholder inclusion.

However, as I argued in Chapter 4, *Opening the Black Box of AI, Only to be Disappointed*, and in Chapter 5, *Beyond Participatory AI,* such stakeholder inclusion will only be successful if we make changes at the institutional level. Chapter 4 specifically discussed the limitations of a narrow view of AI Ethics. I argued AI systems must be institutionally embedded (i.e., we must be able to trace back their deontic provenance) in order to have normative reasons to act on these systems. This provides citizens of digital societies the possibility to contest the broader chain of development of these systems.

In addition, Chapter 5 advocated for the need to institutionally embed design-for-values methodologies. When we talk about power asymmetries, we necessarily talk about decision-making power. Such decision-making power, however, primarily rests with the developers and deployers. Simply providing citizens with the possibility to have a say is not sufficient if this say depends on the goodwill of the developers. Indeed,

designing for non-domination requires that such decision-making power rests under public control, which in turn requires that stakeholders must be able to *force* AI developers to take their interests into consideration. Therefore, stakeholder inclusion during the development of particular AI applications is only meaningfully non-dominating when supported by institutional structures such as regulation and legislation tailored to stakeholder inclusion.

In Chapter 6, *Making Sense of Digital Domination*, I furthermore argued that even if we address concerns with *digital* domination (i.e., improve regulation in the field of AI), we risk overlooking the socio-economic structures that gave rise to digital domination in the first place. This source of domination is neatly captured in Hussain's (2023) phrasing of 'anarchic innovation' (i.e., the lack of regulation for new technological innovations). Ultimately, the concern is with the neo-liberal ideology we find in economic market structures that play a significant role in constituting digital domination. This position, in combination with the assumption that citizens have a 'freedom to innovate', allows individuals and private companies to develop and deploy AI systems according to *their* best interests, without being forced to consider the public's best interests. Here, specifically, the domination complaint addresses the broader societal structures that have enabled uncontrolled development and deployment of AI systems in the first place.

Addressing non-domination on both an application and institutional level thus requires fundamental institutional reforms. It requires a responsible approach to innovation as endorsed in the EU that puts societal interest first and profits second. By putting society first, the innovation is more likely to *fit* society in the sense of Winner's 'political ergonomics' discussed in Chapter 4. Current capitalistic systems cannot support such a change in priority. Some have even argued that republicanism and capitalism are necessarily incompatible (Bryan 2020). Before drawing such a harsh conclusion, a good start is to first see to what extent changing a neo-liberal economic order to a form of 'enlightened' capitalism will mitigate concerns of domination with regard to innovation more generally. In an enlightened economic order, businesses strive for the public's best interest. Part of this requires putting business interests (i.e., profit-maximization) aside (Mayer 2018). Evidently, such an economic order requires drastic institutional changes both economically and legally. However, as Pettit (1997) and Gädeke (2020) show, achieving non-domination requires such institutional changes.

Let me return to the overarching question that informed this dissertation. I asked how the freedom of citizens in modern, digital societies is undermined by the

uncontrolled power of developers and deployers of AI systems to interfere in their lives. Republican theory shows how uncontrolled power and freedom are related by actively integrating the distribution of decision-making power into its conception of freedom as non-domination. Freedom in the digital age urges us to rethink our conceptual background of freedom. When operating from a dominant normative framework, such as a liberal conception of freedom as non-interference, our attention and solutions are also primarily provided to fit this particular normative context. Actively interrogating dominant conceptions in society paves the way for 'out-of-the-box' solutions that quite literally cannot be provided from within the operative conceptual background. Rethinking our conception of freedom is thus a necessary way to better safeguard freedom and democracy in the digital age.

Limitations and reflections

A few of the claims and assumptions made in this dissertation deserve deeper reflection. In the following paragraphs, I elaborate on potential limitations that arise with my categorization of the 'affected' and the 'shapers', and reflect on world events from the past four years such as technological advancements like ChatGPT, the Covid-19 pandemic, and the Russian invasion in Ukraine.

Who is affected?

The first limitation relates to who precisely falls under the 'affected' group. In this PhD, my primary interest has been on the relationship between those that shape an AI system, and those affected by one. I have assumed that 'affected' implies *directly* affected (see Chapter 2), however one might object this is too narrow of an interpretation. To see this, let's consider the example of Cambridge Analytica (CA) I discuss in Chapter 3. People's data was gathered and they were presumably manipulated into a particular political preference. Even if the nudging was not as effective as people may have feared when the scandal broke, the issue remains that this *could* have happened in the first place.

This raises the question regarding who precisely were the directly affected people here. Was this only those whose data is gathered, or only those who were targeted (i.e., those still deciding on their vote)? For the sake of argument, let's assume CA did have the intended effect. Are then only the people who were successfully nudged to vote for the republican party affected? Moreover, what about all US citizens who either did not have Facebook, whose data was not collected, or who were not

successfully nudged? They still live in a country where people's voting decisions affect who their president will be. Are they then not also affected? Even further, given the US is such a big player in the economy with relevant veto rights and global power, one might argue that all citizens of the world are affected given the effects have a global reach.

What the CA case shows is that, ultimately, there are degrees of being affected. Arguably, a person experienced more invasion when their data was taken, was nudged, and was *successfully* nudged than someone whose data was taken, but was not nudged. After all, they did not experience any direct consequences from the data being taken. In addition, a non-Facebook user who now lives under the rule of someone elected by means of digital nudging (again, assuming this was the case), is more affected than a person living on a farm in New Zealand. It is only when the US President does something global that affects this person. Who is affected in what way is therefore context-dependent. As Pettit (2005) explicitly highlights, domination comes in degrees. One way to see how domination comes in degrees—and what indicates the severity of the domination—is to show how affected someone is by a particular AI in a specific context.

Political economy

Whereas the previous limitation focused on who precisely are the 'affected' in the power relation, this limitation focuses on who precisely are the 'shapers'. In Chapter 2, my primary focus is with developers and deployers broadly construed whose design decisions have an affect on the behaviour of the AI system. I have assumed the developers and deployers imply those people who have relevant decision power with regard to how a system behaves. However, there is a whole array of people involved in the development and deployment. Given the broader social setting of AI systems, who precisely has 'relevant decision power' requires further elaboration.

As Crawford (2021) points out in her book *Atlas of AI*, there are various people involved who even make possible the 'existence' of AI. These people include those who harvest the minerals used for chips, or those who are required to run these software programs and crowd workers (e.g., amazon mechanical Turks or Prolific) that do intensive labour required for dataset labelling.[73] Numerous people have pointed out, especially regarding labelling of datasets by crowd workers, that people part of these groups stand themselves in an inferior position of power compared to the

[73] Sometimes, the affected and shapers become mixed, such as with 'Captcha'.

companies for which they work. Often, crowd workers earn poorly, yet their economic opportunities are so limited they have little to no bargaining power. I do not extensively discuss how these relations of power factor into the relation of power I am interested in: between the 'shapers' and 'affected.'

Although in Chapter 5, I briefly touch on these relations when I discuss that participatory AI design requires engagement with the broader political economy that structures AI development (e.g., workers that filter content for data sets in turn used for specific AI systems), I have remained ambiguous on whether we can speak of people with these types of less fundamental decision-making power to be in a dominating position with regards to the affected as well. I believe it would be too permissive to say that they are, precisely for the same reasons why I consider the 'affected' to be dominated: as I argued in Chapter 5, they are not in the position to control the development in a meaningful way. That is, they have no meaningful say in whether the system will be developed and/or deployed. This is then what I have in mind with 'relevant' decision power.

Changing worldview

Finally, a limitation and reflection. I started my dissertation in November 2020, during the Covid-19 pandemic. The pandemic increased reliance on online technologies, which was already significantly high at that time. It furthermore showed how digital technologies such as the internet are necessary to continue functioning in society. The pandemic highlighted vulnerabilities of people of lower socio-economic classes that did not have a suitable work or study place at home. Although in the dissertation I do not focus on digital technologies in such a general sense (i.e., including email or the internet), my arguments can be extended to them as well because of such arising vulnerability.

Other examples of such vulnerability and technological innovations are, for instance, when ChatGPT was released in November 2022. ChatGPT forced educational institutions to, quite rapidly, change their way of assessing and examining their students. An easy reply is to say that technology always causes societal change, and that we should be flexible. Yet what I saw was a vulnerability realized. As a response to ChatGPT, educational institutions started to incorporate AI-detection tools that were too hastily made and/or not accurate enough to be used in a successful manner. These systems would sometimes flag innocent students as having committed plagiarism. ChatGPT's release illustrated the vulnerability of society in general, which

nudged me towards the idea that the problem is rather with innovation than with just technical and social limitations of AI applications in general.

This idea was enforced when Elon Musk's role in the Ukraine-Russia war became public. In February 2022, Russia invaded Ukraine, and Musk graciously offered Ukraine to make use of his SpaceX satellites. There are two noteworthy elements to this case. For one, although initially Musk offered to help Ukraine, after the war continued he realized the costs were getting to high for him and wanted to pull out. The US government had to step in an make an agreement with Musk in order to continue providing Ukraine with Starlink access. Note that Musk could have withdrawn Ukraine's use of his satellites if he pleased. In addition, in September 2023, Musk's biography suggested a prominent decision-making role of Musk in the Ukrainian-Russian war. Ukraine had inquired to make use of his satellites for a military offense on Russia. Musk refused to turn on his satellites for this specific event. Musk states that "[i]f I had agreed to their request [to turn on Starlink], then SpaceX would be explicitly complicit in a major act of war and conflict escalation" (Creamer 2023).

Here, the noteworthy element is that Musk believed he was not involved because he did not actively turn on the satellite. The thing is, he *could* have done so if he wanted to. Musk's involvement in the Ukraine-Russian military conflict thus is an extreme case of a country arbitrarily depending on the wishes and desires of one man. Although this case is not directly related to how AI dominates, because of the possibility of one man to become so powerful economically, he also had gathered immense socio-political power. And the reason why he had become so powerful economically, socially, and politically, was precisely because that is a possibility in our current socio-economic order. Indeed, Musk is a product of the great iconic garage innovator. Both Musk and ChatGPT, thus, caused me to strongly believe that the issue is not just in AI systems' technical limitations that constitute a potential for domination, as what I did when I started my PhD. The world's development over the course of four years led me to re-assess my initial intuitions and hypotheses.

Future research

Building on this dissertation's limitations and conclusion of a non-dominating approach to innovation, I suggest the following lines of future research. First, it is already widely known innovations bear significant costs for society. The EU's approach to Responsible Innovation actively supports to reduce costs. While this line

of research is absolutely vital, we need to focus on the broader socio-economic structures when it comes to responsible innovation. These structures as they stand, especially in strong liberal societies such as the US, currently risk being out of public control. Yet if we take seriously democratic ideals which essentially requires public control, these underlying societal structures, indeed, undermine democracy. The line of research I thus strongly advocate is to investigate what a non-dominating socio-economic order would look like. Relevant questions would be to ask how precisely a socio-economic order leads to domination, what is necessary for a socio-economic order to be *forced* to track the best interests of society, and to what extent governmental regulation and people's freedom to innovate are to be co-shaped.

Second, building on this, we must be critical to venture capitalist arrangements, where start-ups can grow because of private equity. Similar as how proponents of republican theory are often critical of lobbying where non-governmental organizations may influence politicians, we should be critical when private agents can 'lobby' for which AI systems should be developed. This requires some serious philosophical argumentation, however. Would it mean that only public agencies embedded in governmental institutions can make such decisions? From a republican perspective, providing governments with too much power is problematic. Governments are already powerful, and we must be weary to not create relations of domination between the state and society. One potential fruitful route in my view is to explore the value of an *associative* democracy. An associative democracy adopts a bottom-up form of democracy, leaving much decision-making power in the hands of citizens. Such a distribution of decision-making power allows for public control without providing governments too much power that causes similar risks of domination. This line of research would thus inquire how associative democracies can reduce the illegitimacy of venture capitalist and hence contribute towards a non-dominating socio-economic order.

Third and finally, despite my claim that non-domination starts with non-dominating socio-economic structures, that is not to say that current initiatives in Responsible AI development to increase contestability, transparency, or public inclusion are not viable research projects. Particularly, it would be interesting to see how initiatives such as including contestability in ethical frameworks, such as the Australian government has done (Lyons et al. 2021) and initiatives in 'contestability-by-design' (Alfrink et al. 2022) may be a fruitful approach for addressing digital domination. In addition, initiatives in participatory AI design and work on Explainable AI are essential for improving accountability and contestability. Although

as I argued in Chapter 5 and 6, neither of these approaches are sufficient to addressing digital domination, they do contribute to accountability and contestability. Contestation is at the heart of republican theory, and any further research to improve on this is welcome.

Bibliography

Adams, R. (2021). Can artificial intelligence be decolonized?. *Interdisciplinary Science Reviews*, *46*(1-2), 176-197.

Akter, S., McCarthy, G., Sajib, S., Michael, K., Dwivedi, Y. K., D'Ambra, J., & Shen, K. N. (2021). Algorithmic bias in data-driven innovation in the age of AI. *International Journal of Information Management*, *60*, 102387.

Alfano, M., Carter, J. A., & Cheong, M. (2018). Technological seduction and self-radicalization. Journal of the American Philosophical Association, 4(3), 298–322.

Amsterdam UMC. (2019, March 27). *Pacmed en amsterdam UMC gaan met machine learning IC-Zorg Verbeteren*. Amsterdam UMC, Locatie VUmc. https://www.vumc.nl/nieuws/nieuwsdetail/pacmed-en-amsterdam-umc-gaan-met-machine-learning-ic-zorg-verbeteren.htm

Anguiano, D. (2023, September 27). *Hollywood writers agree to end five-month strike after new studio deal*. The Guardian. https://www.theguardian.com/culture/2023/sep/26/hollywood-writers-strike-ends-studio-deal

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2022). Machine bias. In *Ethics of data and analytics* (pp. 254-264). Auerbach Publications.

Arnstein, S. R. 1969. A ladder of citizen participation. *Journal of the American Institute of planners,* 35(4): 216-224.

Aytac, U. (2022). Digital Domination: Social Media and Contestatory Democracy. *Political Studies*, 00323217221096564.

Banerjee, D., & Rao, T. S. (2022). The dark shadow of marital rape: Need to change the narrative. *Journal of psychosexual health*, *4*

Baehr, J. (2013). Educating for Intellectual Virtues: From Theory to Practice. Journal of Philosophy of Education, 47(2), 248–262. https://doi.org/10.1111/1467-9752.12023

Baker, W. D. W. P. of P. R. (2019). *The Structure of Moral Revolutions: Studies of Changes in the Morality of Abortion, Death, and the Bioethics Revolution*.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, *104*, 671.

Bartneck, C. Lütge, C. Wagner, A. & Welsh, S. (2021). Privacy issues of AI. In *An introduction to ethics in robotics and AI*, edited by C. Bartneck, C. Lütge, A. Wagner, & S. Welsh, 61-70. Springer.

Benn, C., & Lazar, S. (2022). What's wrong with automated influence. *Canadian Journal of Philosophy*, *52*(1), 125-148.

Benjamin, R. (2019). Assessing risk, automating racism. *Science*, *366*(6464), 421-422.

Benjamin, R. (2020). *Race After Technology: Abolitionist Tools for the New Jim Code.* Polity.

Bennett, M., & Claassen, R. (2022). Taming the Corporate Leviathan: How to Properly Politicise Corporate Purpose?. In *Wealth and Power* (pp. 145-165). Routledge.

Berlin, I., 1969, 'Two Concepts of Liberty', in I. Berlin, *Four Essays on Liberty*, London: Oxford University Press. New ed. in Berlin 2002.

BEUC. (2022, September 28). *EU liability rules to be modernised but contain AI services blind spot for consumers*. https://www.beuc.eu/press-releases/eu-liability-rules-be-modernised-contain-ai-services-blind-spot-consumers

Bijker, W. (1995). Sociohistorical Technology Studies. In *Handbook of Science and Technology Studies.* S. Jasanoff, G.E. Markle, J. C. Petersen and Pinch, T. (Eds.). Sage Publications.

Birhane, A., Isaac, W., Prabhakaran, V., Diaz, M., Elish, M. C., Gabriel, I., & Mohamed, S. (2022, October). Power to the people? Opportunities and challenges for participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (pp. 1-8).

Birhane, A., Ruane, E., Laurent, T., S. Brown, M., Flowers, J., Ventresque, A., & L. Dancy, C. (2022, June). The forgotten margins of AI ethics. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 948-958).

Blair, A. (Ed.). (2019). Studies in Critical Thinking. University of Windsor. https://doi.org/10.22329/wsia.08.2019

Bolte, L., Vandemeulebroucke, T., & van Wynsberghe, A. (2022). From an ethics of carefulness to an ethics of desirability: Going beyond current ethics approaches to sustainable AI. *Sustainability*, *14*(8), 4472.

Brandom, R. B. (1994). *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Harvard University Press.

Brandstedt, E. (2014). An interview with professor Simon Caney. *De Ethica*, *1*(1), 71-84.

Bratteteig, T., & Wagner, I. (2012, August). Disentangling power and decision-making in participatory design. In *Proceedings of the 12th Participatory Design Conference: Research Papers-Volume 1* (pp. 41-50).

Bryan, A. 2020. The economic implications of Republican political thought. PhD dissertation, School of Law, King's College London, London, UK.

Brzozowski, J. (2008). On locating composite objects. *Oxford studies in metaphysics*, *4*, 193-222.

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big data & society*, *3*(1), 2053951715622512.

Burri, T. (2022). The New Regulation of the European Union on Artificial Intelligence: Fuzzy Ethics Diffuse into Domestic Law and Sideline International Law. *The Cambridge Handbook of Responsible Artificial Intelligence, CUP*.

Büthe, T., & Mattli, W. (2011). The new global rulers: The privatization of regulation in the world economy. Princeton University Press.

Busuioc, M. (2020). Accountable Artificial Intelligence: Holding Algorithms to Account. *Public Administration Review*. DOI: https://doi.org/10.1111/puar.13293.

Bykvist, K. (2017). Moral uncertainty. *Philosophy Compass*, *12*(3), e12408. https://doi.org/10.1111/phc3.12408

Calvert, S., Mecacci, G., Heikoop, D., & Santoni de Sio, F. (2018). *Full platoon control in Truck Platooning: A Meaningful Human Control perspective* (p. 3326). https://doi.org/10.1109/ITSC.2018.8570013

Cammaerts, B., & Mansell, R. (2020). Digital platform policy and regulation: Toward a radical democratic turn. *International journal of communication*, *14*, 135-154.

Cappelen, H. (2018). *Fixing Language: An Essay on Conceptual Engineering*. Oxford University Press.

Carnap, R. (1962). *Logical Foundations of Probability* (2nd edition). The University of Chicago Press.

Carnegie Council. (2021, November 10). *Why are we failing at the ethics of ai?*. Carnegie Council for Ethics in International Affairs. https://www.carnegiecouncil.org/media/article/why-are-we-failing-at-the-ethics-of-ai

Carter, I. (1999). *A Measure of Freedom*. Oxford: Oxford University press.

Carter, I. (2003). Positive and negative liberty. Retrieved from https://plato.stanford.edu/entries/liberty-positive-negative/

Carter, I. and Kramer, M.H. (2008). How Changes in One's Preferences Can Affect One's Freedom (and How They Cannot): A Reply to Dowding and van Hees. *Economics and Philosophy, 24*(1), 81-96.

Castelfranchi, C. (2003). The micro-macro constitution of power. *Protosociology*, *18*, 208-265. DOI: https://doi.org/10.5840/protosociology200318/198.

Castelfranchi, C. (2011). The" Logic" of Power. Hints on How my Power Becomes his Power. *Proceedings of SNAMAS track within AISB 2011*.

Chaum, D. (1985). Security without identification: Transaction systems to make big brother obsolete. *Communications of the ACM*, *28*(10), 1030-1044.

Chrisman, M. (2015). *The Meaning of "Ought": Beyond Descriptivism and Expressivism in Metaethics*. Oxford University Press.

Cocking, D., & Van den Hoven, J. (2018). *Evil Online*. John Wiley & Sons.

Coglianese, C., & Lehr, D. (2016). Regulating by robot: Administrative decision making in the machine-learning era. *Geo. LJ*, *105*, 1147-1223.

Cohen, J. (2018). Reflections on deliberative democracy. In *Political Philosophy in the Twenty-First Century* (pp. 217-235). Routledge.

Cohen, J. E. (2019). *Between truth and power*. Oxford University Press.

Costanza-Chock, S. (2018). Design justice: Towards an intersectional feminist framework for design theory and practice. *Proceedings of the Design Research Society*.

Costanza-Chock, S. 2020. *Design justice: Community-led practices to build the worlds we need*. The MIT Press.

Crawford, K. (2021). *The Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Crawford, K., & Schultz, J. (2014). Big data and due process: Toward a framework to redress predictive privacy harms. *BCL Rev.*, *55*, 93.

Crawford, K., Dobbe, R., Dryer, T., Fried, G., Green, B., Kaziunas, E., Kak, A. et al. (2019). *AI now 2019 report*. Retrieved December 18, 2019. Retrieved from https://ainowinstitute.org/AI_Now_2019_Report.pdf.

Creamer, E. (2023, September 12). *Elon Musk biographer admits suggestion SpaceX head blocked Ukraine drone attack was wrong*. The Guardian. https://www.theguardian.com/books/2023/sep/12/elon-musk-biographer-admits-suggestion-spacex-head-blocked-ukraine-drone-attack-was-wrong

Dauvergne, P. (2022). Is artificial intelligence greening global supply chains? Exposing the political economy of environmental costs. *Review of International Political Economy*, *29*(3), 696-718.

Davis, J., & Nathan, L. P. (2015). Value sensitive design: Applications, adaptations, and critiques. *Handbook of ethics, values, and technological design: Sources, theory, values and application domains*, 11-40.

De Dijn, A. (2020). *Freedom: An unruly history*. Harvard University Press.

De Zwart, H. (2023, October 28). *Judgement of the Dutch Institute for Human Rights shows how difficult it is to legally prove algorithmic discrimination*. Racism and Technology Center. https://racismandtechnology.center/2023/10/28/judgement-of-dutch-institute-for-human-rights-shows-how-difficult-it-is-to-legally-prove-algorithmic-discrimination/

del Bosque , M. (2023, February 8). *Facial recognition bias frustrates black asylum applicants to us, advocates say*. The Guardian. https://www.theguardian.com/us-news/2023/feb/08/us-immigration-cbp-one-app-facial-recognition-bias

Delgado, F., Yang, S., Madaio, M., & Yang, Q. (2023, October). The participatory turn in ai design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (pp. 1-23).

Denton, E., Hanna, A., Amironesei, R., Smart, A., Nicole, H., & Scheuerman, M. K. (2020). Bringing the people back in: Contesting benchmark machine learning datasets. *arXiv preprint arXiv:2007.07399*.

Diakopoulos, N. (2015). Algorithmic accountability: Journalistic investigation of computational power structures. *Digital journalism*, *3*(3), 398-415.

Dignam, A. (2020). Artificial intelligence, tech corporate governance and the public interest regulatory response. *Cambridge Journal of Regions, Economy and Society*, *13*(1), 37-54.

DiSalvo, C. (2022). *Design as democratic inquiry: putting experimental civics into practice*. MIT Press.

Dobbe, R., Gilbert, T. K., & Mintz, Y. (2021). Hard choices in artificial intelligence. *Artificial Intelligence*, *300*, 103555.

Donia, J., & Shaw, J. A. (2021). Co-design and ethical artificial intelligence for health: An agenda for critical research and practice. *Big Data & Society*, *8*(2), 20539517211065248.

Durán, J. M., & Formanek, N. (2018). Grounds for trust: Essential epistemic opacity and computational reliabilism. *Minds and Machines*, *28*, 645-666.

Dyer-Witheford, N. Kjøsen, A. M. & Steinhoff, J. (2019). *Inhuman Power: Artificial Intelligence and the future of capitalism*. Pluto Press.

Ennis, R. H. (1962). A concept of critical thinking. Harvard Educational Review.

Erman, E., & Furendal, M. (2024). Artificial intelligence and the political legitimacy of global governance. *Political Studies*, *72*(2), 421-441.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

European Commission. (2019). *EGTAI: the ethics guidelines for trustworthy artificial intelligence.* Retrieved from https://ec.europa.eu/futurium/en/ai-alliance-consultation

European Commission. (2021). *Proposal for a Regulation laying down harmonised rules on artificial intelligence*. Proposal for a Regulation laying down harmonised rules on artificial intelligence | Shaping Europe's digital future. Retrieved from https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence

European Commission. (2024, January 26). *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts.* EU Artificial Intelligence Act. https://artificialintelligenceact.eu/the-act/

European Group on Ethics (EGE). (2018, December 19). *Future of Work, Future of Society*. Research and innovation. https://research-and-innovation.ec.europa.eu/knowledge-publications-tools-and-data/publications/all-publications/future-work-future-society_en#files

European Parliament. (2012, October 26). Charter of Fundamental Rights of the European Union. *Official Journal of the European Union*. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:12012P/TXT&from=EN

European Parliament. (2024). *Artificial Intelligence Act*. https://artificialintelligenceact.eu/the-act. Accessed 13-05-2024.

Feldstein, S. (2019). *The global expansion of AI surveillance* (Vol. 17, No. 9). Washington, DC: Carnegie Endowment for International Peace.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press.

Fisher, A. (2019). What critical thinking is. In A. Blair (Ed.), Studies in Critical Thinking (pp. 7–32). University of Windsor.

Fisher III, W. W. (2009). The implications for law of user innovation. *Minn. L. Rev.*, *94*, 1417.

Folland, A. (2022). The Harm Principle and the Nature of Harm. *Utilitas*, *34*(2), 139-153.

Floridi, L. (2019). *The logic of information: A theory of philosophy as conceptual design*. Oxford University Press.

Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. *Machine learning and the city: Applications in architecture and urban design*, 535-545.

Forestal, J. (2022). *Designing for democracy: How to build community in digital environments*. Oxford University Press.

Frankfurt, H. G. (2009). *On bullshit*. Princeton University Press.

Fricker, M. (2007). Epistemic Injustice: Power and the Ethics of Knowing. In *Epistemic Injustice*. Oxford University Press. https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780198237907.001.0001/acprof-9780198237907

Friedman, B., Kahn, P., & Borning, A. (2002). Value sensitive design: Theory and methods. *University of Washington technical report*, (2-12).

Friedman, B., Kahn, P. H., Borning, A., & Huldtgren, A. (2013). Value sensitive design and information systems. *Early engagement and new technologies: Opening up the laboratory*, 55-95.

Friedman, B., Harbers, M., Hendry, D. G., van den Hoven, J., Jonker, C., & Logler, N. (2021). Eight grand challenges for value sensitive design from the 2016 Lorentz workshop. *Ethics and Information Technology*, *23*, 5-16.

Fuchs, C. (2015). The digital labour theory of value and Karl Marx in the age of Facebook, YouTube, Twitter, and Weibo. In *Reconsidering value and labour in the digital age* (pp. 26-41). London: Palgrave Macmillan UK.

Fuchs, C. (2019). *Rereading Marx in the age of digital capitalism.* Pluto Press.

Fukuyama, F., Richman, B., & Goel, A. (2021, January/February). How to save democracy from technology. Retrieved March 04, 2021, from https://www.foreignaffairs.com/articles/united-states/2020-11-24/fukuyama-how-save-democracy-technology?utm_medium=email_notifications&utm_source=reg_confirmation&utm_campaign=reg_guestpass

Gabriel, I. (2022). Toward a theory of justice for artificial intelligence. *Daedalus*, *151*(2), 218-231.

Gädeke, D. (2020). Does a mugger dominate? Episodic power and the structural dimension of domination. *Journal of Political Philosophy*, *28*(2), 199-221.

Galvagna, C. (2023, March 30). *Discussion paper: Inclusive AI governance.* Ada Lovelace Institute. https://www.adalovelaceinstitute.org/report/inclusive-ai-governance/

Gebru, T. (2020). Race and gender. *The Oxford handbook of ethics of AI*, 251-269.

Gebru, T. (2022, May 26). *Disrupting big tech – Timnit Gebru*. YouTube. https://www.youtube.com/watch?v=dENwLu1pQb4&t=4s

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

Gerdes, A. (2022a). The tech industry hijacking of the AI ethics research agenda and why we should reclaim it. *Discover Artificial Intelligence*, *2*(1), 25.

Gerdes, A. (2022b). A participatory data-centric approach to AI ethics by design. *Applied Artificial Intelligence,* 36(1), 2009222.

Gigerenzer, G. (2018). The bias bias in behavioral economics. *Review of Behavioral Economics*, *5*(3-4), 303-336.

Gisondi, M. A., Barber, R., Faust, J. S., Raja, A., Strehlow, M. C., Westafer, L. M., & Gottlieb, M. (2022). A deadly infodemic: social media and the power of COVID-19 misinformation. *Journal of medical Internet research*, *24*(2), e35552.

Goodman, N. (1978). *Ways of Worldmaking* (First Printing edition). Hackett Publishing Company, Inc.

Gould, C. C. (1990). *Rethinking democracy: Freedom and social co-operation in politics, economy, and society*. Cambridge University Press.

Gould, C. C. (2019). How democracy can inform consent: cases of the Internet and bioethics. *Journal of Applied Philosophy*, *36*(2), 173-191.

Greenbaum, J. (2017). A design of one's own: towards participatory design in the United States. In *Participatory Design* (pp. 27-37). CRC Press.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI—Explainable artificial intelligence. *Science robotics*, *4*(37), eaay7120.

Habermas, J. (2022). *Ein neuer Strukturwandel der Öffentlichkeit und die deliberative Politik: Platz 1 der Sachbuchbestenliste der WELT*. Suhrkamp Verlag.

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, *30*(1), 99-120. DOI: https://doi.org/10.1007/s11023-020-09517-8

Haslanger, S. (2012). *Resisting Reality: Social Construction and Social Critique*. OUP USA.

Hatta, M. (2020). The right to repair, the right to tinker, and the right to innovate. *Annals of Business Administrative Science*, *19*(4), 143-157.

Haugaard, M. (2012). Rethinking the four dimensions of power: domination and empowerment. *Journal of Political Power, 5*(1), 33-54.

Haugaard, M., & Pettit, P. (2017). A conversation on power and republicanism: an exchange between Mark Haugaard and Philip Pettit. *Journal of Political Power*, *10*(1), 25-39.

Heikkilä, M. (2022, April 13). *Dutch scandal serves as a warning for Europe over risks of using algorithms*. POLITICO. https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/

Helbing, D., Frey, B.S., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., Van den Hoven, J., Zicari, R.V. & Zwitter, A. (2019). Will democracy survive big data and artificial intelligence?. In *Towards digital enlightenment* (pp. 73-98). Cham: Springer.

Hevelke, A., & Nida-Rümelin, J. (2015). Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis. *Science and Engineering Ethics*, *21*(3), 619–630. https://doi.org/10.1007/s11948-014-9565-5

Hildebrandt, M. (2011). Who needs stories if you can get the data? ISPs in the era of big number crunching. *Philosophy & Technology, 24*(4), 371–390. https://doi.org/10.1007/s13347-011-0041-8

Hildebrandt M (2021). The issue of bias. The framing powers of machine learning. In *Machines We Trust. Perspectives on Dependable AI.* Pelillo, M., Scantamburlo, T. (Eds.). MIT Press.

Himmelreich, J. (2019). Responsibility for Killer Robots. *Ethical Theory and Moral Practice*, *22*(3), 731–747. https://doi.org/10.1007/s10677-019-10007-9

Himmelreich, J. (2023). Against "democratizing AI". *AI & SOCIETY*, *38*(4), 1333-1346.

Hirota, Y., Nakashima, Y., & Garcia, N. (2022). Gender and Racial Bias in Visual Question Answering Datasets. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1280–1292. https://doi.org/10.1145/3531146.3533184

Hitchcock, D. (2018). Critical Thinking. In Edward N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (2018th ed.). Metaphysics Research Lab, Stanford University.

Hodder, I. (2014). The entanglements of humans and things: A long-term view. *New literary history*, *45*(1), 19-36.

Hoeksema, B. (2023). Digital Domination and the Promise of Radical Republicanism. *Philosophy & Technology*, *36*(1), 17.

Hofmann, B., Haustein, D., & Landeweerd, L. (2017). Smart-Glasses: Exposing and Elucidating the Ethical Issues. *Science and Engineering Ethics*, *23*(3), 701–721. https://doi.org/10.1007/s11948-016-9792-z

Hussain, W. (2023). *Living with the Invisible Hand: Markets, Corporations, and Human Freedom*. Oxford University Press.

Ienca, M. (2023). On Artificial Intelligence and Manipulation. *Topoi*, *42*(3), 833-842.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389-399. DOI: https://doi.org/10.1038/s42256-019-0088-2

Johnson, D. G. (1997). Is the global information infrastructure a democratic technology?. *Acm Sigcas Computers and Society*, *27*(3), 20-26.

Johnson, D. G. (2007). Democracy, technology, and information societies. In *The Information Society: Innovation, Legitimacy, Ethics and Democracy In honor of Professor Jacques Berleur sj: Proceedings of the Conference "Information Society: Governance, Ethics and Social Consequences", University of Namur, Belgium 22–23 May 2006* (pp. 5-16). Springer US.

Johnson, D. G. (2021). Algorithmic accountability in the making. *Social Philosophy and Policy*, *38*(2), 111-127.

Johnson, D. G., & Verdicchio, M. (2024). The sociotechnical entanglement of AI and values. *AI & SOCIETY*, 1-10.

Johnson, S., & Acemoglu, D. (2023). *Power and progress: Our thousand-year struggle over technology and prosperity*. Hachette UK.

Johnstone, J. (2007). Technology as empowerment: a capability approach to computer ethics. Ethics and Information Technology, 9(1), 73–87. https://doi.org/10.1007/s10676-006-9127-x

Jolly, J. (2020, May 25). *German court rules against Volkswagen in 'dieselgate' scandal*. The Guardian. https://www.theguardian.com/business/2020/may/25/german-court-rules-against-volkswagen-dieselgate-scandal.

Joss, S., & Bellucci, S. (2002). Participatory technology assessment. *European Perspectives. London: Center for the Study of Democracy*.

Kafali, Ö., Ajmeri, N., & Singh, M. P. (2019). DESEN: Specification of sociotechnical systems via patterns of regulation and control. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, *29*(1), 1-50.

Kalluri, P. (2020). Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature*, *583*(7815), 169-169.

Keyes, O., Hutson, J., & Durbin, M. (2019, May). A mulching proposal: Analysing and improving an algorithmic system for turning the elderly into high-nutrient slurry. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems* (pp. 1-11).

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2018). Discrimination in the Age of Algorithms. *Journal of Legal Analysis*, *10*, 113-174.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2020). Algorithms as discrimination detectors. *Proceedings of the National Academy of Sciences*, *117*(48), 30096-30100.

Kramer, M. H. (2003). *The Quality of Freedom*. Oxford: Oxford University Press.

Kranzberg, M. (1986). Technology and history: "Kranzberg's laws". *Technology and culture*, *27*(3), 544-560.

Kudina, O., & Verbeek, P. P. (2019). Ethics from within: Google Glass, the Collingridge dilemma, and the mediated value of privacy. *Science, Technology, & Human Values*, *44*(2), 291-314.

Laborde, C., & Maynor, J. (Eds.). (2009). *Republicanism and political theory*. John Wiley & Sons.

Laborde, C. (2010). Republicanism and global justice: a sketch. *European journal of political theory, 9*(1), 48-69.

Laidler, J. (2019, March 4). *Harvard professor says surveillance capitalism is undermining democracy*. Harvard Gazette. https://news.harvard.edu/gazette/story/2019/03/harvard-professor-says-surveillance-capitalism-is-undermining-democracy/

Laux, J., Wachter, S., & Mittelstadt, B. (2024). Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance, 18*(1), 3-32.

Lazar, S. (2023, January 25). *Lecture I: Governing the algorithmic city*. YouTube. Retrieved April 24, 2023, from https://www.youtube.com/watch?v=MzRWdpB39qw

Lazar, S. (2024). Automatic Authorities: Power and AI. *arXiv preprint arXiv:2404.05990*.

Leavitt, S. (2006). "A private little revolution": The home pregnancy test in American culture. *Bulletin of the History of Medicine*, *80*(2), 317–345. https://doi.org/10.1353/bhm.2006.0064

Lehtonen, M. (2004). The environmental–social interface of sustainable development: capabilities, social capital, institutions. *Ecological economics*, *49*(2), 199-214.

Lente, H. van, Swierstra, T., & Joly, P.-B. (2017). Responsible innovation as a critique of technology assessment. *Journal of Responsible Innovation*, *4*(2), 254–261. https://doi.org/10.1080/23299460.2017.1326261

Lipton, Z. C. (2018). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31-57.

List, C. (2006). Republican freedom and the rule of law. *Politics, Philosophy & Economics*, *5*(2), 201-220.

List, C., & Valentini, L. (2016). Freedom as independence. *Ethics*, *126*(4), 1043-1074.

Lovett, F. (2010). *A general theory of domination and justice*. Oxford University Press.

Lovett, F. (2012). What counts as arbitrary power?. *Journal of Political Power*, *5*(1), 137-152.

Lukes, S. (2021). *Power: A radical view*. Bloomsbury Publishing.

Luitse, D., & Denkena, W. (2021). The great transformer: Examining the role of large language models in the political economy of AI. *Big Data & Society*, *8*(2), 20539517211047734.

Lutzke, L., Drummond, C., Slovic, P., & Árvai, J. (2019). Priming critical thinking: Simple interventions limit the influence of fake news about climate change on Facebook. *Global Environmental Change, 58*. https://doi.org/10.1016/j.gloenvcha.2019.101964

Ma, A. (2018, March 19). *Everyone is talking about Cambridge Analytica, the trump-linked data firm that harvested 50 million facebook profiles - here's what's going on*. Business Insider Nederland. https://www.businessinsider.nl/cambridge-analytica-a-guide-to-the-trump-linked-data-firm-that-harvested-50-million-facebook-profiles-2018-3?international=true&r=US

Maas, J. J. C. (2020, September 22). *The Power of Tech Companies: Towards a non-dominating technology sector* [Master's Thesis]. University of Twente. http://essay.utwente.nl/83612/

Maas, J. (2022). A Neo-republican Critique of AI Ethics. *Journal of Responsible Technology*, 100022.

Maas, J. 2023. Machine learning and power relations. *AI & SOCIETY*, *38*(4): 1493-1500. doi.org/10.1007/s00146-022-01400-7.

Macnish, K., & Galliott, J. (Eds.). (2020). *Big Data and Democracy*. Edinburgh University Press.

Madiega, T. (2023, February). *Artificial Intelligence Liability directive - European Parliament*. Artificial Intelligence Liability Directive. https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/739342/EPRS_BRI(2023)739342_EN.pdf

Marwick, A. E., & boyd, d. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New media & society*, *13*(1), 114-133.

Marx, K. (1859). *A Contribution to the Critique of Political Economy*. Translated by S. W. Ryazanskaya. London: Lawrence & Wishart.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and information technology*, *6*(3), 175-183. https://doi.org/10.1007/s10676-004-3422-1

Mau, S. (2023). *Mute compulsion: A Marxist theory of the economic power of capital*. Verso Books.

Mayer, C. (2018). *Prosperity: Better business makes the greater good*. Oxford University Press.

Mhlambi, S., & Tiribelli, S. (2023). Decolonizing AI ethics: Relational autonomy as a means to counter AI Harms. *Topoi*, 1-14.

Milana, C., & Ashta, A. (2021). Artificial intelligence techniques in finance and financial markets: a survey of the literature. *Strategic Change*, *30*(3), 189-209.

Mill, J. S. (1998). *On liberty and other essays*. Oxford University Press, USA.

Mill, J. S. (2003). *On liberty*. New Haven: Yale University Press.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature machine intelligence*, *1*(11), 501-507.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2). DOI: https://doi.org/10.1177/2053951716679679

Moloney, P. (2011). John Stuart Mill on savagery, slavery and civilization. *Slavery and Civilization*.

Mosco, V. (2009). *The Political Economy of Communication.* London: Sage

Morozov, E. (2014). *To save everything, click here: Technology, Solutionism and the urge to fix problems that don't exist*. Penguin Books.

Mozur, P. (2018, October 15). *A genocide incited on Facebook, with posts from Myanmar's military*. The New York Times. https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html

Muldoon, J., & Raekstad, P. (2023). Algorithmic domination in the gig economy. *European Journal of Political Theory*, *22*(4), 587-607.

Muller, M., & Liao, Q. V. (2017). Exploring ai ethics and values through participatory design fictions. *Human Computer Interaction Consortium*.

Naeem, S. B., Bhatti, R., & Khan, A. (2021). An exploration of how fake news is taking over social media and putting public health at risk. *Health Information & Libraries Journal*, *38*(2), 143-149.

Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376*(2133), 20180089.

Netherlands Institute for Human Rights. (2023). *Drie Oordelen in Zaken Van Toeslagenouders: Geen directe maar wel indirecte discriminatie door Belastingdienst/Toeslagen*. Nieuwsbericht | College voor de Rechten van de Mens. https://www.mensenrechten.nl/actueel/nieuws/2023/10/2/toeslagen-oordelen

Neyland, D., & Möllers, N. (2017). Algorithmic IF… THEN rules and the conditions and consequences of power. *Information, Communication & Society*, *20*(1), 45-62.

Nickel, P. J., Kudina, O., & Poel, I. van de. (2020). *Moral Uncertainty in Technomoral Change: Bridging the Explanatory Gap*.

Noble, S. U. (2018). *Algorithms of oppression*. New York: New York University Press.

Noorman, M., & Swierstra, T. (2023). Democratizing AI from a sociotechnical perspective. *Minds and Machines*, 1-24.

Nussbaum, M. C. (2007). *Frontiers of justice: Disability, nationality, species membership*. Harvard University Press.

Nyholm, S. (2018). The ethics of crashes with self-driving cars: A roadmap, I. *Philosophy Compass*, *13*(7), e12507. https://doi.org/10.1111/phc3.12507

Nyholm, S. (2023). Minding the Gap (s): Different Kinds of Responsibility Gaps Related to Autonomous Vehicles and How to Fill Them. In *Connected and Automated Vehicles: Integrating Engineering and Ethics* (pp. 1-18). Cham: Springer Nature Switzerland.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447-453.

Oosterlaken, I., & Van den Hoven, J. (2011). ICT and the capability approach. *Ethics and Information Technology, 13*, 65-67.

Owen, R., Macnaghten, P., & Stilgoe, J. (2020). Responsible research and innovation: From science in society to science for society, with society. In *Emerging Technologies: Ethics, Law and Governance* (pp. 117-126). Routledge.

Ovide, S. (2021, February 16). The state house versus big tech. Retrieved March 05, 2021, from https://www.nytimes.com/2021/02/16/technology/the-state-house-versus-big-tech.html

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. London: Penguin.

Perarnaud, C. (2023, April 25). *With the AI act, we need to mind the standards gap – CEPS*. CEPS. https://www.ceps.eu/with-the-ai-act-we-need-to-mind-the-standards-gap/

Perrigo, B. (2023, June 20). *Exclusive: Openai lobbied E.U. to water down AI regulation*. Time. https://time.com/6288245/openai-eu-lobbying-ai-act/

Pettit, P. (1996). Freedom as Antipower. *Ethics*, *106*(3), 576–604.

Pettit, P. (1997). *Republicanism: A theory of freedom and government*. Oxford: Oxford University Press.

Pettit, P. (2002). Keeping republican freedom simple: on a difference with Quentin Skinner. *Political theory*, *30*(3), 339-356.

Pettit, P. (2005). The domination complaint. *Nomos*, *46*, 87-117.

Pettit, P. (2009). Law and liberty. In *Legal republicanism: national and international perspectives.* Besson, S., & Martí, J. L. (Eds.). Oxford University Press.

Pettit, P. (2011). The instability of freedom as noninterference: the case of Isaiah Berlin. *Ethics*, *121*(4), 693-716.

Pettit, P. (2012). *On the People's Terms. A Republican Theory and Model of Democracy*. Cambridge: Cambridge University Press

Pogge, T. (1989). *Realizing Rawls*. Cornell University Press.

Ponce, A. (2021). The AI Regulation: entering an AI regulatory winter? Why an ad hoc directive on AI in employment is required. *Why an ad hoc directive on AI in employment is required (June 25, 2021). ETUI Research Paper-Policy Brief*.

Powles, J., & Nissenbaum, H. (2018, December 7). *The seductive diversion of "solving" bias in artificial intelligence*. Medium. https://onezero.medium.com/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53

Putnam, H. (1988). *Representation and Reality* (1st edition). Mit Pr.

Rahwan, I. (2018). Society-in-the-loop: programming the algorithmic social contract. *Ethics and information technology*, *20*(1), 5-14.

Rawls, J. (1999). *A Theory of Justice*. Oxford University Press.

Rambachan, A., Kleinberg, J., Mullainathan, S., & Ludwig, J. (2020). *An economic approach to regulating algorithms* (No. w27111). National Bureau of Economic Research.

Rességuier, A., & Rodrigues, R. (2020). AI ethics should not remain toothless! A call to bring back the teeth of ethics. *Big Data & Society*, *7*(2), 2053951720942541.

Richardson, H. S. (2002). *Democratic autonomy: Public reasoning about the ends of policy*. Oxford University Press, USA.

Risse, M. (2023). *Political Theory of the Digital Age: Where Artificial Intelligence Might Take Us*. Cambridge University Press.

Robertson, T., & Simonsen, J. (2012). Challenges and opportunities in contemporary participatory design. *Design Issues*, *28*(3), 3-9.

Robeyns, I. (2005). The capability approach: a theoretical survey. *Journal of human development*, *6*(1), 93-117.

Roose, K. (2023, January 12). *Don't ban chatgpt in schools. teach with it.* The New York Times. https://www.nytimes.com/2023/01/12/technology/chatgpt-schools-teachers.html

Rudy-Hiller, F. (2018). The Epistemic Condition for Moral Responsibility. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2018). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2018/entries/moral-responsibility-epistemic/

Rushkoff, D. (2010). *Program or be programmed: Ten commands for a digital age*. Or Books.

Sander, B. (2019). Freedom of expression in the age of online platforms: The promise and pitfalls of a human rights-based approach to content moderation. *Fordham Int'l LJ*, *43*, 939.

Santoni de Sio, F., & Van den Hoven, J. (2018). Meaningful Human Control over Autonomous Systems: A Philosophical Account. *Frontiers in Robotics and AI*, *5*. https://doi.org/10.3389/frobt.2018.00015

Santoni de Sio, F., & Mecacci, G. (2021). Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. *Philosophy & Technology*, 1-28. DOI: https://doi-org.tudelft.idm.oclc.org/10.1007/s13347-021-00450-x

Sartori, L., & Bocca, G. (2023). Minding the gap (s): public perceptions of AI and socio-technical imaginaries. *AI & society*, *38*(2), 443-458.

Scantamburlo, T., & Grandi, G. (2023). A 'Little Ethics' for Algorithmic Decision-Making. *3442*, *CEUR Workshop Proceedings*.

Schaake, M. (2024). *The Tech Coup: How to Save Democracy from Silicon Valley*. Princeton University Press.

Schaffer, J. (2016). Grounding in the image of causation. *Philosophical studies*, *173*(1), 49-100.

Scholz, T. (2016). Platform cooperativism: Challenging the corporate sharing economy.

Schyns, C. (2023, February). *The lobbying ghost in the Machine*. Corporate Europe Observatory. https://corporateeurope.org/en/2023/02/lobbying-ghost-machine

Schuler, D., & Namioka, A. (Eds.). (1993). *Participatory design: Principles and practices*. CRC press.

Sclove, R. (1995). *Democracy and technology*. Guilford Press.

Searle, J. R. (1995). *The construction of social reality*. Simon and Schuster.

Seger, E., Ovadya, A., Garfinkel, B., Siddarth, D., & Dafoe, A. (2023). Democratising AI: Multiple Meanings, Goals, and Methods. *arXiv preprint arXiv:2303.12642*.

Segun, S. T. (2021). Critically engaging the ethics of AI for a global audience. *Ethics and Information Technology*, 99-105.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019, January). Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 59-68).

Sen, A. (1983). Liberty and social choice. *The Journal of Philosophy*, *80*(1), 5-28.

Sen, A. K. 2001. *Development as freedom*. Oxford University Press.

Shapiro, I. (2012). On non-domination. *University of Toronto Law Journal, 62*(3), 293-336.

Sharon, T. (2021). From hostile worlds to multiple spheres: towards a normative pragmatics of justice for the Googlization of health. *Medicine, Health Care and Philosophy*, *24*(3), 315-327.

Siegmann, C., & Anderljung, M. (2022). The Brussels effect and artificial intelligence: How EU regulation will impact the global AI market. *arXiv preprint arXiv:2208.12645*.

Simonsen, J., & Robertson, T. (Eds.). (2012). *Routledge international handbook of participatory design*. Routledge.

Singh, M. P. (2014). Norms as a basis for governing sociotechnical systems. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *5*(1), 1-23.

Skinner, Q. (1998). *Liberty before Liberalism*, Cambridge: Cambridge University Press.

Skinner, Q. (2002). 'A Third Concept of Liberty', *Proceedings of the British Academy*, 117(237): 237–68.

Sloane, M., & Moss, E. (2019). AI's social sciences deficit. *Nature Machine Intelligence*, *1*(8), 330-331.

Sloane, M., Moss, E., Awomolo, O., & Forlano, L. (2022, October). Participation is not a design fix for machine learning. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (pp. 1-6).

Sparrow, R. (2007). Killer Robots. *Journal of Applied Philosophy*, *24*(1), 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x

Spelliscy, C., Hubbard, S., Schneider, N., & Vance-Law, S. (2024). Toward Equitable Ownership and Governance in the Digital Public Sphere. *Stan. J. Blockchain L. & Pol'y*, *7*, 110.

Srnicek, N. (2018). Platform monopolies and the political economy of AI. In *Economics for the Many* (pp. 152-163). Verso.

Stanley, J. (2015). *How propaganda works*. Princeton University Press.

Stilgoe, J. (2023). We need a Weizenbaum test for AI. *Science*, *381*(6658), eadk0176.

Sunstein, C. R. (2018). Is social media good or bad for democracy. *SUR-Int'l J. on Hum Rts.*, *27*, 83.

Susskind, J. (2022). *The digital republic: on freedom and democracy in the 21st century*. Bloomsbury publishing.

Swierstra, T. (2015). Identifying the normative challenges posed by technology's 'soft' impacts. *Etikk i Praksis - Nordic Journal of Applied Ethics*, *1*, 5–20. https://doi.org/10.5324/eip.v9i1.1838

Swierstra, T., Stemerding, D., & Boenink, M. (2009). Exploring Techno-Moral Change: The Case of the ObesityPill. In P. Sollie & M. Düwell (Eds.). *Evaluating New Technologies: Methodological Problems for the Ethical Assessment of Technology Developments.* (pp. 119–138). Springer Netherlands. https://doi.org/10.1007/978-90-481-2229-5_9

Tamahori, L. (1994). *Once Were Warriors*. Fine Line Features.

Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.

Torrance, A. W., & Von Hippel, E. (2015). The right to innovate. *Mich. St. L. Rev.*, 793.

Trogdon, K. (2018). Inheritance arguments for fundamentality. *Reality and its structure: Essays in fundamentality*, 182-98.

Umbrello, S., & van de Poel, I. (2021). Mapping value sensitive design onto AI for social good principles. *AI and Ethics*, 1-14. DOI: https://doi-org.tudelft.idm.oclc.org/10.1007/s43681-021-00038-3.

Urbinati, N. (2002). *Mill on democracy: from the Athenian polis to representative government.* University of Chicago Press.

Vaassen, B. (2022). AI, Opacity, and Personal Autonomy. *Philosophy & Technology*, *35*(4), 88.

Van den Hoven, J. (2017). The Design Turn in Applied Ethics. In J. Van den Hoven, S. Miller, & T. Pogge (Eds.). *Designing in Ethics* (pp. 11–31). Cambridge University Press. https://doi.org/10.1017/9780511844317.002

Van den Hoven, M. J. (2017[1994]). Towards ethical principles for designing politico-administrative information systems. In *Computer Ethics* (pp. 193-213). Routledge.

Van de Poel, I. (2013). Why New Technologies Should be Conceived as Social Experiments. *Ethics, Policy & Environment*, *16*(3), 352–355. https://doi.org/10.1080/21550085.2013.844575

Van Maanen, G. (2022). AI ethics, ethics washing, and the need to politicize data ethics. *Digital Society*, *1*(2), 9.

Veluwenkamp, H. (2021). Inferentialist Truth Pluralism. *Ethical Theory and Moral Practice*. https://doi.org/10.1007/s10677-020-10145-5

Veluwenkamp, H., & van den Hoven, J. (2023). Design for values and conceptual engineering. *Ethics and Information Technology*, *25*(1), 2.

Viganò, E. (2023). The Right to be an Exception to Predictions: a Moral Defense of Diversity in Recommendation Systems. *Philosophy & Technology*, *36*(3), 59.

Voinea, C., Vică, C., Mihailov, E., & Savulescu, J. (2020). The Internet as Cognitive Enhancement. *Science and Engineering Ethics.* Advance online publication. https://doi.org/10.1007/s11948-020-00210-8

Waelen, R. (2022). Why AI ethics is a critical theory. *Philosophy & Technology*, *35*(1), 9.

Wasko, J. (2004). The Political Economy of Communications. In *The SAGE handbook of media studies,* McQuail, D., Schlesinger, P., Downing, J. D., & Wartella, E. (Eds.). Sage Publications.

Wedgwood, R. (2001). Conceptual Role Semantics for Moral Terms. *Philosophical Review*, *110*(1), 1–30. https://doi.org/10.1215/00318108-110-1-1

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., ... & Schwartz, O. (2018). *AI now report 2018* (pp. 1-62). New York: AI Now Institute at New York University.

Wieringa, M. (2020, January). What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 1-18). DOI: https://doi-org.tudelft.idm.oclc.org/10.1145/3351095.3372833.

Williams, J. (2018). *Stand out of our light: Freedom and resistance in the attention economy*. Cambridge: Cambridge University Press.

Winner, L. (1980). Do artifacts have politics?. *Daedalus*, 121-136.

Winner, L. (1987). Political Ergonomics: Technological Design and the Quality of Public Life. *Wissenschaftszentrum Berlin für Sozialforschung*. IIUG dp. 87-7.

Winner, L. (1993). Upon opening the black box and finding it empty: Social constructivism and the philosophy of technology. *Science, technology, & human values*, *18*(3), 362-378.

Wylie, C. (2019). *Mindf*ck*. Profile Books.

Yeung, K. (2017). 'Hypernudge': Big Data as a mode of regulation by design. *Information, Communication & Society*, 20(1): 118-136.

Young, I. M. (1990). *Justice and the Politics of Difference*. Princeton University Press.

Yudkowsky, E. (2004). Coherent extrapolated volition. *Singularity Institute for Artificial Intelligence*

Zerilli, J. (2021). *A citizen's guide to artificial intelligence*. MIT Press.

Zhang, Y., Gowder, P., & Rogers, B. (2023, September 18). *Dismantle or democratize big tech?*. LPE Project. https://lpeproject.org/blog/dismantle-or-democratize-big-tech/

Zimmermann, A., Di Rosa, E., and Kim, H. 2020. Technology can't fix algorithmic injustice. Boston Rev. http://bostonreview.net/science-nature-politics/annette-zimmermann-elena-di-rosa-hochan-kim-technology-cant-fx-algorithmic. Accessed: 2024-03-12.

Zuboff, S. (2016). The Secrets of Surveillance Capitalism. Retrieved from: https://9x0rg.com/Shoshana-Zuboff-Secrets-of-Surveillance-Capitalism.pdf.

Zuboff, S. (2019a). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power.* New York: Public Affairs.

Zuboff, S. (2019b). "We Make Them Dance": Surveillance Capitalism, the Rise of Instrumentarian Power, and the Threat to Human Rights. In *Human Rights in the Age of Platforms*. Jørgensen, R.F. (Ed.), 3-51.

The Simon Stevin Series in Ethics of Technology is an initiative of the 4TU Centre for Ethics and Technology. 4TU.Ethics is a collaboration between Delft University of Technology, Eindhoven University of Technology, University of Twente, and Wageningen University & Research. Contact: info@ethicsandtechnology.eu

Books and Dissertations

Volume 1: Lotte Asveld, '*Respect for Autonomy and Technology Risks*', 2008

Volume 2: Mechteld-Hanna Derksen, *'Engineering Flesh, Towards Professional Responsibility for 'Lived Bodies' in Tissue Engineering',* 2008

Volume 3: Govert Valkenburg, *'Politics by All Means. An Enquiry into Technological Liberalism'*, 2009

Volume 4: Noëmi Manders-Huits, *'Designing for Moral Identity in Information Technology'*, 2010

Volume 5: Behnam Taebi, *'Nuclear Power and Justice between Generations. A Moral Analysis of Fuel Cycles'*, 2010

Volume 6: Daan Schuurbiers, *'Social Responsibility in Research Practice. Engaging Applied Scientists with the Socio-Ethical Context of their Work'*, 2010

Volume 7: Neelke Doorn, *'Moral Responsibility in R&D Networks. A Procedural Approach to Distributing Responsibilities'*, 2011

Volume 8: Ilse Oosterlaken, *'Taking a Capability Approach to Technology and Its Design. A Philosophical Exploration'*, 2013

Volume 9: Christine van Burken, *'Moral Decision Making in Network Enabled Operations'*, 2014

Volume 10: Faridun F. Sattarov, *'Technology and Power in a Globalising World, A Political Philosophical Analysis'*, 2015

Volume 11: Gwendolyn Bax, *'Safety in large-scale Socio-technological systems. Insights gained from a series of military system studies'*, 2016

Volume 12: Zoë Houda Robaey, *'Seeding Moral Responsibility in Ownership. How to Deal with Uncertain Risks of GMOs'*, 2016

Volume 13: Shannon Lydia Spruit, *'Managing the uncertain risks of nanoparticles. Aligning responsibility and relationships'*, 2017

Volume 14: Jan Peter Bergen, *Reflections on the Reversibility of Nuclear Energy Technologies*, 2017

Volume 15: Jilles Smids, *Persuasive Technology, Allocation of Control, and Mobility: An Ethical Analysis*, 2018

Volume 16: Taylor William Stone, *Designing for Darkness: Urban Nighttime Lighting and Environmental Values*, 2019

Volume 17: Cornelis Antonie Zweistra, *Closing the Empathy Gap: Technology, Ethics, and the Other*, 2019

Volume 18: Ching Hung, *Design for Green: Ethics and Politics for Behavior-Steering Technology*, 2019

Volume 19: Marjolein Lanzing, *The Transparent Self: a Normative Investigation of Changing Selves and Relationships in the Age of the Quantified Self*, 2019

Volume 20: Koen Bruynseels, *Responsible Innovation in Data-Driven Biotechnology*, 2021

Volume 21: Naomi Jacobs, *Values and Capabilities: Ethics by Design for Vulnerable People*, 2021

Volume 22: Melis Baş, *Technological Mediation of Politics. An Arendtian Critique of Political Philosophy of Technology*, 2022

Volume 23: Mandi Astola, *Collective Virtues. A Response to Mandevillian Morality*, 2022

Volume 24: Karolina Kudlek, *The Ethical Analysis of Moral Bioenhancement. Theoretical and Normative Perspectives*, 2022

Volume 25: Chirag Arora, *Responsibilities in a Datafied Health Environment*, 2022

Volume 26: Agata Gurzawska, *Responsible Innovation in Business. A Framework and Strategic Proposal*, 2023

Volume 27: Rosalie Anne Waelen, *The Power of Computer Vision. A Critical Analysis*, 2023

Volume 28: José Carlos Cañizares Gaztelu, *Normativity and Justice in Resilience Strategies*, 2023

Volume 29: Martijn Wiarda, *Responsible Innovation for Wicked Societal Challenges: An Exploration of Strengths and Limitations*, 2023

Volume 30: Leon Walter Sebastian Rossmaier, *mHealth Apps and Structural Injustice*, 2024

Volume 31: Haleh Asgarinia, *Privacy and Machine Learning-Based Artificial Intelligence: Philosophical, Legal, and Technical Investigations*, 2024

Volume 32: Caroline Bollen, *Empathy 2.0: What it means to be empathetic in a diverse and digital world*, 2024

Volume 33: Iris Loosman, *Rethinking Informed Consent in mHealth*, 2024

Volume 34: Benjamin Hofbauer, *Governing Prometheus. Ethical Reflections On Risk & Uncertainty In Solar Climate Engineering Research*, 2024

Volume 35: Madelaine Ley, *It's not (just) about the robots: care and carelessness across an automated supply chain*, 2024

Volume 36: Arthur Gwagwa, *Re-imagining African Unity in a Digitally Interdependent World*, 2024

Volume 37: Jonne Maas, *Freedom in the Digital Age: Designing for Non-Domination*, 2025

Simon Stevin (1548-1620)

'Wonder en is gheen Wonder'

This series in the philosophy and ethics of technology is named after the Dutch / Flemish natural philosopher, scientist and engineer Simon Stevin. He was an extraordinarily versatile person. He published, among other things, on arithmetic, accounting, geometry, mechanics, hydrostatics, astronomy, theory of measurement, civil engineering, the theory of music, and civil citizenship. He wrote the very first treatise on logic in Dutch, which he considered to be a superior language for scientific purposes. The relation between theory and practice is a main topic in his work. In addition to his theoretical publications, he held a large number of patents, and was actively involved as an engineer in the building of windmills, harbours, and fortifications for the Dutch prince Maurits. He is famous for having constructed large sailing carriages.

Little is known about his personal life. He was probably born in 1548 in Bruges (Flanders) and went to Leiden in 1581, where he took up his studies at the university two years later. His work was published between 1581 and 1617. He was an early defender of the Copernican worldview. He died in 1620, but the exact date and the place of his burial are unknown. Philosophically he was a pragmatic rationalist. For him, wonder about a phenomenon, however mysterious, should be the starting point for seeking understanding or even ultimate explanation through human reasoning. Hence the dictum 'Wonder is no Wonder' that he used on the cover of several of his books.

The effects of digital technologies on freedom and democracy have garnered increasing attention in recent years. Many have raised concerns about surveillance capitalism, technofeudalism, and general threats to constitutional democracies—with a special convergence on the worry that uncontrolled power of online platforms undermines people's freedom. However, it remains unclear how 'freedom' should be understood, what the relation is between freedom and uncontrolled power, and to what extent these worries extend beyond online platforms. In this dissertation, I argue that these problems are best answered by appealing to a neo-republican account of freedom as non-domination, where 'domination' is understood as a condition of living under an agent's uncontrolled power. In the context of AI systems used in core societal sectors such as healthcare, I show that domination of a system's (in)direct end-users by the system's developers occurs in at least three ways: (1) the distribution of decision-making power, (2) technical limitations of AI systems, and (3) underlying societal structures that empower developers and disempower end-users. To safeguard freedom in the digital age, I propose that AI development requires the explicit intention to 'design for non-domination'. This requires us to consider the broader societal contexts within which these systems operate, such as current regulatory initiatives and the political economy.

4TU. Centre for Ethics and Technology

TU Delft Technische Universiteit Delft
TU/e Eindhoven University of Technology
University of Twente.
Wageningen University & Research