# Machine learning and power relations

Maas, Jonne

**DOI**
10.1007/s00146-022-01400-7

**Publication date**
2022

**Document Version**
Final published version

**Published in**
AI and Society

**Citation (APA)**
Maas, J. (2022). Machine learning and power relations. *AI and Society*, *38*, 1493-1500.
https://doi.org/10.1007/s00146-022-01400-7

AI & SOCIETY
https://doi.org/10.1007/s00146-022-01400-7

STUDENT FORUM

# Machine learning and power relations

Jonne Maas[1]

Received: 14 June 2021 / Accepted: 7 February 2022

**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.

**Keywords** 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 et al. 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 and Lehr 2016; Crawford 2021; Kalluri 2020; Sloane and Moss 2019). The debate, hence, seems to invoke a moral intuition 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 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)

✉ Jonne Maas
j.j.c.maas@tudelft.nl

[1] Department: Technology, Policy, and Management, Delft University of Technology, Delft, The Netherlands

Published online: 25 February 2022

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 (Sect. 2) and elaborate on the different actors (developers, users, end-users) involved in an ML system (Sect. 3). In Sect. 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 (Sect. 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[1] (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 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. Therefore, contrary to a dictator who has unaccountable power due to lack in effective accountability mechanisms, a democratic government, though 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

[1] 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.

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.[2]

## 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.[3] 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 and 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.[4] 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 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, 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 and Floridi 2016).[5]

Moreover, the involvement of each actor depends on the context and type of algorithm that is developed, so to isolate

[2] 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 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.

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

[4] 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.

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

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.

## 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 to achieve her desired goal G (Castelfranchi 2003, 213). Therefore, 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 et al. 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–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 (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,[6] 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 and 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 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

[6] This softens the claim that the 'shapers' wield power, as any exercise of power 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).

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.[7]

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 et al. 2020; Kleinberg et al. 2018; 2020). Although these scholars do not explicitly address morally problematic power 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

[7] 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.

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,[8] 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 et al. 2002), participatory design (PD) (Simonsen and Robertson 2012) or Responsible Research and Innovation (RRI) (Owen et al. 2020), and other democratic initiatives for technological innovation like participatory Technology Assessment (pTA) (Joss and 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 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 and 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 first 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. Second, 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.

**Acknowledgements** The research reported in this work was partially supported by the EU H2020 ICT48 project „Humane AI Net" under contract # 952026. The support is gratefully acknowledged. I would also like to thank two anonymous reviewers, the editorial board of the special issue Embedding AI in Society, and Jeroen van den Hoven for their insightful comments on earlier drafts of this paper.

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

## References

Busuioc M (2020) Accountable artificial intelligence: holding algorithms to account. Public Adm Rev. https://doi.org/10.1111/puar.13293

Castelfranchi C (2003) The $icro-macro constitution of power. Protosociology 18:208–265. https://doi.org/10.5840/protosociology200318/198

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

Crawford K (2021) The Atlas of AI. 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

Davis J, Nathan LP (2015) Value sensitive design: Applications, adaptations, and critiques. Handbook of ethics, values, and technological design: Sources, theory, values and application domains, 11–40.

Diakopoulos N (2015) Algorithmic accountability: journalistic investigation of computational power structures. Digit Journal 3(3):398–415

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.

Castelfranchi C (2011) The" Logic" of Power. Hints on how my power becomes his power. Proceedings of SNAMAS track within AISB 2011.

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

Gädeke D (2020) Does a mugger dominate? Episodic power and the structural dimension of domination. J Polit Philos 28(2):199–221

Hagendorff T (2020) The ethics of ai ethics: an evaluation of guidelines. Mind Mach 30(1):99–120. https://doi.org/10.1007/s11023-020-09517-8

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

Hildebrandt M (2021). he issue of bias. The framing powers of machine learning. In: Machines We Trust. Perspectives on Dependable AI. MIT Press.

Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. Nat Mach Intell 1(9):389–399. https://doi.org/10.1038/s42256-019-0088-2

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

Kafali Ö, Ajmeri N, Singh MP (2019) DESEN: Specification of sociotechnical systems via patterns of regulation and control. ACM Trans Softw Eng Methodol (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

Kleinberg J, Ludwig J, Mullainathan S, Sunstein CR (2018) Discrimination in the age of algorithms. J Legal Anal 10:113–174

Kleinberg J, Ludwig J, Mullainathan S, Sunstein CR (2020) Algorithms as discrimination detectors. Proc Natl Acad Sci 117(48):30096–30100

Lipton ZC (2018) The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue 16(3):31–57

Lovett F (2010) A general theory of domination and justice. Oxford University Press

Maas J (2022) A neo-republican critique of ai ethics. J Responsible Technol 100022.

Matthias A (2004) The responsibility gap: ascribing responsibility for the actions of learning automata. Ethics Inf Technol 6(3):175–183

Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L (2016) The ethics of algorithms: mapping the debate. Big Data Soc 3(2). https://doi.org/10.1177/2053951716679679.

Neyland D, Möllers N (2017) Algorithmic IF THEN rules and the conditions and consequences of power. Inform Commun Soc 20(1):45-62

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.

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

Pettit P (2011) The instability of freedom as noninterference: the case of Isaiah Berlin. Ethics 121(4):693–716

Pettit P (1997) Republicanism: a theory of freedom and government. 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.

Santoni de Sio F, Mecacci G (2021) Four responsibility gaps with artificial intelligence: Why they matter and how to address them. Philosophy Technol 1–28. DOI: https://doi-org.tudelft.idm.oclc.org/https://doi.org/10.1007/s13347-021-00450-x.

Segun ST (2021) Critically engaging the ethics of AI for a global audience. Ethics Inform Technol 99–105.

Simonsen J, Robertson T (Eds.) (2012) Routledge international handbook of participatory design. Routledge.

Singh MP (2014) Norms as a basis for governing sociotechnical systems. ACM Trans Intell Syst Technol (TIST) 5(1):1–23

Sloane M, Moss E (2019) AI's social sciences deficit. Nat Mach Intell 1(8):330–331

Umbrello S, van de Poel I (2021) Mapping value sensitive design onto AI for social good principles. AI and Ethics 1–14. https://doi-org.tudelft.idm.oclc.org/https://doi.org/10.1007/s43681-021-00038-3.

Whittaker M, Crawford K, Dobbe R, Fried G, Kaziunas E, Mathur V, Schwartz O (2018) AI now report 2018. AI Now Institute at New York University, New York, pp 1–62

Wieringa M (2020) 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/https://doi.org/10.1145/3351095.3372833.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.