frontiers | Frontiers in Human Dynamics

ORIGINAL RESEARCH
published: 12 April 2022
doi: 10.3389/fhumd.2022.703510

Distribution of Forward-Looking Responsibility in the EU Process on AI Regulation

*Maria Hedlund**

Department of Political Science, Lund University, Lund, Sweden

Artificial Intelligence (AI) is beneficial in many respects, but also has harmful effects that constitute risks for individuals and society. Dealing with AI risks is a future-oriented endeavor that needs to be approached in a forward-looking way. Forward-looking responsibility is about who should do what to remedy or prevent harm. With the ongoing EU policy process on AI development as a point of departure, the purpose of this article is to discuss distribution of forward-looking responsibility for AI development with respect to what the obligations entail in terms of burdens or assets for the responsible agents and for the development of AI. The analysis builds on the documents produced in the course of the EU process, with a particular focus on the early role of the European Parliament, the work of the High-Level Expert Group on AI, and the Commission's proposal for a regulation of AI, and problematises effects of forward-looking responsibility for the agents who are attributed forward-looking responsibility and for the development of AI. Three issues were studied: ethics by design, Artificial General Intelligence (AGI), and competition. Overall, the analysis of the EU policy process on AI shows that competition is the primary value, and that the perspective is technical and focused on short-term concerns. As for ethics by design, the question of which values should be built into the technology and how this should be settled remained an issue after the distribution of responsibility to designers and other technical experts. AGI never really was an issue in this policy process, and it was gradually phased out. Competition within the EU process on AI is a norm that frames how responsibility is approached, and gives rise to potential value conflicts.

Keywords: artificial intelligence (AI), forward-looking responsibility, responsibility, European Union (EU), policy process

OPEN ACCESS

Edited by:
Olaf Witkowski,
Cross Labs, Japan

Reviewed by:
Daniel Schiff,
Georgia Institute of Technology, United States
Martin Sand,
Delft University of Technology, Netherlands

***Correspondence:**
Maria Hedlund
maria.hedlund@svet.lu.se

Specialty section:
This article was submitted to Digital Impacts, a section of the journal Frontiers in Human Dynamics

Received: *30 April 2021*
Accepted: *10 March 2022*
Published: *12 April 2022*

Citation:
Hedlund M (2022) Distribution of Forward-Looking Responsibility in the EU Process on AI Regulation. Front. Hum. Dyn. 4:703510. doi: 10.3389/fhumd.2022.703510

INTRODUCTION

Artificial Intelligence (AI) is developing quickly, with applications in practically all areas in society. In health care, image recognition is used for diagnosis and prognostics (Asan and Choudhury, 2021), in the transport sector, automated vehicles are enabled by AI technology (Ma et al., 2020), recommender algorithms are central for the functioning of social media (Bucher, 2018), search engines (Haider and Sundin, 2020), music (Melchiorre et al., 2021), video (Abul-Fottouh et al., 2020), and other media services (Alhamid et al., 2015), to name just a few areas in which the efficiency of AI makes life convenient. However, AI also has harmful effects on society and individuals and so implies risks. Algorithmic bias

and discrimination (O'Neill, 2016; Sirbu et al., 2019), violation of privacy (Stahl, 2016, 2020; Helm, 2018), negative effect on democratic processes (Bartlett, 2018; Deibert, 2019; Stahl, 2020), mass surveillance (Stahl, 2016; Diamond, 2019; Zuboff, 2019), and the potential outsmarting of humans, should AI achieve superintelligence (Bostrom, 2014; Tegmark, 2017; Russell, 2019), are some of the risks connected with AI.

While it is urgent that we as a society deal with these and other risks with AI, it is also important that the responsibility to do that is distributed in a way that enables AI to be developed in a desirable direction (Cath et al., 2018). In the ongoing policy process on regulation of AI technology in the European Union (EU)[1], the goal is to make AI "human-centric" (EC, 2019), and the steps to do so include the allocation of responsibility to different actors to take certain actions needed to make sure that the development of AI is beneficial for humans, society, and the economy. Responsibility so recognized is future-oriented, focusing on who should do what to lead the development in a desirable direction. This is a forward-looking understanding of responsibility, meaning that the responsible agent has an obligation to see to it that some state of affairs will materialize (Van de Poel, 2015a). Forward-looking responsibility could refer to remediation of past harm, or to ensure a good future by preventing future harm. With the ongoing EU policy process on AI development as a point of departure, the purpose of this article is to discuss distribution of forward-looking responsibility for AI development with respect to what the obligations entail in terms of burdens or assets for the responsible agents and for the development of AI.

One reason for concentrating on forward-looking responsibility in this article is that the EU's Responsible Research and Innovation (RRI) approach, which is a model for responsible development of AI, referred to as a vision to be applied to the development of AI systems (Dignum, 2019, p. 49; Gorgoni, 2020), is explicitly forward-looking regarding societal implications of new technology (c.f., Van de Poel and Sand, 2018; Horizon, 2020). Another reason to focus on forward-looking responsibility is that the object of analysis, the EU's ongoing policy process on AI development, is a political process. A political process is inherently future oriented, and ascribing responsibilities to agents in such a process is an act of attributing obligations for the future to those agents (Bexell and Jönsson, 2021, p. 29).

To some extent, AI has become an umbrella term, sometimes referring to digital technology in a general sense (Kim et al., 2021; Piorkowski et al., 2021). In this article, AI refers to "systems that display intelligent behavior by analyzing their environment and taking actions—with some degree of autonomy—to achieve specific goals" (AI HLEG, 2019b, p. 1) and are defined as non-organic cognitive systems that can think and act rationally and similarly to humans (Russell and Norvig, 2010). These understandings of AI underlines that AI differs from other technologies in its ability to "learn" and to act autonomously[2].

The EU policy process on regulation of AI has been in progress since 2016, when the European Parliament called for action on autonomous machines and urged the Commission to take an initiative on regulation of AI (EP, 2016a). Early in 2017, the Commission announced that it had begun work with regulation of AI (EC, 2017), and in the spring of 2018, a High-Level Expert Group on AI (AI HLEG) was set up. In the end of 2018, the AI HLEG released draft ethical guidelines (AI HLEG, 2018), which were the basis of a stakeholder consultation process, the result of which was considered by the High-Level group in the final version of ethical guidelines, published in April 2019 (AI HLEG, 2019a). In the beginning of 2020, the Commission published a White Paper on AI (EC, 2020c), which was also the object of a consultation process (EC, 2020a) and the Proposal for a regulation was released in April 2021 (EC, 2021d)[3]. As this is an ongoing process, presented as an "approach" (EC, 2021d) to AI, no binding decisions have yet been taken. This means that at this stage, it is not possible to tell how responsibility will be implemented in practice.

In the following, a notion of forward-looking responsibility is delineated, pointing at the political aspect of distribution of responsibility and how responsibility can constrain as well as empower agents. Possible justifications for the allocation of forward-looking responsibility are outlined, succeeded by a section on methodology. Next, forward-looking responsibility in the EU process on AI is analyzed in terms of (i) responsibility as remedying harm, illustrated by ethics by design, (ii) responsibility as preventing harm, illustrated by Artificial General Intelligence (AGI), and (iii) competition as constraining and enabling factor for responsibility. Finally, in the conclusion the result of the analysis is summed up, implications for the outlined notion of forward-looking responsibility are touched upon, and options for further research suggested.

FORWARD-LOOKING RESPONSIBILITY

Responsibility is a concept with many meanings depending on context (Sondermann et al., 2017), but a basic distinction is that between backward-looking and forward-looking responsibility (Van de Poel et al., 2015). Backward-looking, or retrospective, responsibility is about blame or praise for something that has

[1]This process, "A European Approach to Artificial Intelligence" (EC, 2018a, 2022), deals with AI technology as such. In parallel, the EU is also dealing with automated vehicles, and with the responsibility of big tech companies in the digital sphere. These processes are related to AI, but not with the AI technology as primary focus, and are not included in the current study (see text footnote 6).

[2]It should be noted that autonomy refers to different things in different contexts. Autonomy within AI and robotics refers to the machine's ability to perceive the environment, learn from experience, and act independently of an external operator (Hakli and Mäkelä, 2019, p. 264). Even when a human has decided what the machine should do, in a technical understanding it is considered autonomous as long as the execution of the task takes place without human guidance (Jain and Prathiar, 2010). From a philosophical perspective, an algorithm following a routine is not autonomous, but is just following directions set up by a human. In contrast, a philosophical, conceptual understanding of autonomy stresses the capacity to act independently without some external power and to make one's own choices (Haselager, 2005, p. 518–519). Hence, machines cannot be responsible, even if they in a technical sense autonomously can make decisions or give recommendations.

[3]This short presentation of the ongoing policy process is not exhaustive. For an overview, see e.g., https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence.

happened and asks who should be accountable (Sneddon, 2005). We are normally accountable for what we have caused, given that we were aware of the moral nature of the action and that we were not acting under coercion or ignorance (Thompson, 1987; Miller, 2001; Douglas, 2009). Forward-looking, or prospective, responsibility deals with the future and can be understood as responsibility as obligation, meaning that an agent "ought to see to it" that some state of affairs is fulfilled (Van de Poel, 2015a, p. 27). This is not a requirement that the responsible agent herself achieves the outcome by her own actions, but refers to the supervisory duties of the responsible agent to bring about a particular outcome (Van de Poel, 2015a, p. 28–29). When someone is responsible in a forward-looking way, she is required to undertake specific tasks to ensure that appropriate actions are taken (Bexell and Jönsson, 2021). The future state of affairs that this requirement refers to could either be about remedying a bad situation that has already happened (Miller, 2001; Young, 2006), for instance paying for diminishing CO_2 emissions, or about realizing a good situation in the future (Erskine, 2014; Van de Poel, 2015a), for instance by leading the development of AI in a safe and democratic direction.

Understanding responsibility as obligation in this way points to how forward-looking responsibility, albeit qualitatively different from backward-looking responsibility, nevertheless has some connection to responsibility in the backward-looking sense. While ensuring a desirable state of affairs is entirely forward-looking, remedying harm is backward-looking to the extent that it deals with something that has happened in the past, even though the perspective is on the future. One implication of this is that an agent with forward-looking responsibility is not necessarily the agent that has caused the harm that needs to be remedied. Forward-looking responsibility may also have a more direct connection to a backward-looking perspective. One argument is that forward-looking responsibility is only meaningful if the agent can be accountable, would she not fulfill her forward-looking obligations (Goodin referred to by Van de Poel, 2015a, p. 40). However, she can be accountable also for other reasons. In other words, forward-looking responsibility does not require backward-looking responsibility, nor does it necessarily imply backward-looking responsibility (Van de Poel, 2015a; Pereboom, 2021; c.f., Alfano, 2021). Moreover, as will be discussed below, assignment of forward-looking responsibility is justified on partly other grounds than responsibility in the backward-looking sense. In this article, focus is primarily on forward-looking responsibility[4].

Another important aspect of forward-looking responsibility is the responsibility of collectives. Collective responsibility is a contested matter, as it is questionable whether collectives qualify as moral agents. Collectives cannot have intentions or care about the outcome of their actions in the same sense as individuals. Moreover, collectives do not have the control over actions that is crucial for moral responsibility (Mäkäle, 2007; Hedlund, 2012). However, for some kinds of collectives such as organized groups, it could be appropriate to talk about collective responsibility, meaning that the collective is responsible as a single body (Thompson, 1987). Organizations such as states, companies, or universities, can make decisions and act in ways that are not reducible to the sum of the actions of the individuals within them (Van de Poel, 2015b, p. 57). When such collectives in some sense have a collective aim (Van de Poel, 2015b, p. 56) and when institutional rules and meaningful decision-making procedures are established (Held, 1970), it is reasonable to assert that they have collective responsibility. However, for collectives with looser ties between the individuals constituting the collective, such as train passengers or networks, the problem of many hands may arise. Consider for instance a network of several contract holders, responsible for separate parts of the same project. While each contract holder is responsible for her particular contribution, the outcome of a collective decision is constituted by these contributions taken together (and may well add up to more than the sum of the individual contributions). Whereas the collective is responsible, none of the individual contract holders could reasonably be responsible for the whole project, and we are faced with the problem of many hands (Van de Poel, 2015b, p. 52; Zwart, 2015).

It is important to bear in mind that being responsible in a forward-looking sense can constrain but also enable or empower agents (Ulbert, 2017). Hence, attributing someone responsibility is a political act "embedded in the role of function of power" (Bexell and Jönsson, 2021, p. 29). As such, it must be justified, and the normative basis for such a justification must be understood in the context in which the distribution of responsibility comes about (Ulbert and Sondermann, 2017). In the context of EU regulation of AI, one norm that seems to be taken as a given is that innovation and competition are primary values. This is a normative context that may affect how responsibility plays out, which will be further elaborated in the analysis.

Dealing with harmful effects of an emerging technology is a future-oriented endeavor, and to be effective, the distribution of forward-looking responsibility should take place in a way that enables the desirable state of affairs, whether it is about remedying harm or about preventing future harm. As a basic requirement, forward-looking responsibility has to be assigned to an actor or a group of actors who has the capacity to act responsibly. In other words, the actor needs to be a moral agent, that is, be aware of the moral nature of an action and be able to make moral judgments (Held, 1970; Hakli and Mäkelä, 2019). A moral agent also needs to be able to act freely, although this freedom always is to some extent constrained by social practices (Sondermann et al., 2017). A further requirement for forward-looking responsibility is that the agent has the ability to act in a way that can bring about the desired outcome (Van de Poel, 2015a, p. 35). This is a requirement that points to a causality condition. Although forward-looking responsibility does not necessarily belong to the agent who has caused some harm to be remedied, or, as in the case of ensuring a good future, does not need to be connected to some past events, forward-looking responsibility brings causality into focus, although in a forward-looking sense. Without the ability to make a difference, forward-looking responsibility would not make much sense, but the prospective perspective of causality

[4]"Responsibility" and "forward-looking responsibility" are used interchangeably when referring to forward-looking responsibility. When referring to backward-looking responsibility, this is explicitly stated.

makes it more adequate to talk about efficacy (Graafland, 2003; c.f., Van de Poel, 2015a).

As mentioned above, attribution of forward-looking responsibility differs from how backward-looking responsibility is attributed, since past events play a different role, if any role at all. While for backward-looking responsibility, it is crucial to tie the agent to some blameworthy or praiseworthy state-of-affairs, for forward-looking responsibility the causal relation to past events does not need to play a role. In contrast, causality related to future events, denoted as efficacy, is pivotal. However, capacity to bring the circumstances in the specific direction that the obligation sets out is a basic requirement for forward-looking responsibility, it does not in itself say anything about which particular agent or group of agents that should be responsible in a forward-looking sense. Many candidates may meet that requirement, and some principles must underpin the allocation of forward-looking responsibility. Cost may be one such principle. While efficiency is the rationale for the capacity principle as it is presented here, it could also be relevant to consider the cost for actors with capacity to bring about the desired outcome (Miller, 2001). This means that under some circumstances, it could be justifiable to allocate responsibility to an agent who, while having the required capacity, is not the most capable among possible responsible agents. Other principles suggested in the responsibility literature include equality (Lake, 2001; Markandya, 2011), desert (Anton, 2015; Messina, 2021), connectedness (Young, 2006; Bexell and Jönsson, 2021), and ability (Graafland, 2003; Hayward, 2012). In the particular context of EU regulation of AI, involving many tech companies and technical experts[5], beyond cost, it could be fruitful to consider contribution, benefit, and position as bases for allocation of forward-looking responsibility.

The *contribution principle* means that an agent who has caused a bad outcome is responsible to put it right (Young, 2006; Bexell and Jönsson, 2021). This principle, often referred to in the debate on climate change mitigation as the principle of "polluter pays" (Page, 2008; e.g., Moellendorf, 2009), pays attention to causal connections between an agent and problematic past events. In the context of AI development, many different actors, from individual AI developers to big tech companies, have directly or indirectly contributed to negative effects of the AI technology, although to different extents. According to the principle, responsibility should be differentiated according to the actual contribution, which makes causal claims crucial (Chasek et al., 2016).

The *benefit principle* states that those who benefit from something that have negative effects for society also should be responsible for these effects, for instance by paying to reduce the negative effects (Jakob et al., 2021). Also this principle is common in the context of climate change mitigation, known as the "beneficiaries pay" principle (Hayward, 2012). As for the contribution principle, to be applicable, the benefit principle requires some estimation of the benefits. While this may be difficult for benefits gained in the past, it is even trickier for benefits in the future. This circumstance is highly valid in the AI context, as it is widely considered impossible to anticipate the future development of science and technology (Carrier, 2021). Still, it could be argued that the benefit principle is justifiable from a moral point of view (Mapel, 2005; Hayward, 2012).

The *position principle* asserts that an agent being in a position that enables her to act, has a moral responsibility to remedy a bad situation or to prevent future harm, regardless of past causality relations (Miller, 2001; Young, 2006; c.f., Nihlén Fahlquist, 2009). Position as a justification for allocating responsibility could be defended by efficiency: agents with relevant resources are arguably best placed to take the necessary steps. In other words, the agent who is able to respond—"response-able"—should be responsible (Fukuda-Parr and McNeill, 2015). The argument for this principle is the seriousness of a particular situation and the urgency of doing something about it. Power, knowledge, and other resources could place an agent in a position that makes her more able to deal with the problem than agents who lack those resources, and according to this principle, this position makes her obligated to act (Gilbert, 1993; Persson et al., 2021). A problem with the position principle is the separation from causal connections to past events, which could constitute a legitimacy problem (Hedlund, 2012). One way to handle this problem is to regard how agents, by being part of social structures that make certain outcomes possible, to varying extents contribute to those outcomes. Moreover, as Young (2011) argues, the moral dimension of solidarity justifies why agents in a position to act should be responsible in a forward-looking way.

Forward-looking responsibility can also be conceptualized in terms of virtues. Responsibility as virtue is a kind of forward-looking responsibility that focuses on the person who acts, rather than on the action itself (Williams, 2008; Schmidt, 2021). Responsibility as virtue pays attention to how to *be* good, not just to act in a good way, and virtues direct attention to "the importance of practicing and cultivating a predisposition to act in a certain way over time" (Nihlén Fahlquist, 2019, p. 215). While fulfilling obligations that has been allocated to you is a way of acting responsibly in a forward-looking way, *taking* responsibility because that is the right thing to do is a way of actively engaging ethically with how things could be done to ensure a better future, that is, to act virtuously (Nihlén Fahlquist, 2019, p. 216). People who have a willingness to take on responsibility are arguably more likely to take on and fulfill responsibility in a forward-looking way (Nihlén Fahlquist, 2015, p. 187). For this reason, virtue ethics could be used as a heuristic to develop virtuous attitudes and to ground the application of responsibility for doing good (Daly, 2021; Schmidt, 2021).

METHODS

Research on responsibility ranges from theoretical deliberation on principles of attribution of responsibility (Graafland, 2003;

[5]While actors from the AI industry, academia, and civil society were involved in the consultation process, and the Responsible Research & Innovation approach (RRI) explicitly emphasizes the importance of a multi-actor and public engagement, in the AI High Level Expert Group on Artificial Intelligence (AI HLEG) there was a dominance of participants from the AI industry (EC Futurium; EC Regexpert; Metzinger, 2019).

Hedlund, 2012; Neuhäuser, 2014; Gunnemyr, 2020) to empirical studies on distribution of responsibility in specific issues such as climate change (Droz, 2021; Persson et al., 2021), biotechnology (Danaher, 2016), or agriculture (Doorn et al., 2011), at global (Moelle, 2017), regional (Karageorgiou, 2019), or national (Alcaniz and Hellwig, 2011) level. This study deals with distribution of forward-looking responsibility in the ongoing policy process on AI in the EU. The empirical material considered covers the main documents produced in the course of this process and which are pertinent for the question of forward-looking responsibility. The documents are public documents that constitute part of early phases of the EU decision-making process on AI in which the European Parliament (EP), the European Commission (EC), and permanent (EGE) and temporary (AI HLEG) expert groups play a key role. At the time of writing, the EC (2021d) proposal is the most recent document. This document constitutes the formal initiative on AI regulation, to which national governments (the Council of the EU) and the EP will take a stand and finally decide on new legislation (European Council). Such processes normally take a couple of years, and while we still do not know what the final outcome will be, the preparatory work of any policy process is of utmost importance for how it will develop, as the framing of an issue in early stages of a policy process often steer the direction of the subsequent process as well as which aspects will be considered, and how (Weiss, 1989; Lindblom and Woodhouse, 1993; Rochefort and Cobb, 1994; Kingdon, 1995; Carragee and Roefs, 2004). The documents used in the analysis are listed in **Table 1**[6].

The documents have been studied regarding risks with AI and how they are proposed to be met by the EU, and in particular how allocation of responsibility plays out. The analytical framework that guided the reading of the policy documents is the concept of forward-looking responsibility as it is defined here: an obligation to see to it that some state of affairs is brought about, and the three principles for distribution of forward-looking responsibility: the contribution principle, the benefit principle, and the position principle (**Table 2**). Forward-looking responsibility as virtue amounts to the willingness of actors to take on responsibility and cannot be regarded as a principle for distribution analogous to contribution, benefit, and position (or cost). It may however play a role in how policy-makers take into account future implementation and how responsibility is approached by agents who are allocated responsibility.

The method applied was a document analysis that combines elements of content analysis and thematic analysis (Bowen, 2009; Mathias, 2021). In the first stage, I skimmed through the documents to identify meaningful and relevant passages. Next, I carefully re-read and reviewed these passages with regard to distribution of forward-looking responsibility for AI development and application and the obligations this gives rise to. These passages pertain to several issues, for instance, autonomous vehicles, human-machine interaction, algorithm bias, transparency, and robot laws. To the analysis, issues were selected based on three aspects of the outlined notion of forward-looking responsibility. First, forward-looking responsibility understood in terms of remedying harm is used to interpret responsibility allocations brought to the fore by the ethics by design approach in the studied policy process. Second, forward-looking responsibility understood in terms of preventing harm is used to discuss the gradual disappearing in the process of the question of Artificial General Intelligence (AGI). Third, drawing on the understanding of how the normative context affects the justification of distribution of responsibility, the norm of competition is discussed as a constraining and enabling factor for forward-looking responsibility.

RESPONSIBILITY AS REMEDYING HARM: ETHICS BY DESIGN

A crucial issue for the responsibility of AI is the design and development of AI systems. Forward-looking responsibility might be of special importance at this stage, as this is when the character of the later use of the AI system will be set. For instance, the specific instructions included in an algorithm tells it how to complete a task and under which conditions. If these instructions are altered, the system will function in another way (Bucher, 2018). In the EU process on AI, the High-Level Group presents the idea of "ethics by design," namely that norms could be incorporated in algorithms and architectures of AI systems (AI HLEG, 2019a, 21) to make the behavior of AI systems ethical (Dignum, 2019, 6). Ethical principles are complied with to different degrees in different populations, and it is tricky to translate ethical principles to precise systems and algorithms. That is particularly difficult when autonomous decision-making algorithms meet moral dilemmas based on independent and possibly conflicting value systems (Etzioni and Etzioni, 2016). Despite these difficulties, many AI researchers believe that it is possible to develop ethical frameworks that could guide AI system reasoning and decision-making (Dignum, 2019, ch. 5). While it is beyond the scope of this article to judge this possibility, from a responsibility perspective it is important to point out some principal implications of deliberately incorporating ethical values in AI technology.

One issue with ethics by design is to decide on the principles that should be built into the technology. Recalling that agents in the position to make a difference should be

[6]In parallel to this process on regulation of AI, other policy processes with bearing on AI are also taking place within the EU. By the introduction of the Digital Markets Act and the Digital Services Act, the EU is upgrading current rules on digital services with the purpose to balance the dominance on digital platforms of big companies such as Google, Facebook and Amazon (EC, 2021c; EP, 2022). While the products and services of these companies to a large extent build on AI technology, the AI as such is not the focus of this regulation and hence not included in this study. Automated vehicles is another product depending on AI technology, and a number of expert reports on automated vehicles have been produced within/for the EU. For instance, the report *Ethics of Connected and Automated Vehicles* (EC, 2020b) is explicitly discussing forward-looking responsibility in a way that relates to the analysis of this study. However, so far, automated vehicles are not specifically included in the policy process on AI referred to above and in focus of this paper. Thus, neither that report is included in this study.

TABLE 1 | Documents used in the analysis.

Document	Organization	Date	References
Draft report with recommendations to the Commission on Civil Law Rules on Robotics	European Parliament	2016, May 31	EP, 2016a
European Civil Law Rules in Robotics	European Parliament	2016, October	EP, 2016b
The future of robotics and artificial intelligence in Europe	European Commission	2017, February 16	EC, 2017
Civil Law Rules on Robotics	European Parliament	2017, February 16	EP, 2017
The European AI Landscape	European Commission	2018, January	EC, 2018b
Statement on Artificial Intelligence, Robotics and "Autonomous" Systems	European Group on Ethics in Science and New Technologies	2018, March	EGE, 2018
Draft Ethics Guidelines for Trustworthy AI	Artificial Intelligence High Level Expert Group	2018, December 18	AI HLEG, 2018
Building Trust in Human-Centric Artificial Intelligence	European Commission	2019, April 4	EC, 2019
Ethics Guidelines for Trustworthy AI	Artificial Intelligence High Level Expert Group	2019, April 8	AI HLEG, 2019a
A Definition of AI: Main Capabilities and Disciplines	Artificial Intelligence High Level Expert Group	2019, April 8	AI HLEG, 2019b
White Paper on Artificial Intelligence	European Commission	2020, February 19	EC, 2020c
Setting up a new committee on artificial intelligence	European Parliament	2020, June 20	EP, 2020
Public Consultation on the AI White Paper	European Commission	2020, November	EC, 2020a
Proposal for a Regulation on Artificial Intelligence	European Commission	2021, April 21	EC, 2021d
Fostering a European Approach to Artificial Intelligence	European Commission	2021, April 21	EC, 2021a
Artificial Intelligence: MEPs discuss ways to boost EU competitiveness	European Parliament	2021, April 23	EP, 2021

TABLE 2 | Bases for distribution of forward-looking responsibility.

Principle	Rationale	Who is responsible? Examples
Contribution	Causality in the past	Actors are responsible to remedy a bad situation to which they have contributed
Benefit	Beneficial in the future	Actors are responsible for negative effects from which they will benefit
Position	Best placed to fulfill obligation	Actors with power, knowledge and other relevant resources are responsible to remedy a bad situation or prevent future harm

assigned forward-looking responsibility, a technical solution would certainly empower developers and designers, who have the expert knowledge required to be attributed forward-looking responsibility. This may however increase the influence of developers and designers at the expense of the general public. From a democratic perspective, it is crucial that the public is included in deliberations about values (Dahl, 1989; Chambers, 2003; Urbinati, 2014), and that is not least important regarding values supposed to guide a technology that is integrated in practically all aspects of people's lives (Chambers and Gastil, 2021; Gastil and Broghammer, 2021; Buhmann and Fieseler, 2022). Although the prospect of reaching consensus on value questions is always difficult, on national as well as on local level, the global reach of AI technology makes such an endeavor even harder. Moreover, consensus may not even be a goal to strive for, as pluralism of values may reflect our moral reality (Hedlund, 2014; Søbirk Petersen, 2021). On the other hand, it could be argued that technology is never neutral, but always incorporates judgments (Jasanoff, 2016), and that designers and engineers always make decisions about the technology they develop that have consequences for users (Verbeek, 2006). From such a perspective, ethics by design might be a way to make this explicit. However, attempts to assign forward-looking responsibility *via* ethics by design run into challenges that make operationalisation difficult (Schiff et al., 2021; Georgieva et al., 2022).

Furthermore, even if ethics could be perfectly designed into AI systems, it would hit only narrow technical aspects of AI, and disregard the societal situatedness of the technological systems in question (Hagendorff, 2021). AI systems could still be used in ways that cause environmental damages, foster social oppression, and support unethical business models (Van Dijck et al., 2018; Buruk et al., 2020; Crawford, 2021). Certainly, the Commission points to the importance of AI systems that do not cause harm to society and the environment (EC, 2021d, p. 24), but the proposed measures to avoid risks are primarily directed at the technical systems as such. To a large extent, these measures are phrased in terms of responsibility. As is discussed above, to be responsible in a forward-looking way you need to be in the position to make a difference, and it is quite obvious that AI developers are in such a position when it comes to the design of AI applications.

This is also the line of argument when the Commission discusses the distribution of obligations for high-risk AI systems[7]. Each obligation, the Commission states, "should be addressed to the agent(s) who is (are) best placed to address any potential risk," and illustrates with developers of AI, who are best placed to do so in the development phase (EC, 2020c, p. 22).

The explicit reference to the agents "best placed" is very much in line with the theoretical requirements for distribution of forward-looking responsibility discussed above. However, this does not preclude that there are backward-looking aspects here as well, such as requirements of post-market monitoring (EC, 2021d, p. 74–75) or requirements to pay attention to the whole AI lifecycle (EC, 2021d, p. 3, 4, 30, 46–47, 49). Nevertheless, although those measures refer to retrospective control, they are expressed as an obligation that some agent has in the future. As an example, the Commission states that it is important that providers "take the responsibility" to establish a robust post-marketing system to identify, analyse, estimate, evaluate and manage risks with high-risk AI systems that they place on the market (EC, 2021d, p. 31, 47), and for the requirement to ensure safety and protecting fundamental rights "throughout the whole AI system's lifecycle" (EC, 2021d, p. 3), providers as well as designers and developers are attributed responsibility in a forward-looking sense.

As these examples illustrate, forward-looking responsibility may come with constraints, but it may also come with benefits. Obviously, for providers, the requirements of post-market monitoring of risks are extensive, and would probably be costly in the short run. On the other hand, post-market auditing could play a positive role in marketing and public relations (Mökander and Floridi, 2021). For designers and developers, responsibility to ensure traceability of data and processes involved in the production of an AI model would most probably be a burden. In the absence of a common approach, and as there seems to currently be a lack of appropriate tools (Mora-Cantallops et al., 2021), taking the step from principles to practice can be difficult (Schneiderman, 2020). Moreover, even though there is an awareness of AI ethics and checklists at national levels, aligning them with the EU-level guidance can be challenging (Georgieva et al., 2022).

However, ethics by design may also open up possibilities for designers. Being attributed forward-looking responsibility can be empowering, by implying opportunity for influence and control (Ulbert, 2017; Persson and Hedlund, 2021), and as will be discussed in the section on competition, the EU is bringing out their ethical requirements on AI systems as a competitive advantage. By being assigned responsibility for the design of ethical AI, experts in the field not only have the task to design AI technology ethically. By merit of their expertise, they already are in a position to do so (Douglas, 2009), but with this responsibility, it can be argued that they also are empowered to influence the development of AI technology. One consequence of this is that it is important to pay attention to who these designers are. For instance, the gender imbalance in the tech industry with an overwhelming majority of men[8], and the observation that women have a tendency to focus on different things than men, may increase the risk that certain perspectives will be lost (Hagendorff, 2020; Resseguier and Rodrigues, 2021). Another aspect of attributing responsibility to designers is where they work. In the context of AI development, a few big tech companies dominate the global market (Van Dijck et al., 2018), and the risk of empowering the already powerful may possibly speak against justification through position. These aspects of designer responsibility—who designers are and where they work—further emphasize the importance of a public discussion about which values and principles to build into AI systems. It may also point to a need to foster agents to take on responsibility for the good of society.

RESPONSIBILITY AS PREVENTING HARM: ARTIFICIAL GENERAL INTELLIGENCE

AI systems today surpass humans on specialized tasks such as for example different games or facial recognition (narrow AI). Although restricted to one particular task, narrow AI can have the potential to lead to practically irreversible change in a specific domain of society. For instance, widespread use of AI-driven surveillance seems to have a potential to irreversibly change the capacity of authoritarian governments to suppress dissent (Bakir, 2015; Brucato, 2015; Greutzemacher and Whittlestone, 2022). Other AI technologies, although narrow in the sense that they have one particular task, may have "the potential to lead to practically irreversible change that is broad enough to impact most important aspects of life and society" (Greutzemacher and Whittlestone, 2022, p. 6). One example of such transformative AI (TAI) could be natural language understanding, which has gone through dramatic advances in the last years and could potentially lead the way to practical human–machine interaction via language user interfaces. Combined with powerful decision-making machines, this could potentially result in systems that can replace the majority of current jobs (Greutzemacher and Whittlestone, 2022). A further step in the ladder of risk is a more general artificial intelligence, with human-level intelligence or beyond. General Artificial Intelligence (AGI) has for a long time been the object of imaginative interpretations in film and literature. So far, this has mostly been speculative, but lately, we have seen academic discussions on how to make sure a potential AGI would not "take over" and constitute a threat to human life as we know it. A "take over" scenario could refer to a sudden transition in which AI very soon gets much

[7]The Commission proposes four levels of risk: unacceptable risk (e.g., social scoring by governments), high risk (e.g., remote biometric identification of people), limited risk (e.g., manipulation by chat bot), and minimal risk (no need of legislation) (EC, 2021b). See also the section on AGI below.

[8]According to *AI Index Annual Reports*, men make up at average 80% of AI professors, AI PhD recipients has remained virtually constant at 20% women since 2010 in the US (Perrault et al., 2019), and men constitute 83.9% and women 16.1% of tenure track faculty at computer science departments at top universities around the world 2019–2020 (Zhang et al., 2021). In a recent global survey about the gender distribution among software developers in 2021, 91.7% of developers are men (Statista, 2022). Gender imbalance is just one example of the biased composition of AI developers, that also include geography and ethnicity (Zhang et al., 2021).

more intelligent and develops superintelligence (Bostrom, 2014; Tegmark, 2017; Russell, 2019). In such a future there is no self-evident room for humans, who, in the best-case scenario, will function as the machines' slaves or pets, but, more probably, will be wiped out in the endless striving for resources of this superintelligence. While the AI community is divided regarding the potential of AGI (Sotala and Yampolskiy, 2015), another objection to such a scenario is that it does not necessarily take a superintelligence to outmaneuver humans. As the notion of TAI bears witness to, superiority in some area may be sufficient and general intelligence is not a necessary way to AGI or radically transformative AI (c.f., Russell, 2019; Häggström, 2021; Greutzemacher and Whittlestone, 2022). The difficulty is to make sure that the superintelligence would realize goals that are in the interest of humans (Bostrom, 2014, p. 127; Dafoe, 2018). Discussions on how to solve this so-called control problem abound, but many of the suggestions have proved to be unsafe or would not work at all (Russell, 2019, p. 145–170).

However, the possibility of superintelligence that outsmart humans is not a question that is addressed in the European Union policy process on AI. In fact, early in the EU policy process on AI the Commission explicitly rejects issues of a "speculative" nature, but stresses the need to "address concrete problems we are facing today" rather than dealing with "unrealistic expectations from the technology" (EC, 2017). Nevertheless, the draft ethical guidelines from the High-Level Group, published in the following year, point to potential concerns with a possible development of Artificial Moral Agents or self-improving AGI in a long-distant future. Such technology, the group states, "would potentially present a conflict with maintaining responsibility and accountability in the hands of humans" (AI HLEG, 2018, p. 13). While not explicitly alluding to a "AI taking over" scenario, the draft guidelines do show a forward-looking perspective on responsibility when raising this concern. Interestingly, this concern is toned down in the final version of the ethical guidelines. Now it is referred to as something that can be "hypothesized," that "some" consider that AGI or superintelligence[9] "can be examples" of long-term concerns, but that "many others" believe them to be unrealistic (AI HLEG, 2019a, p. 35). However, both the draft and the final version of these guidelines emphasize that the concerns should be kept "into consideration" (AI HLEG, 2018, p. 13; AI HLEG, 2019a, p. 35). In contrast to this, in a workshop on AI organized by the Commission, several EU member states put forward "a need to discuss realistic future scenarios, and found debates surrounding dystopian applications of AI, e.g., killer robots, to be hindering beneficial AI development" (EC, 2018b, p. 29)[10].

In the Commission's *White paper on Artificial Intelligence* (EC, 2020c), published the year after the final guidelines, the concept of high-risk AI is introduced. High-risk AI refers to AI technology with "significant risks" in relation to safety, consumer rights and fundamental human rights, and "exceptional instances" such as the use of AI systems for recruitment, remote biometric identification (e.g., video surveillance with face recognition) and "other intrusive surveillance technologies" (EC, 2020c, p. 17–18). This understanding of high-risk AI is kept in the proposal for regulation of AI, which is tailored to concrete situations of concern that can "reasonably be anticipated in the near future" (EC, 2021d, p. 3). As this proposal constitutes the basis for a possible legislation on AI, it is arguably the most important document in this policy process so far. However, although hypothetical and "speculative" developments of AGI and superintelligence are not discussed, the proposal includes a longer-term perspective, emphasizing that the regulation should be flexible and enable it to be "dynamically adapted as the technology evolves and new concerning situations emerge" (EC, 2021d, p. 3). This future orientation is further emphasized by the notion of a "future-proof" definition of AI, "taking into account the fast technological and market developments related to AI" (EC, 2021d, p. 4).

Although "future-proof" may give connotations to a mark that is stamped on a product for marketing purposes, it indicates responsibility in a forward-looking sense. However, this makes the gradual dismissal in this process of the risks of potential AGI somewhat remarkable. Admittedly, as with all technology development (c.f., Carrier, 2021), there is a great uncertainty of the emergence of a future AGI. However, the characteristics of AI that distinguishes it from other technologies—the capacity to "learn" and to act "autonomously," pointed out above—make it even more complicated to anticipate the possible emergence of AGI. Moreover, the problem of value alignment, that is, the difficulty of ensuring that an AI system's goals and behavior should be aligned with human values (Bostrom, 2014; Sierra et al., 2021), needs to be solved before a potential AGI is a reality, if a disastrous development is to be avoided. If not, there is a risk that AGI will really become human's "final invention" (Barrat, 2013). Even if the probability of AGI were diminutive, provided that the stakes are so high that Bostrom and others put forward, it would not have been unexpected to see it included in considerations of a "future-proof" AI. In terms of forward-looking responsibility as preventing future harm, the potential emergence of AGI seems to be an issue of concern.

How, then, could potential AGI be approached with regard to forward-looking responsibility? Without claiming to provide a definite answer, some tentative thoughts may give a hint to the complexity of this matter. The capacity principle in combination with the position principle may appear to be best fitting to distribute obligations to see to it that AGI, if it materializes, will be aligned with human values, but also the contribution principle may apply and, perhaps, the benefit principle. Computer scientists and other AI developers are arguably the agents that have the knowledge and skills that can make a difference, they have contributed to the technology that has given rise to the speculation and worry in the first place, and they would, perhaps, also gain from future AGI, should they succeed. On the other hand, the large uncertainty connected to potential AGI and how, if at all, it could be controlled by humans, may speak against the

[9]Other examples listed are Artificial Consciousness and Artificial Moral Agents (AI HLEG, 2019a, p. 35).

[10]Although this is not the place to make assumptions of why people have a specific opinion, this attitude may be a reflection of the focus on growth and competition, which is discussed below.

allocation of responsibility on the basis of technological expertise only. While the technological experts would need to do the hands-on work, should it be done at all, it could be argued that it is the responsibility of policy-makers to regulate questions of existential threat. Referring to the precautionary principle (e.g., EP, 2015), political representatives could decide to regulate development of AGI, for instance, by prohibiting it altogether, or by allocating resources for a development of safe AGI. However, the compliance of a prohibition would be hard to control, and even if resources would be destined to development of safe AGI, agents who are not covered by these resources may go ahead without any concern for safety. It seems important in this context that regulation is combined with the nurturing of virtues of responsibility, making agents willing to take responsibility for the common good[11].

While forward-looking responsibility certainly gets attention in this policy process, it is clear that the focus is more short-term than long-term (Prunkl and Whittlestone, 2020)[12]. In a short-term perspective, emphasis is on issues that society is already facing or is likely to face very soon, such as data privacy and algorithmic bias. These issues are concrete and specific, and directly related to progress in machine learning which have enabled real-world applications. In contrast, in a long-term perspective, focus is on issues emerging from advanced AI systems that will arise far into the future, or are less certain, such as AGI or radical transformative AI (Prunkl and Whittlestone, 2020, p. 2). Admittedly, it is easier to pay attention to the more obvious and probabilistically likely scenarios that short-term issues give rise to, than to concentrate on uncertain and long-term concerns, and maybe especially so in the political sphere; the late awakening regarding climate change may be a case that illustrates this point. In this particular policy process on AI development, it is not unlikely that political considerations have played a role when questions on AGI came to be left out. For one thing, as pointed out above, governmental representatives explicitly stated that the discussion should be about scenarios that are realistic and not be hampered by hypothetical applications, an attitude that is clearly short-term. It is well-established that political representatives tend to favor issues that are within reach for their influence, that is, within their terms of office (Beckman, 2008; Randers, 2015). Moreover, so far in this policy process, actors from the AI industry have been well-represented, and it is not improbable that they prefer to focus on technologies and applications that they are currently working on, and, as we saw in the ethics by design section, issues that can be met by technical means. In other words, power dynamics in the policy process may explain why forward-looking responsibility is not a matter of concern when it comes to potential risks that possibly lie very far in the future.

[11]This is the approach of the expert report on automated vehicles, mentioned above, which states that "legislation alone may be insufficient" and that the fostering of "a culture of responsibility" is crucial in the development and design of connected and automated vehicles (EC, 2020b, p. 16, 53, 56–57).

[12]Prunkl and Whittlestone (2020) use the terms "near-term" and "long-term".

COMPETITION—CONSTRAINING AND ENABLING FACTOR FOR RESPONSIBILITY

As indicated above, the primacy of innovation and competition is a norm that appears to be taken as a given in this policy process. Although the efforts to develop AI in an ethical and responsible way get slightly different focus in the course of this policy process, one thing remains constant: safety and ethics always need to have the highest priority, but—not to the extent that they hamper development and competition. Formulations that ethics and safety must not restrain, but should facilitate innovation, or that measures to ensure safety at the same time must encourage innovation and growth, are, in different variants, frequently found in the documents both in the early and in the recent phases of the process. For instance, the Commission states that "any measure proposed or considered in this context should not stifle innovation" (EC, 2017), and in its most recent document, that proposes a regulation for AI, the Commission emphasizes the importance of ensuring "a legal framework that is innovation-friendly" (EC, 2021d, p. 34). Ensure ethics—but not at any price, seems to be the message. Strongly related to this emphasis on innovation is the urge for competition and leadership.

A common line of argument in the course of this policy process is the emphasis on the importance of investing in research and development to become or to remain the world leader. In the early stage of the process, discussions about AI are mostly found in the Parliament. While European leadership on AI is not the core concern for the Parliament at that time, it is nevertheless mentioned, although in cautious and somewhat down-to-earth terms. Europe "could," it is said, "become the global leader" on AI (EP, 2016b, p. 25). When the Commission takes the initiative on questions of AI, leadership takes a more prominent role. In a reply to the Parliament, the Commission stresses that it has since long recognized the need for significant investment in AI, which is necessary for the EU "to stay in the lead" and "maintain a strong leadership role" (EC, 2017). Later, in its White Paper on AI, the Commission asserts that the EU is "well-positioned to exercise global leadership" in promoting ethical use of AI (EC, 2020c, p. 8). Now also the European Parliament takes action for competition, and sets up a special committee on Artificial Intelligence in a Digital Age to analyse the impact of AI on the EU economy (EP, 2020, 2021). According to the Stanford University AI index report, the US and China clearly lead the AI race (Stanford, 2019), and the chair of AIDA underlines the hurry to act: "If we wake up soon enough, we can make up this gap" (IIIM, 2020).

While it is repeatedly stated that ethical considerations must not impede competition, ethics instead becomes a strategy for competitive leadership. For the Commission, the EU is particularly well-positioned to "create world-leading AI capabilities" (EC, 2021a, p. 3) and "exercise global leadership" in promoting ethical use of AI (EC, 2020c, p. 2, 8). One basis for the Commission's confidence in "ethical AI" is the extensive work to produce ethical guidelines for AI that began with the publication of *Statement on Artificial Intelligence, Robotics, and "Autonomous" Systems* from the European Group on Ethics in Science and New Technologies (EGE, 2018). The EGE worries

about "ethics shopping" and underlines a "great urgency" of a coordinated, balanced approach in the regulation of AI in which "Europe should play an active and prominent role" (EGE, 2018, p. 12, 14). This statement also informs the ethical guidelines of the High-Level Expert Group on Artificial Intelligence (AI HLEG), which suggests "responsible competitiveness" as a unique approach to AI that enables Europe "to position itself as a global leader in cutting-edge AI" (AI HLEG, 2019a, p. 5)[13].

From the perspective of responsibility, the focus on competition may give rise to value conflicts, referring to the impossibility of complying with one value without violating another (Søbirk Petersen, 2021). One such example is that larger datasets to work against bias may increase privacy violation (Jobin et al., 2019, p. 396). Another is that competition as a value may conflict with safety and other ethical concerns. In terms of forward-looking responsibility, it is important not only to find out what various stakeholders want from AI, but also to find ways to deal with conflicting values (Dafoe, 2018, p. 49). The risk of potential value conflicts is recognized by the Commission, mentioning that several stakeholders warn the Commission to avoid "conflicting obligations" (EC, 2021d, p. 8). For instance, the obligation of business actors to identify and manage risks with AI could be resource-demanding and constrain their competitiveness, as well as an expectation to contribute to EU's competitiveness may limit their capacity to ensure ethical AI. For developers, the responsibility to make AI ethical by integrating values in the design of AI systems may likewise constrain their competitiveness, maybe especially so for "micro or small enterprise" (EC, 2021d, p. 40). We can see that costs of fulfilling responsibilities on, say, safety, may affect the agent's competitive capacity. However, while the Commission does not explicitly address how this potential value conflict should be met in terms of forward-looking responsibility, its reference to "responsible innovation" (EC, 2021d, p. 11, 34) could be a way to reconcile the possible friction and turn it into an asset in line with the "responsible competitiveness" approach discussed in the previous paragraph.

On the other hand, as pointed out by Van de Poel and Sand (2018), "responsible innovation" does not only suggest that innovators carry additional responsibility, but also that this responsibility (e.g., backward-looking responsibility for product failure) could diminish the willingness of engineers and innovators to learn about such effects to prevent future failure, thus, discourage them to take forward-looking responsibility. Instead, van de Poel and Sand proposes some kind of insurance for the risks with innovation. For the individual innovator, this could reduce the cost and increase the willingness to take on responsibility to prevent risks and, thereby, facilitate their contribution to the competitive race. Hence, the norm of innovation and competition could be both a constraining and an enabling factor for forward-looking responsibility, although the enabling component appears to need some promotion, illustrated by the joining of "responsibility" with "competition" and "innovation" in the EU documents.

However, the strategy to make ethics a comparative advantage and this coupling of responsibility with competition imbues the priority of keeping up in the technically advanced AI race. Although the goal is said to be AI that is "trustworthy", indicating a technology that is reliable, focus is on how this trustworthiness can ensure a functioning single market (EC, 2021d, p. 9) and support the objective of global leadership (EC, 2021d, p. 18). While the focus on ethics, human rights, and responsibility in this competition is welcome from a humanly and societal perspective, the point of departure is how AI can be advantageous, not the other way round. In other words, the question of what society we want and whether AI could be helpful in this regard, is not asked. So interpreted, ethics and responsibility can be seen as a means to a higher end, that is, competition.

CONCLUSION

I have argued that responsibility in a forward-looking sense is particularly relevant in the context of handling risks with AI. Overall, the analysis of the EU policy process on AI shows that competition is the primary value, and that the perspective is technical and focused on short-term concerns. In the case of ethics by design, operationalisation may be challenging both for technical and for normative reasons. The question of which values should be built into the technology and how this should be settled remained an issue after the distribution of responsibility to designers and other technical experts. Moreover, the technical focus and the biased composition of designers may have the effect that certain issues are disregarded, pointing as well to the importance of inclusive deliberation about values, as to a need to foster agents to a sense of responsibility for the good of society. In addition, while specific obligations allocated to designers such as ensuring traceability could constrain their work, obligations of a more general character could increase the influence of designers on the development of AI.

Further, the discussion on the potential threat from a superintelligent AI, or AGI, showed that forward-looking responsibility in terms of ensuring good is not only theoretically fitting to this particular problem, but also a moral obligation that decision-makers should take on. Arguments that a potential AGI is unlikely, is very far in the future, or that some AI experts do not believe it to ever be a reality are not entirely convincing in that regard, and even if the probability of AGI is extremely low, the stakes are too high to dismiss the question. However, although AGI or similar was mentioned by the High-Level Expert Group, it never really was an issue in this policy process, and it was gradually phased out, primarily by the influence of actors from the political sphere. Tentative reflections on how potential AGI could be approached with regard to forward-looking responsibility, had it been included in this policy

[13]This orientation toward competition is particularly interesting as it echoes of the previous discussion within the EU in the lengthy process on biological patents in the 1980's and 1990's, in which ethics was included to put the public at ease first after strong opposition against one-sided focus on competition and growth (Busby et al., 2008; Mohr et al., 2012). Now, in the process on AI, the Commission speaks about using ethics as a means for competition and promoting European values worldwide (EC, 2021a, p. 8), which is a notable contrast to the, as it seemed, reluctant adaption of ethical values in the process of biological patents some decades ago.

process, showed difficulties with reliance on regulation only, and suggested that it is combined with nurturing of virtues of responsibility to make agents willing to take responsibility for the common good.

Moreover, I have demonstrated how competition within the EU process on AI is a norm that frames how responsibility is approached, and gives rise to potential value conflicts. Competition can be seen as a primary goal, making responsibility a means to this end. The strategy to frame ethics as a comparative advantage, strengthens this interpretation. From this point of view, forward-looking obligations can induce costs that constrain competitiveness of business actors and developers, at least in the short run. On the other hand, the obligation to contribute to the EU's competitiveness can constrain their capacity to realize responsible AI.

Finally, some reflections on the concept of forward-looking responsibility as it was conceptualized in this study. The distinction between remedying harm and preventing harm was tentatively tried out, but needs to be further developed in a theory of forward-looking responsibility. Although well-established as principles, the particular combination of principles for distribution of forward-looking responsibility applied in this study was adapted to the context of AI regulation in the EU. The principle of contribution, with its connection to causality, resembles how backward-looking responsibility is attributed: you are normally, under certain circumstances, accountable for things to which you can be causally connected. In the forward-looking variant, you are responsible to remedy past harms that you may have caused. Following this, the principle of contribution is "logical" from a moral point of view. The principle of benefit, on the other hand, is more normative, as it is based on the moral standpoint that those who gain from something are responsible, whether responsibility is seen as a benefit or as a burden. Still, there is some connection to the issue at hand. For position, no such connection is necessary. This makes position the potentially most controversial basis for distribution of responsibility. Nevertheless, if we want forward-looking responsibility to make a difference to real-world situations, position is arguably the principle that best meets the requirement of efficiency, although in some cases, costs could justify a relief of the efficiency requirement. However, in the context of AI development, position, like the other principles discussed here, often point out technical expertise as the responsible agents. As the analysis illustrates, this might be a problem for the appointed agents if the responsibility constrains them, but it could also be a benefit, as responsibility also is empowering. For society, however, the biased composition of these technical experts in AI implies a risk that some perspectives will be excluded, perhaps especially so in ethics by design, which potentially leaves room for designers to influence which values to build into AI. It is argued that public deliberation on the values that should guide design could broaden the view and reduce the risk that certain perspectives are omitted, and that responsibility as virtue could be a viable approach to deal with the risk that individual interests get prominence over the societal good.

A related matter that was only touched upon here is the question of inclusion in the policy process. In the EU policy process on AI, there is a dominance of those who have contributed, those who will benefit, and those who are in a position to take on forward-looking responsibility. How this has affected the distribution of forward-looking responsibility is a topic of future studies.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data are publicly available at various websites. All relevant details are included in the reference list.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

This study was supported by Marianne and Marcus Wallenberg Foundation MMW 2018.0020.

REFERENCES

Abul-Fottouh, D., Song, M. Y., and Gruzd, A. (2020). Examining algorithmic biases in YouTube's recommendations of vaccine videos. *Int. J. Med. Informat.* 140:104175. doi: 10.1016/j.ijmedinf.2020.104175

AI HLEG (2018). *Draft Ethics Guidelines for Trustworthy AI*. Brussels: European Commission: High-Level Expert Group on AI.

AI HLEG (2019a). *Ethics Guidelines for Trustworthy AI*. Brussels: European Commission: High-Level Expert Group on AI.

AI HLEG (2019b). *A Definition of AI: Main Capabilities and Disciplines*. Brussels: European Commission: High-Level Expert Group on AI.

Alcaniz, I., and Hellwig, T. (2011). Who's to blame? The distribution of responsibility in developing democracies. *Br. J. Polit. Sci.* 41, 389–411. doi: 10.1017/S0007123409990317

Alfano, M. (2021). Towards a genealogy of forward-looking responsibility. *Monist* 104, 498–509. doi: 10.1093/monist/onab015

Alhamid, M. F., Rawashdeh, M., Osman, H. A., Hossain, M. S., and El Saddik, A. (2015). Towards context-sensitive collaborative media recommender system. *Multimedia Tools Appl.* 74, 11399–11428. doi: 10.1007/s11042-014-2236-3

Anton, A. L. (2015). *Moral Responsibility and Desert of Praise and Blame*. Lamham, MD: Lexington Books.

Asan, O., and Choudhury, A. (2021). Research trends in artificial intelligence applications in human factors health care: mapping review. *JMIR Hum. Fact.* 8:28236. doi: 10.2196/28236

Bakir, V. (2015). 'Veillant panoptic assemblage': mutual watching and resistance to mass surveillance after Snowden. *Media Commun.* 3, 12–25. doi: 10.17645/mac.v3i3.277

Barrat, J. (2013). *Our Final Invention: Artificial Intelligence and the End of the Human Era. Dunne/St. Martin's*.

Bartlett, J. (2018). How AI could kill democracy. London: New Statesman, 28–32.

Beckman, L. (2008). Do global climate change and the interest of future generations have implications for democracy? *Environ. Polit.* 17, 610–624. doi: 10.1080/09644010802193500

Bexell, M., and Jönsson, K. (2021). *The Politics of the Sustainable Development Goals: Legitimacy, Responsibility, and Accountability*. London, New York, NY: Routledge. doi: 10.4324/9781003043614

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qualitat. Res. J.* 9, 27–40. doi: 10.3316/QRJ0902027

Brucato, B. (2015). The new transparency: police violence in the context of ubiquitous surveillance. *Media Commun.* 3, 39–55. doi: 10.17645/mac.v3i3.292

Bucher, T. (2018). If… then: algorithmic power and politics. *Oxford Scholarship Online*. doi: 10.1093/oso/9780190493028.001.0001

Buhmann, A., and Fieseler, C. (2022). Deep learning meets deep democracy: deliberative governance and responsible innovation in Artificial Intelligence. *Bus. Ethics Quart.* 42, 1–34. doi: 10.1017/beq.2021.42

Buruk, B., Ekmekci, P. E., and Arda, B. (2020). A critical perspective on guidelines for responsible and trustworthy artificial intelligence. *Med. Health Care Philos.* 23, 387–399. doi: 10.1007/s11019-020-09948-1

Busby, H., Hervey, T., and Mohr, A. (2008). Ethical EU law? The influence of the European Group on Ethics in Science and New Technologies. *Eur. Law Rev.* 33, 803–842.

Carragee, K. M., and Roefs, W. (2004). The neglect of power in recent framing research. *J. Commun.* 54, 214–233. doi: 10.1111/j.1460-2466.2004.tb02625.x

Carrier, M. (2021). How to conceive of science for the benefit of society: prospects of responsible research and innovation. *Synthese* 198 (Suppl.19), 4749–4768. doi: 10.1007/s11229-019-02254-1

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., and Floridi, L. (2018). Artificial intelligence and the 'good society': the US, EU, and UK approach. *Sci. Eng. Ethics* 24, 505–528. doi: 10.1007/s11948-017-9901-7

Chambers, S. (2003). Deliberative democratic theory. *Ann. Rev. Polit. Sci.* 6, 307–326. doi: 10.1146/annurev.polisci.6.121901.085538

Chambers, S., and Gastil, J. (2021). Deliberation, democracy, and the digital landscape. *Polit. Stud.* 69, 3–6. doi: 10.1177/0032321719901123

Chasek, P. S., Wagner, L. M., Leone, F., Lebada, A.-M., and Risse, N. (2016). Getting to 2030: negotiating the post-2015 sustainable development agenda. *Rev. Eur. Commun. Int. Environ. Law* 25, 5–14. doi: 10.1111/reel.12149

Crawford, K. (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven, CT: Yale University Press. doi: 10.12987/9780300252392

Dafoe, A. (2018). *AI Governance: A Research Agenda*. Oxford: Centre for the Governance of AI, Future of Humanity Institute, University of Oxford.

Dahl, R. A. (1989). *Democracy and Its Critics*. New Haven, CT: Yale University Press.

Daly, D. J. (2021). Virtue ethics and action guidance. *Theol. Stud.* 82, 565–582. doi: 10.1177/00405639211055177

Danaher, J. (2016). Human enhancement, social solidarity and the distribution of responsibility. *Ethical Theor. Moral Pract.* 19, 359–378. doi: 10.1007/s10677-015-9624-2

Deibert, R. J. (2019). The road to digital unfreedom: three painful truths about social media. *J. Democr.* 30, 25–39. doi: 10.1353/jod.2019.0002

Diamond, L. (2019). The threat of postmodern totalitarianism. *J. Democr.* 30, 20–24. doi: 10.1353/jod.2019.0001

Dignum, V. (2019). *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Berlin: Springer. doi: 10.1007/978-3-030-30371-6

Doorn, N., Raven, R. P. J. M., and Royakkers, L. M. M. (2011). Distribution of responsibility in socio-technical networks: the Promest case. *Technol. Anal. Strategic Manag.* 23, 453–417. doi: 10.1080/09537325.2011.558403

Douglas, H. E. (2009). *Science, Policy, and the Value-free Ideal*. Pittsburg, PA: University of Pittsburgh Press. doi: 10.2307/j.ctt6wrc78

Droz, L. (2021). Distribution of responsibility for climate change within the milieu. *Philosophies* 6, 1–15. doi: 10.3390/philosophies6030062

EC (2017). *The Future of Robotics and Artificial Intelligence in Europe*. European Commission. Blog post. Available online at: https://ec.europa.eu/digital-single-market/en/blog/future-robotics-and-artificial-intelligence-europe (accessed March 25, 2022).

EC (2018a). *Artificial Intelligence for Europe. Communication COM. 237 Final*. Available online at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0237andfrom=EN (accessed March 25, 2022).

EC (2018b). *The European AI Landscape. Workshop Report*. Brussels: EC.

EC (2019). *Building Trust in Human-Centric Artificial Intelligence. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions*. Brussels: European Commission.

EC (2020a). *Public Consultation on AI White Paper – Final report*. Brussels: European Commission.

EC (2020b). *Ethics of Connected and Automated Vehicles: Recommendations on Road Safety, Privacy, Fairness, Explainability and Responsibility. Horizon 2020 Commission Expert Group*. Brussels: European Commission.

EC (2020c). *White Paper on Artificial Intelligence – A European Approach to Excellence and Trust*. Brussels: European Commission.

EC (2021a). *Fostering a European Approach to Artificial Intelligence. Communication*. Brussels: European Commission.

EC (2021b). *New rules for Artificial Intelligence". Questions and Answers*. Available online at: https://ec.europa.eu/commission/presscorner/detail/en/QANDA_21_1683 (accessed March 25, 2022).

EC (2021c). *The Digital Services Act package*. Available online at: https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package (accessed March 25, 2022).

EC (2021d). *Proposal for a Regulation on Artificial Intelligence – A European Approach*. Brussels: European Commission.

EC (2022). *A European Approach to Artificial Intelligence*. Available online at: https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence (accessed March 25, 2022).

EC Futurium. *AI HLEG – steering group of the European Alliance*. European Commission. Available online at: https://ec.europa.eu/futurium/en/european-ai-alliance/ai-hleg-steering-group-european-ai-alliance.html (accessed March 25, 2022).

EC Regexpert. *European Commission Register of Commission Expert Groups and Other Similar Entities. High-Level Expert Group on Artificial Intelligence (E03591)*. Available online at: https://ec.europa.eu/transparency/expert-groups-register/screen/expert-groups/consult?do=groupDetail.groupDetailandgroupID=3%20591 (accessed March 25, 2022).

EGE (2018). *Statement on Artificial Intelligence, Robotics, and 'Autonomous' Systems*. Brussels: The European Group of Ethics in Science and New Technologies.

EP (2015). *The Precautionary Principle: Definitions, Applications and Governance*. European Parliament. Available online at: https://www.europarl.europa.eu/thinktank/en/document/EPRS_IDA(2015)573876

EP (2016a). *Committee on Legal Affairs. Draft report with recommendations to the Commission of civil Law Rules on Robotics. 2015/2103(INL)*. Strasbourg: European Parliament.

EP (2016b). *Legal Affairs, European Civil Law Rules in Robotics*. PE 571.379. Strasbourg: European Parliament.

EP (2017). *Civil Law Rules on Robotics Resolution. P8_TA (2017)0051*. Strasbourg: European Parliament.

EP (2020). *Setting up a special committee on artificial intelligence in a digital age, and defining its responsibilities, numerical strength and term of office. (2020/2684(RSO))*. Strasbourg: European Parliament.

EP (2021). *Artificial Intelligence: MEPs discuss ways to boost EU competitiveness". Press release, March 23, 2021*. Strasbourg: European Parliament. Available online at: https://www.europarl.europa.eu/news/en/press-room/20210322IPR00511/artificial-intelligence-meps-discuss-ways-to-boost-eu-competitiveness (accessed March 25, 2022).

EP (2022). *EU Digital Markets Act and Digital Services Act Explained*. Strasbourg: European Parliament.

Erskine, T. (2014). Coalitions of the willing and responsibilities to protect: informal associations, enhanced capacities, and shared moral burdens. *Ethics Int. Affairs* 28, 115–145. doi: 10.1017/S0892679414000094

Etzioni, A., and Etzioni, O. (2016). AI assisted ethics. *Ethics Inform. Technol.* 18, 149–156. doi: 10.1007/s10676-016-9400-6

European Council. *The Ordinary Legislative Procedure*. Available online at: https://www.consilium.europa.eu/en/council-eu/decision-making/ordinary-legislative-procedure/ (accessed March 25, 2022).

Fukuda-Parr, S., and McNeill, D. (2015). Post 2015: a new era of accountability? *J. Glob. Ethics* 11, 10–17. doi: 10.1080/17449626.2015.1004738

Gastil, J., and Broghammer, M. (2021). Linking theories of motivation, game mechanics, and public deliberation to design an online system for participatory budgeting. *Polit. Stud.* 69, 7–25. doi: 10.1177/0032321719890815

Georgieva, I., Lazo, C., Timan, T., and van Veenstra. A. F. (2022). From AI ethics to data science practice: a reflection and a gap analysis based on recent frameworks and practical experience. *AI Ethics*. 127:3. doi: 10.1007/s43681-021-00127-3

Gilbert, M. (1993). Agreements, coercion, and obligation. *Ethics* 103, 679–706. doi: 10.1086/293548

Gorgoni, G. (2020). Stay human: the quest for responsibility in the algorithmic society. *J. Ethics Legal Technol.* 2, 31–47. doi: 10.14658/pupj-jelt-2020-1-2

Graafland, J. J. (2003). Distribution of responsibility, ability and competition. *J. Bus. Ethics* 45, 133–147. doi: 10.1023/A:1024188916195

Greutzemacher, R., and Whittlestone, J. (2022). The transformative potential of Artificial Intelligence. *Futures* 135, 1–11. doi: 10.1016/j.futures.2021.102884

Gunnemyr, M. (2020). Why the social connection model fails: participation is neither necessary nor sufficient for political responsibility. *Hypatia* 35, 567–586. doi: 10.1017/hyp.2020.40

Hagendorff, T. (2020). The ethics of AI ethics: an evaluation of guidelines. *Minds Machines* 30, 99–120. doi: 10.1007/s11023-020-09517-8

Hagendorff, T. (2021). Blind spots in AI ethics. *AI Ethics*. 122:8. doi: 10.1007/s43681-021-00122-8

Häggström, O. (2021). *Tänkande maskiner: Den artificiella intelligensens genombrott [Thinking machines: The breakthrough of artificial intelligence]*. Stockholm: Fri Tanke.

Haider, J., and Sundin, O. (2020). *Invisible Search and Online Search Engines: The Ubiquity of Search in Everyday Life*. London: Routledge. doi: 10.4324/9780429448546

Hakli, R., and Mäkelä, P. (2019). Moral responsibility of robots and hybrid agents. *Monist* 102, 259–275. doi: 10.1093/monist/onz009

Haselager, W. F. G. (2005). Robotics, philosophy and the problems of autonomy. *Pragmat. Cogn.* 13, 515–532. doi: 10.1075/pc.13.3.07has

Hayward, T. (2012). Climate change and ethics. *Nat. Climate Change* 2, 843–848. doi: 10.1038/nclimate1615

Hedlund, M. (2012). Epigenetic responsibility. *Med. Stud.* 3, 171–183. doi: 10.1007/s12376-011-0072-6

Hedlund, M. (2014). Ethics expertise in political regulation of biomedicine: the need of democratic justification. *Crit. Pol. Stud.* 8, 282–299. doi: 10.1080/19460171.2014.901174

Held, V. (1970). "Can a random collection of individuals be morally responsible?" in *Five Decades of Debate in Theoretical and Applied Ethics*, eds L. May and S. Hoffman (Lanham, MD: Rowman & Littlefield Publishers, Inc.). p. 89–100.

Helm, P. (2018). Treating sensitive topics online: A privacy dilemma. *Ethics. Inform. Technol.* 20, 303–313. doi: 10.1007/s10676-018-9482-4

Horizon (2020). *Responsible Research and Innovation*. Available online at: https://ec.europa.eu/programmes/horizon2020/en/h2020-section/responsible-research-innovation (accessed March 25, 2022).

IIIM (2020). *Icelandic Institute for Intelligent Machines*. Available online at: https://www.iiim.is/2020/10/europe-aims-to-catch-up-in-ai-race/#more-5931 (accessed March 25, 2022).

Jain, L. C., and Prathiar, D. K. (2010). *Intelligent Autonomous Systems*. Berlin/Heidelberg: Springer.

Jakob, M., Ward, H., and Steckel, J. C. (2021). Sharing responsibility for trade-related emissions based on economic benefits. *Glob. Environ. Change* 66, 1–8. doi: 10.1016/j.gloenvcha.2020.102207

Jasanoff, S. (2016). *The Ethics of Invention: Technology and the Human Future*. New York, NY; London: W. W. Norton and Company.

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

Karageorgiou, E. (2019). The distribution of asylum responsibilities in the EU: Dublin, partnerships with third countries, and the question of solidarity. *Nordic J. Int. Law* 2019, 315–358. doi: 10.1163/15718107-08803003

Kim, B., I., Koopmanschap, M. H. R., Mehrizi, M., and Huysman, and, E., Ranschaert (2021). How does the radiology community discuss the benefits and limitations of artificial intelligence for their work? A systematic discourse analysis. *Eur. J. Radiol.* 2021:136. doi: 10.1016/j.ejrad.2021.109566

Kingdon, J. W. (1995). *Agendas, Alternatives, and Public Policies*. New York, NY: Harper Collins.

Lake, C. (2001). *Equality and Responsibility*. Oxford; New York, NY: Oxford University Press.

Lindblom, C. E., and Woodhouse, E. J. (1993). *The Policy-Making Process*. Upper Saddle River, NJ: Prentice Hall.

Ma, Y., Wang, Z., Yang, H., and Yang, L. (2020). Artificial intelligence applications in the development of autonomous vehicles: a survey. *IEEE/CAA J. Automat. Sin.* 7, 315–329. doi: 10.1109/JAS.2020.1003021

Mäkäle, P. (2007). Collective agents and moral responsibility. *J. Soc. Philos.* 38, 456–468. doi: 10.1111/j.1467-9833.2007.00391.x

Mapel, D. R. (2005). Political obligation, and benefits across borders. *Polity* 37, 426–442. doi: 10.1057/palgrave.polity.2300022

Markandya, A. (2011). Equity and distributional implications of climate change. *World Dev.* 39, 1051–1060. doi: 10.1016/j.worlddev.2010.01.005

Mathias, G. (2021). The ambivalence of the psychosocial in Norwegian education: a policy document analysis. *Nordic J. Stud. Educ. Pol.* 7, 65–77. doi: 10.1080/20020317.2021.1958994

Melchiorre, A. B., Rekabsaz, N., Parada-Cabaleiro, E., Brandl, S., Lesota, O., and Schedl, M., Investigating gender fairness of recommendation algorithms in the music domain. *Inform. Proces. Manag.* (2021) 58, 1–27. doi: 10.1016/j.ipm.2021.102666

Messina, J. P. (2021). Reasonable pluralism about desert-presupposing moral responsibility: a conditional defense. *J. Value Inq.* 55, 189–208. doi: 10.1007/s10790-020-09746-1

Metzinger, T. (2019). *EU Guidelines: Ethics Washing Made in Europe*. Washington, DC: Der Tagespiegel.

Miller, D. (2001). Distributing responsibilities. *J. Polit. Philos.* 9, 453–471. doi: 10.1111/1467-9760.00136

Moelle, M. P. (2017). *The International Responsibility of International Organisations: Cooperation in Peacekeeping Operations*. Cambridge: Cambridge University Press. doi: 10.1017/9781316415757

Moellendorf, D. (2009). Treaty norms and climate change mitigation. *Ethics Int. Affairs* 23, 247–266. doi: 10.1111/j.1747-7093.2009.00216.x

Mohr, A., Busby, H., Hervey, T., and Dingwall, R. (2012). Mapping the role of official bioethics advice in the governance of biotechnologies in the EU: the European Group on Ethics' Opinion on commercial cord blood banking. *Sci. Public Pol.* 39, 105–117. doi: 10.1093/scipol/scs003

Mökander, J., and Floridi, L. (2021). Ethics-based auditing to develop trustworthy AI. *Minds Machines* 21:8. doi: 10.1007/s11023-021-09557-8

Mora-Cantallops, M., Sánchez-Alonso, S., García-Barriocanal, E., and Sicilia, M.-A. (2021). Traceability for trustworthy AI: a review of models and tools. *Big Data Cogn. Comput.* 5, 20–34. doi: 10.3390/bdcc5020020

Neuhäuser, C. (2014). Structural injustice and the distribution of forward-looking responsibility. *Stud. Philos.* 38, 232–251. doi: 10.1111/misp.12026

Nihlén Fahlquist, J. (2009). Moral responsibility for environmental problems: individual or institutional? *J. Agri. Environ. Ethics* 22, 109–124. doi: 10.1007/s10806-008-9134-5

Nihlén Fahlquist, J. (2015). *Responsibility as a virtue and the problem of many hands" in Ibo Van de Poel, Lambèr Royakkers and Sjoerd D. Zwart, Moral Responsibility and the Problem of Many Hands*. London; New York, NY: Routledge, 187–208.

Nihlén Fahlquist, J. (2019). Public health and the virtues of responsibility, compassion and humility. *Public Health Ethics* 12, 213–224. doi: 10.1093/phe/phz007

O'Neill, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York, NY: Broadway Books.

Page, E. A. (2008). Distributing the burdens of climate change. *Environ. Polit.* 17, 556–575. doi: 10.1080/09644010802193419

Pereboom, D. (2021). Undivided forward-looking moral responsibility. *Monist* 104, 484–497. doi: 10.1093/monist/onab014

Perrault, R., Shoham, Y., Brynjolfsson, E., Clark, J., Etchemendy, J., Grosz, B., et al. (2019). *The AI Index 2019 Annual Report, AI Index Steering Committee*. Stanford, CA: Human-Centered AI Institute, Stanford University.

Persson, E., Eriksson, K., and Knaggård, Å. (2021). A fair distribution of responsibility for climate adaptation: translating principles of distribution from an international to a local context. *Philosophies* 6, 1–16. doi: 10.3390/philosophies6030068

Persson, E., and Hedlund, M. (2021). The future of AI in our hands? To what extent are we as individuals morally responsible for guiding the development

of AI in a desirable direction? *AI Ethics*. 21:5. doi: 10.1007/s43681-021-00125-5

Piorkowski, D., Park, S., Wang, A. Y., Wang, D., Muller, M., and Portnoy, F. (2021). How AI developers overcome communication challenges in a multidisciplinary team: a case study. *arXiv*. 2021:3449205. doi: 10.1145/3449205

Prunkl, C., and Whittlestone, J. (2020). Beyond near- and long-term: Toward a clearer account of research priorities in AI ethics and and society. *arXiv [Preprint]. arXiv: 2001.04335v2*. Available online at: https://arxiv.org/pdf/2001.04335.pdf

Randers, J. (2015). *The Tyranny of the Short-Term: Why Democracy Struggles With Issues Like Climate Change*. Democracy Audit. Available online at: (accessed March 25, 2022). http://www.democraticaudit.com/2015/02/09/the-tyranny-of-the-short-term-why-democracy-struggles-with-issues-like-climate-change/

Resseguier, A., and Rodrigues, R. (2021). Ethics as attention to context: recommendations for the ethics of artificial intelligence. *Open Res. Europe* 13260:2. doi: 10.12688/openreseurope.13260.2

Rochefort, D.A., and Cobb, R. W. (1994). *The Politics of Problem Definition: Shaping the Policy Agenda*. Lawrence, Kansas: University Press of Kansas.

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. New York, NY: Viking.

Russell, S., and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach, 3rd Edn*. London: Pearson Education.

Schiff, D., Fanti, A., Rakova, B., Lennon, M., and Ayesh, A. (2021). Explaining the principles to Practices gap in AI. *IEEE Technol. Soc. Magazine* 40, 81–94. doi: 10.1109/MTS.2021.3056286

Schmidt, J. A. (2021). "Virtuous engineers: ethical dimensions of technical decisions," in *Science, Technology, Virtues*, eds E. Ratti and T. A. Stapleford (Oxford: Oxford Scholarship Online), 1.

Schneiderman, B. (2020). Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Trans. Interact. Intellig. Syst.* 10, 1–31. doi: 10.1145/3419764

Sierra, C., Osman, N., Noriega, P., Sabater-Mir, J., and Perello-Moragues, A. (2021). *Value Alignment: A Formal Approach*. Available online at: https://arxiv.org/pdf/2110.09240.pdf (accessed March 25, 2022).

Sirbu, A., Pedreschi, D., Giannotti, F., and Kertész, J. (2019). Algorithmic bias amplifies opinion fragmentation and polarization: a bounded confidence model. *PLoS ONE* 14, 1–20. doi: 10.1371/journal.pone.0213246

Sneddon, A. (2005). Moral responsibility: the difference of Strawson, and the difference it should make. *Ethical Theor. Moral Pract.* 8, 239–264. doi: 10.1007/s10677-005-2484-4

Søbirk Petersen, T. (2021). Ethical guidelines for the use of artificial intelligence and value conflict challenges. *Nordic J. Appl. Ethics* 15, 25–40. doi: 10.5324/eip.v15i1.3756

Sondermann, E., Ulbert, C., and Finkenbusch, P. (2017). "Introduction: moral agency and the politics of responsibility," in *Moral Agency and the Politics of Responsibility*, eds C. Ulbert, P. Finkenbusch, E. Sondermann, and T. Debiel (London: Routledge), 1–18. doi: 10.4324/9781315201399-1

Sotala, K., and Yampolskiy, R. V. (2015). Responses to catastrophic AGI risk: a survey. *Physica Scripta* 90, 1–33. doi: 10.1088/0031-8949/90/1/018001

Stahl, T. (2016). Indiscriminate mass surveillance and the public sphere. *Ethics Inform. Technol.* 18, 33–39. doi: 10.1007/s10676-016-9392-2

Stahl, T. (2020). Privacy in public: a democratic defence. *Moral Philos. Polit.* 7, 73–96. doi: 10.1515/mopp-2019-0031

Stanford (2019). Stanford University AI Index Report. Available online at: https://hai.stanford.edu/sites/default/files/ai_index_2019_report.pdf

Statista (2022). *Software Developer Gender Distribution Worldwide as of 2021*. Available online at: https://www.statista.com/statistics/1126823/worldwide-developer-gender/ (accessed March 25, 2022).

Tegmark, M. (2017). *Life 3.0: Being Human in the Age of Artificial Intelligence*. New York, NY: Alfred A. Knopf.

Thompson, D. F. (1987). *Political Ethics and Public Office*. Cambridge, MA; London: Harvard University Press.

Ulbert, C. (2017). "In search of equity: practices of differentiation and the evolution of a geography of responsibility," in *Moral Agency and the Politics of Responsibility*, eds C. Ulbert, P. Finkenbusch, E. Sondermann, and T. Debiel (London: Routledge), 105–121. doi: 10.4324/9781315201399-7

Ulbert, C., and Sondermann, E. (2017). "Conclusion: practising the politics of responsibility," in *Moral Agency and the Politics of Responsibility*, eds C. Ulbert, P. Finkenbusch, E. Sondermann, and T. Debiel (London: Routledge), 196–202. doi: 10.4324/9781315201399-13

Urbinati, N. (2014). *Democracy Disfigured*. Cambridge, MA: Harvard University Press. doi: 10.4159/harvard.9780674726383

Van de Poel, I. (2015a). "Moral responsibility," in *Moral Responsibility and the Problem of Many Hands*, eds I. Van de Poel, L. Royakkers and S.D. Zwart (London; New York, NY: Routledge), 13–49. doi: 10.4324/9781315734217

Van de Poel, I. (2015b). "The problem of many hands," in *Moral Responsibility and the Problem of Many Hands*, eds I. Van de Poel, L. Royakkers, and S. D. Zwart (London; New York, NY: Routledge), 50–92.

Van de Poel, I., Royakkers, L., and Zwart, S. D. (2015). *Moral Responsibility and the Problem of Many Hands*. London; New York, NY: Routledge.

Van de Poel, I., and Sand, M. (2018). Varieties of responsibility: two problems of responsible innovation. *Synthese* 198, 4769–4787. doi: 10.1007/s11229-018-01951-7

Van Dijck, J., Poell, T., and de Waal, M. (2018). *The Platform Society: Public Values in a Connective World*. Oxford: Oxford University Press. doi: 10.1093/oso/9780190889760.001.0001

Verbeek, P.-P. (2006). Materializing morality: design ethics and technological mediation. *Sci. Technol. Hum. Values* 31, 361–380. doi: 10.1177/0162243905285847

Weiss, J. A. (1989). The powers of problem definition: the case of government paperwork. *Pol. Sci.* 22, 97–121. doi: 10.1007/BF00141381

Williams, G. (2008). Responsibility as virtue. *Ethical Theor. Moral Pract.* 11, 455–470. doi: 10.1007/s10677-008-9109-7

Young, I. M. (2006). Responsibility and global justice: a social connection model. *Soc. Philos. Pol.* 23, 102–130. doi: 10.1017/S0265052506060043

Young, I. M. (2011). *Responsibility for Justice*. Oxford: Oxford Scholarship Online. doi: 10.1093/acprof:oso/9780195392388.001.0001

Zhang, D., Mishra, S., Brynjolfsson, E., Etchemendy, J., Ganguli, D., Grosz, B., et al. (2021). *The AI Index 2021 Annual Report*. Stanford, CA: AI Index Steering Committee, Human-Centered AI Institute, Stanford University.

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for the Future at the New Frontier of Power*. London: Profile Books.

Zwart, S. D. (2015). "Responsibility and the problem of many hands in networks," in *Moral Responsibility and the Problem of Many Hands*, eds I. Van de Poel, L. Royakkers, and S. D. Zwart (London; New York, NY: Routledge), 132–166.

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.