Received: 28 June 2022 | Revised: 3 May 2023 | Accepted: 24 May 2023

DOI: 10.1111/ropr.12567

RPR REVIEW OF POLICY RESEARCH

ORIGINAL ARTICLE

Power and politics in framing bias in Artificial Intelligence policy

Inga Ulnicane | **Aini Aden**

De Montfort University, Leicester, UK

Correspondence
Inga Ulnicane, De Montfort University, The Gateway, Leicester LE1 9BH, UK.
Email: inga.ulnicane@dmu.ac.uk

Funding information
EU's Horizon 2020 Research and Innovation Programme, Grant/Award Number: 945539, 785907 and 720270

Abstract

Bias is a key issue in expert and public discussions about Artificial Intelligence (AI). While some hope that AI will help to eliminate human bias, others are concerned that AI will exacerbate it. To highlight political and power aspects of bias in AI, this contribution examines so far largely overlooked topic of framing of bias in AI policy. Among diverse approaches of diagnosing problems and suggesting prescriptions, we can distinguish two stylized framings of bias in AI policy—one more technical, another more social. Powerful technical framing suggests that AI can be a solution to human bias and can help to detect and eliminate it. It is challenged by an alternative social framing, which emphasizes the importance of social contexts, balance of power and structural inequalities. Technological frame sees simple technological fix as a way to deal with bias in AI. For the social frame, we suggest to approach bias in AI as a complex wicked problem, for which a broader strategy is needed involving diverse stakeholders and actions. The social framing of bias in AI considerably expands the legitimate understanding of bias and the scope of potential actions beyond technological fix. We argue that, in the context of AI policy, intersectional bias should not be perceived as a niche issue but rather be seen as a key to radically reimagine AI governance, power and politics in more participatory and inclusive ways.

KEYWORDS

artificial intelligence, bias, intersectionality, policy, politics, power, technological fix

> 'If diversity isn't present in the planners at the planning stage, then we get the same issues we see in the biased data-sets. AI doesn't have a skin colour or a gender – by making it mostly white and mostly male at every stage, we're reinforcing a problem we need to solve.
>
> If AI and AGI[1] really is going to benefit the many and not the few, people invited to the table must include more people of colour, more women, and more people with the humanities background – rather than an overwhelming number of male physicists'
>
> (Winterson, 2021: 235).

INTRODUCTION

Bias is one of the key topics in public, political and scholarly debates about Artificial Intelligence (AI) (Bender et al., 2021; Benjamin, 2019; Broussard, 2023; Caliskan et al., 2017; Courtland, 2018; Criado Perez, 2019; D'Ignazio & Klein, 2020; Noble, 2018; O'Neil, 2016; Zou & Schiebinger, 2018). Increased use of AI in many settings from hiring and credit allocation to crime prediction and courts, has raised major questions about fairness, transparency and accountability. Optimistic suggestions have been made that AI can help to identify, reduce or even eliminate human bias. However, several well-publicized cases suggest the opposite, namely, that AI reflects and even amplifies pre-existing gender, racial, ethnic and other biases exacerbating inequalities and discrimination. In 2016, a widely discussed ProPublica investigative journalism study revealed that software used in the US courts to predict future crimes overestimated the risk of black defendants to reoffend but underestimated the risk of white defendants to commit future crimes (Angwin et al., 2016). Research on commercial gender classification algorithms using facial analysis dataset found that classifiers performed best for lighter-skinned individuals and males overall but performed worst on darker-skinned females (Buolamwini & Gebru, 2018). A study of search engines demonstrated how they reinforce racial and gender stereotypes about black girls (Noble, 2018).

Above-mentioned examples have raised awareness and shaped discussion about bias in AI. They have highlighted that AI as any technology is not just a neutral tool but has major political and societal implications (Winner, 1980), including problematic consequences such as discrimination, increased inequalities and violation of human rights. As technology tends to reflect people who make it, examples of bias in algorithms and AI devices have drawn attention to the lack of gender, racial, ethnic and other types of diversity among AI developers and founders of the major big tech companies (Little & Winch, 2021; West et al., 2019; Young et al., 2021). Recently, a couple of memoirs (Liu, 2020; Wiener, 2020) have provided first-hand accounts of female experiences in the male dominated tech industry, where, for example, the use of male pseudonyms for external correspondence helps to be more effective[2] and gives more authority because men simply respond differently to men (Wiener, 2020).

In the context of raising concerns about bias in AI, many stakeholders, experts and policy-makers are calling for urgent action and present a range of recommendations to tackle bias (Collett & Dillon, 2019; European Commission, 2020a; Koene et al., 2018; UNESCO, 2019, 2020; West et al., 2019; Young et al., 2021). These emerging discussions invite a closer look at political, policy and power aspects of bias in AI. The question of politics of bias in AI has acquired new urgency in the context of illiberal backlash when objections to gender equality in AI policy are expressed at the highest level of the European Union's decision-making (Schopmans & Cupac, 2021).

While research on bias in AI has addressed many ethical, philosophical, social and technical aspects (see e.g., Broussard, 2023; Kordzadeh & Ghasemaghaei, 2022; Simon et al., 2020; Søraa, 2023), so far political, power and policy aspects have received less attention.[3] Against this background, this paper examines how bias is presented and described in AI policy documents. In the context of recent advances in AI, policy-makers and stakeholders around the world have formulated their approaches to AI as one of the key emerging technologies that comes with major economic and social promises but also major concerns about its impact on democracy, welfare state, and human rights (Radu, 2021; Schiff, 2023; Ulnicane, Eke, et al., 2021; Ulnicane, Knight, et al., 2021). Previous research (Jobin et al., 2019; Schiff et al., 2021; Ulnicane, Eke, et al., 2021; Ulnicane, Knight, et al., 2021) indicates that bias and related concepts such as fairness, equity, and inclusion are among major concerns discussed in a number of these documents.

To study how bias is understood in policy documents, we use the policy framing approach that focusses on frames as diagnostic and prescriptive stories about what is wrong and what needs doing (Rein & Schon, 1993, 1996; Schon & Rein, 1994). We examine a number of closely related questions about how bias is framed in AI policy: Which intersectional characteristics—gender, race, ethnicity and other—are included in framing of bias? How do policy documents frame the relationship between AI and bias: will AI eliminate or amplify human bias? And how the causes and impacts of bias in AI as well as the recommendations to tackle it are framed?

Our contribution is developed in the context of the special issue on Politics and Policy of AI. By analyzing framing of bias in AI policy, we contribute to other articles in this special issue which focus on ideational dimension of policy and emerging themes in AI policy documents around the world (af Malmborg, 2022; Kim, 2023; Schiff, 2023). In particular, our contribution is part of a discussion about how AI policy can disadvantage certain groups in society (Giest & Samuels, 2022) and neglects the role of existing power relations (Rönnblom et al., 2023).

The article proceeds as follows: First, it introduces the key concepts for this study—AI, bias and intersectionality; second, it outlines policy framing approach; third, it introduces our methods and data; fourth, it presents our empirical material on framing bias in AI policy; and finally, the two stylized frames of bias in AI—technical and social—are summarized and discussed in conclusions.

KEY CONCEPTS: AI, BIAS, AND INTERSECTIONALITY

Three concepts—AI, bias, and intersectionality—are particularly relevant for our analysis of how bias is framed in AI policy documents. While all three are complex and contested concepts, in this section, we introduce them in a way that is relevant for this study.

Artificial Intelligence

The field of AI has been described as including 'a broad set of approaches, with the goal of creating machines with intelligence' (Mitchell, 2019: 8). The definition of AI, as well as its current state and potential future development remains contested, with ongoing debates and speculations about the gap between the current narrow AI that can perform only narrowly defined tasks and envisaged future general AI that is expected to do everything humans do and possibly more (Mitchell, 2019: 40–41).

Similarly, in AI policy, discussions about the definition of AI are ongoing (European Commission, 2019a). In our analysis, we follow the actors' definitions, namely, how AI is defined and understood in AI policy documents, where it is often used as a broad umbrella term to cover machine learning, robotics, autonomous systems and other subfields (Ulnicane, Knight, et al., 2021). According to a widely-used policy definition

> Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals. AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications).
>
> (European Commission, 2018b: 1).

Our understanding of AI draws on social studies of technology, which approach technologies not just as neutral tools or instruments but as a result of mutual co-shaping between technologies and societies, politics and cultures they are embedded (Jasanoff, 2003; Winner, 1980). Accordingly, we study AI as a result and reinforcement of existing power relations and political, economic, social and cultural settings (Benjamin, 2019; Broussard, 2023; Noble, 2018). This has important implications for understanding bias in AI, which is not just an accidental by-product or technical error but rather a reflection of power structures and unequal social, political and economic systems in which it is developed and used. Langdon Winner summarizes inherent social biases in the development of technology as follows:

> many of the most important examples of technologies that have political consequences are those that transcend the simple categories of "intended" and "unintended" altogether. These are instances in which the very process of technical development is so thoroughly biased in a particular direction that it regularly produces results counted as wonderful breakthroughs by some social interests and crushing setbacks by others. In such cases it is neither correct nor insightful to say, "Someone intended to do somebody else harm." Rather, one must say that the technological deck has been stacked long in advance to favor certain social interests, and that some people were bound to receive a better hand than others.
>
> (Winner, 1980: 125–126).

New technologies including AI are seen as particularly harmful in reproducing existing biases because they are perceived as more objective (Benjamin, 2019; Noble, 2018). Meredith Broussard introduces the term 'technochauvinism' to describe the widespread belief among technology

developers that computational solutions are superior to all other solutions, including human ones (Broussard, 2023). In her book 'More than a Glitch', she challenges the popular perception that technology problems such as reproduction of biases are just glitches, pointing out that 'The biases embedded in technology are more than mere glitches; they're baked in from the beginning. They are structural biases, and they can't be addressed with a quick code update.' (Broussard, 2023: 4).

In popular debates about AI, technology is often offered as a solution to social problems, including bias. Such approach resonates with a long tradition of technological fix that presents technology as a solution to complex societal problems (Johnston, 2018). In his work on solutionism, Evgeny Morozov criticizes the urge to use technology to fix problems that do not exist (Morozov, 2013). He points out that 'in promising almost immediate and much cheaper results, they [technology fixes] can easily undermine support for more ambitious, more intellectually stimulating, but also more demanding reform projects' (Morozov, 2013: 9).

Ruha Benjamin warns that 'the road to inequity is paved with technical fixes' (Benjamin, 2019: 7) because 'tech fixes often hide, speed up, and even deepen discrimination, while appearing to be neutral and benevolent when compared to the racism of a previous era' (Benjamin, 2019: 8). As an alternative, Broussard suggests to think more holistically (Broussard, 2023: 28). We argue that one way to think more holistically is to approach bias as a 'wicked problem', namely a complex social problem which cannot be solved just by scientific and technological measures (Rittel & Webber, 1973) but requires a broader strategy involving diverse stakeholders and actions (Head, 2022; Ulnicane, 2022).

To highlight the political relevance and implications of recent discussions about AI and bias, it is important to place AI in the broader context of debates about emerging technologies (Ulnicane, 2022) characterized by radical novelty, relatively fast growth, prominent impact, and uncertainty and ambiguity (Rotolo et al., 2015). Such emerging technologies are also characterized by hype as well as positive and negative expectations that have a performative function, meaning that irrespective of their accuracy they have important implications for shaping agendas and actions (Ulnicane et al., 2022; Van Lente et al., 2013). Importantly, emerging technologies not only bring social and economic benefits but can also exacerbate social problems (Coad et al., 2021; Garvey, 2021). These characteristics of emerging technologies are important for contextualizing recent discussions about bias in AI.

Bias in AI

Bias in AI is multifaceted phenomenon that includes social, physical and cognitive aspects (Wellner & Rothman, 2020). It is understood as 'prejudice' (European Commission, 2019b: 36), 'a systematic deviation from equality that emerges in the outputs of an algorithm' (Kordzadeh & Ghasemaghaei, 2022: 395) and is seen as 'legally or morally unacceptable within the social context where the system is used, e.g. algorithmic systems that produce outcomes with differential impact strongly correlated with protected characteristics (such as race, gender, sexuality, etc)' (Koene et al., 2018: 39).

An important reference point in recent studies of bias in AI (Simon et al., 2020) is earlier work on bias in computer systems by Friedman and Nissenbaum (1996). They use the term bias with significant moral meaning

> to refer to computer systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favour of other. A system discriminates unfairly if it denies an opportunity or a good or if it assigns an undesirable

> outcome to an individual or group of individuals on grounds that are unreasonable or inappropriate
>
> (Friedman & Nissenbaum, 1996: 332).

They distinguish between three types of biases: pre-existing, technical, and emergent (Friedman & Nissenbaum, 1996). According to them, pre-existing bias originates from social institutions, practices, and attitudes, technical bias stems from technical constraints and considerations but emergent bias arises in a context of use.

Discussions about bias in AI often mention two understandings of bias—one more narrow technical understanding that focusses on data and statistics and another broader social (or socio-technical combining both technical and social aspects, see, e.g., Kordzadeh & Ghasemaghaei, 2022) understanding of bias that also considers historical and political contexts (Collett & Dillon, 2019; West et al., 2019). While some definitions and descriptions of bias in AI largely focus on algorithmic bias and related data bias, it is important to consider biased gender and racial representations in AI devices such as white robots and female voice assistants as well (Cave & Dihal, 2020; Collett & Dillon, 2019; Gruber & Benedikter, 2021).

Intersectionality

A crucial concept for understanding bias in AI is intersectionality. With its focus on multi-layered identity categories such as race, gender, and class, intersectionality provides an important lens for analyzing framing of bias in AI policy. In her pioneering work on intersectionality, Kimberle Crenshaw highlighted 'the need to account for multiple grounds of identity when considering how the social world is constructed' (Crenshaw, 1991: 1245). According to her, it is important to look at convergence of race, gender and class domination to understand how social power works to exclude or marginalize those who are different. Intersectionality allows to move beyond focus on gender inequality and examine how different identity categories interact in disadvantaging and discriminating certain groups (Fothergill et al., 2019; West et al., 2019). Intersectional character of bias in AI has been recognized by the European Commission (2020b), emphasizing that:

> As well as gender and sex, there are other, interconnected factors affecting bias, such as ethnicity, age, socioeconomic status, sexual orientation, geographic location and disability. These all shape a person's or a group's experience and social opportunities, thereby influencing the form of discrimination and inequality they encounter.
>
> (European Commission, 2020b: 2).

Recent literature on AI has focused on a range of biases including gender (D'Ignazio & Klein, 2020; Guevara-Gómez et al., 2021), racial (Benjamin, 2019; Cave & Dihal, 2020), ability (Broussard, 2023; Søraa, 2023; Whittaker et al., 2019), intersection of gender and racial biases (Broussard, 2023; Noble, 2018), and socio-economic biases which can also overlap with other types of biases leading to intersecting categories of race, class and gender (Eubanks, 2019). Little and Winch (2021) have described digital capitalism as a new patriarchy characterized by racialized, gendered and classed shaping of personhood that persists in its business culture, workforce and products.

Intersectional and feminist approaches to AI (Wellner & Rothman, 2020; West, 2020) invite to rethink and challenge discriminatory AI suggesting strategies that are polyvocal, multimodal and experimental (Ciston, 2019) and that promise a fairer, slower, consensual and collaborative AI (Toupin, 2023).

The three concepts of AI, bias and intersectionality introduced here are closely related to a number of other concepts such as fairness, justice, equality (Coeckelbergh, 2022; Wong, 2020) and diversity (Søraa, 2023), as we reveal in the following discussion of frames in AI policy documents.

POLICY FRAMING APPROACH

To analyze policy discussions about bias in AI, we have chosen policy framing approach developed by Martin Rein and Donald Schon (Rein & Schon, 1993, 1996; Schon & Rein, 1994). Policy frames are of great importance in social and political process of designing policy involving multiplicity of actors and shifting contexts (Schon & Rein, 1994). According to Rein and Schon, in policy frames 'facts, values, theories, and interests are intertwined' (Rein & Schon, 1993: 145) and

> framing is a way of selecting, organizing, interpreting, and making sense of a complex reality to provide guideposts for knowing, analysing, persuading, and acting. A frame is a perspective from which an amorphous, ill-defined, problematic situation can be made sense of and acted on
>
> (Rein & Schon, 1993: 146).

In Science and Technology Studies literature, frame analysis is seen as critically important for making sense of potential problems and solutions related to uncertainty of emerging technologies (Jasanoff, 2003).

For our analysis of how policy documents frame bias in AI and ways to deal with it, the conceptualization of policy frames as narratives of 'what is wrong and what needs doing' (Schon & Rein, 1994: 27) is of particular relevance. Rein and Schon define frames as 'diagnostic/prescriptive stories that tell, within a given issue terrain, what needs fixing and how it might be fixed' (Rein & Schon, 1996: 89). They link policy frames to public controversies and pluralism, where in each given issue domain a variety of policy frames compete for meaning, legitimacy, and economic and social resources.

Frames here are derived from policy documents which 'are treated as vehicles of messages, communicating or reflecting official intentions, objectives, commitments, proposals, "thinking", ideology and responses to external events' (Freeman & Maybin, 2011: 157). Analysis of policy documents focuses on rhetorical frames 'constructed from the policy-relevant texts that play important roles in policy discourse, where the context is one of debate, persuasion, and justification' (Rein & Schon, 1996: 90). Crucial for the frame analysis is 'to recognize the non-innocence of how "problems" get framed within policy proposals, how the frames will affect what can be thought about and how this affects possibilities for action' (Bacchi, 2000: 50). Analysis of framing also includes revealing of silences, omissions and politics hidden in the framing (Bacchi, 2000). To summarize, frames in policy-relevant texts are important because they shape legitimization of certain problems, actions and allocation of resources.

EXAMINING BIAS IN AI POLICY

As indicated above, policy documents reflect intentions, objectives and proposals, which affect possibilities for action. Therefore, we analyze how AI policy documents frame bias to identify narratives about the problem and potential solutions that can shape future policy actions.

Our empirical analysis of AI policy documents proceeds in two steps. First, we use an existing dataset of AI policy documents to identify those documents that explicitly discuss bias; and second, we undertake an in-depth analysis of how these documents frame bias. First, we draw on empirical material from previous analysis of 49 AI policy documents issued by national governments, international organizations, think tanks and consultancies in Europe and the US (see Annex 1)[4] (Ulnicane, Knight, et al., 2021). On the basis of this material, we identify documents that explicitly discuss bias. Of the initial 49 documents, a few do not mention bias at all, several only mention it in passing, but some mention issues related to bias such the need to improve fairness (Executive Office of the President, 2016c), equal access and equal opportunities (European Group on Ethics in Science and New Technologies 2018) as well as to encourage diversity and gender balance (European Commission, 2018c; HM Government, 2018) but do not necessarily elaborate on these issues and their relationship to bias.

Second, we focused on those AI policy documents that explicitly mention and meaningfully discuss bias. Both authors closely read these documents, focusing on how they frame bias. This included developing relevant categories for analysis, writing memos about the context of and approach to framing bias in each document as well as analyzing and discussing how these documents frame bias.

Bias in AI policy documents is introduced and discussed in various ways. Some policy documents have longer or shorter sections explicitly dedicated to bias (Campolo et al., 2017; Whittaker et al., 2018; CNIL, 2017; IEEE, 2017), others discuss bias in sections on issues such as social inequality (Crawford & Whittaker, 2016) or diversity in supply of skills (Hall & Pesenti, 2017), but some mention bias in the context of different themes discussed throughout the document (European Commission, 2018a). The annual reports prepared by the AI Now Institute (Campolo et al., 2017; Crawford & Whittaker, 2016; Whittaker et al., 2018) stand out among the analyzed documents with more elaborate discussion of bias.

The studied documents have been developed in various ways: some are written by the government-invited experts who have consulted other experts and organizations (Hall & Pesenti, 2017), others draw on inputs from various workshops and events (CNIL, 2017; Crawford & Whittaker, 2016) and some are result of volunteer contributions (IEEE, 2017). It is possible to observe cross-fertilization and mutual influences among the documents, as they reference other documents. For example, several documents (CNIL, 2017; European Commission, 2018a; IEEE, 2017) refer to discussions of bias in the influential AI Now Reports, written by some of the main internal critics of AI industry (Sadowski & Phan, 2022).

It was an important methodological choice to analyze how bias is framed in AI documents that discuss a broad range of issues related to AI rather than documents which are specifically dedicated to gender and equality in AI (Collett & Dillon, 2019; European Commission, 2020b; UNESCO, 2019; UNESCO, 2020; West et al., 2019; Young et al., 2021). By looking how AI is framed in broader AI policy documents, we can examine bias in the context of wider discussions about opportunities and challenges associated with AI rather than a siloed issue discussed in some dedicated documents.

FRAMING OF BIAS IN AI POLICY: DIAGNOSIS AND PRESCRIPTIONS

To study the framing of bias in AI policy, we examine how policy documents diagnose problems and what prescriptions they make to address them. Following the initial review of discussions of bias in AI policy documents, we formulate the following questions to guide our analysis:

- How do AI policy documents frame bias? Which intersectional characteristics—gender, race, ethnicity and others—are included in framing of bias?
- How do policy documents frame expectations towards AI and bias: will AI eliminate or amplify human bias?
- How do policy documents frame
  - the causes of bias in AI,
  - the consequences of bias in AI, and
  - the recommendations to tackle bias in AI?

In the following, we present findings of our analysis organized according to these questions.

Framing bias and its intersectional characteristics

While the word 'bias' is used in a number of AI policy documents, its meaning is far from clear and straightforward. Several documents talk about bias without explicitly defining and explaining it (the 2015 Panel 2016; BIC/APPGAI, 2017b; European Commission, 2018a; Hall & Pesenti, 2017). 'Bias' often appears next to other terms such as 'discrimination and exclusion' (CNIL, 2017), 'prejudice' (House of Lords, 2018) and 'diversity' (Hall & Pesenti, 2017). The AI Now 2017 report acknowledges that the word "bias" has multiple meanings that occasionally contradict each other (Campolo et al., 2017: 13). This document distinguishes the meaning of word bias in statistics used in many machine learning applications from the popular and social scientific definitions of bias. The former is described as follows

> the idea of "selection bias" refers to errors in estimation that result when some members of a population are more likely to be sampled than others. So when a machine learning program trained to recognize, say, faces of a particular racial group is applied to larger or more diverse populations, it may produce biased results in the sense of having a lower measure of accuracy.
>
> (Campolo et al., 2017: 14)

Similar to the selection bias is the notion of sampling bias mentioned in the report from the Royal Society (2017), which is understood as 'Selection of data or samples in a way that does not represent the true parameters (or distribution) of the population' (The Royal Society, 2017: 122).

In addition to the statistical understanding of bias, the AI Now 2017 report mentions more popular and social understanding of bias that refers to judgment based on preconceived notions or prejudices, as opposed to the impartial evaluation of facts: 'This sense of the word bias is closely linked to normative and ethical perspectives on fairness, and the idea that different groups

should be treated equally' (Campolo et al., 2017: 14). It is also admitted that both—statistical and normative—understandings are related:

> When examining technical systems, there can be temptation to, or vested interest in, limiting discussion of bias to the first more 'neutral' statistical sense of the term. However, in practice there is rarely a clear demarcation between the statistical and the normative definitions: biased models or learning algorithms, as defined statistically, can lead to unequal and unfair treatments and outcomes for different social or racial groups.
>
> (Campolo et al., 2017: 14).

When it comes to intersectional characteristics of bias, documents often discuss bias generally without specifying if it is gender, racial, ethnic or any other type of bias. On some occasions, specific intersectional characteristics are mentioned when talking about diversity, fairness or discrimination. Several documents mention categories of gender, race, sexual orientation and age (e.g., European Commission, 2018a; The 2015 Panel, 2016; UNI Global Union, 2017). The AI Now 2017 report mentions 'persistent gendered, racial and cultural biases' (Campolo et al., 2017: 18) when talking about recent analysis of search results and advertisements. The IEEE (Institute of Electrical and Electronics Engineers) document mentions 'biases that were inadvertently built into systems, such as racism and sexism in search engine algorithms' (IEEE, 2017: 45). The CNIL[5] report discusses racist bias of predictive justice and a face recognition software regarding African-Americans and Asians (CNIL, 2017: 31–32), gender bias in relation to women being less likely to be displayed advertisements for highly paid jobs (CNIL, 2017: 32) and social, racial and gender bias in the recruitment of AI developers (CNIL, 2017: 34).

To summarize, AI policy documents highlight multiple, overlapping and contradictory meanings of bias. These include statistical and technical meanings such as sampling or selection bias as well as social, ethical, normative and popular understandings of bias related to prejudices. It is emphasized that different meanings of bias are interrelated. When AI policy documents directly or indirectly mention various intersectional characteristics, they typically list a range of them such as gender, race, ethnicity, socio-economic background and sexual orientation. While various intersectional characteristics are listed, there is hardly any explicit discussion about their relationship and mutual reinforcement.

Relationship between AI and bias

When it comes to the relationship between bias and AI, documents discuss both options, namely that AI can help to detect, reduce and eliminate human bias as well as it can also reflect, embed and amplify it (Campolo et al., 2017; Hall & Pesenti, 2017). Documents highlight the potential of AI systems, if they are well-designed, to be less biased and fairer than humans, for example, when screening job applications (IEEE, 2017: 158). Moreover, it is suggested that AI offers opportunities to support diversity and help ensure equitable treatment

> AIs can be developed that can detect biases, both in new AI-supported functions, but also in existing, historical systems that still influence decision-making in different

> sectors. AI can address the challenges faced by individuals because of unconscious bias, by bringing these to the surface more effectively than has been done in the past. (Hall & Pesenti, 2017: 56).

However, several documents also highlight that AI can exacerbate human biases, when training data, algorithms, and other design choices that shape AI systems reflect and amplify existing cultural assumptions and inequalities (Campolo et al., 2017).

Examples of bias mentioned in AI policy documents such as the vast majority of humanoid robots having white skin and using female voices (IEEE, 2017: 51) and other cases mentioned in the previous and following sections demonstrate how AI reflects and amplifies existing biases embedded in historical data, cultural assumptions and power relations. However, concrete examples of how AI has helped to eliminate bias are difficult to find in AI policy documents, which might partly be due to the difficulty to detect such instances. To better understand discussion of how and why AI exacerbates bias, we turn to the next question on how policy documents frame reasons for bias in AI.

Reasons for bias in AI

Policy documents outline a number of reasons for bias in AI. These include pre-existing bias in society, lack of diversity in AI workforce, technical problems, and insufficient government regulation and transparency.

Several AI policy documents point out that bias is not a new problem, it has already existed in society for centuries, and AI resurfaces this problem (BIC/APPGAI, 2017b). Data and design of algorithms can reflect these long-standing biases in society (House of Lords, 2018):

> When data reflects biases of either form, there is the risk that AI systems trained on this data will produce models that replicate and magnify those biases. In such cases, AI systems would exacerbate the discriminatory dynamics that create social inequality, and would likely do so in ways that would be less obvious than human prejudice and implicit bias. (Crawford & Whittaker, 2016: 6–7).

As mentioned earlier, an important reason for bias is the lack of diversity of AI workforce. AI policy documents highlight that 'currently, the workforce is not representative of the wider population' (Hall & Pesenti, 2017: 56). AI developers are characterized as being mostly 'male, generally highly paid, and similarly technically educated' (Campolo et al., 2017: 17), with largely homogeneous racial and ethnic backgrounds (Crawford & Whittaker, 2016). It is reminded that 'in the past, gender and ethnic exclusion have been shown to affect the equitability of results from the technology' (Hall & Pesenti, 2017: 56). Homogeneity of composition of AI field 'can limit the perspectives and experiences of AI's creators' (Crawford & Whittaker, 2016: 5) and have negative effects on developing AI in an inclusive and representative way because

> AI applications and the data they rely upon may reflect the biases of their designers and users, who specify the data sources. This threatens to deepen existing social biases, and concentrate AI's benefits unequally among different subgroups of society. (The 2015 Panel, 2016: 43).

Several AI policy documents remind of historical change in the representativeness in this field from its early days when the computer industry had a significant proportion of female workers (House of Lords, 2018), which shrank as the field grow in its prominence (Hicks, 2018). The AI Now 2017 report summarizes this change over time as follows:

> Early programming and data entry work was characterized as secretarial, and was female-dominated. These women were themselves called "computers," and they were often undercompensated and rarely credited. All the while, they were responsible for things like maintaining sophisticated systems that targeted bomb strikes in World War II and tabulating decades of census data. The history of AI reflects this pattern of gender exclusion. The 1956 Dartmouth Summer Research Project on Artificial Intelligence, which initiated the concept of artificial intelligence, was exclusively attended by men. Pioneering work in natural language processing and computational linguistics, key to contemporary AI systems, has been credited to male colleagues and students rather than to Margaret Masterman, who founded the Cambridge Language Research Unit and was one of the leaders in the field. Intentional exclusion and unintentional "like-me" bias is responsible for a continued lack of demographic representation within the AI field and within the tech industry for women, Hispanics, and African Americans.
>
> (Campolo et al., 2017: 17).

Diversity of the workforce is seen as a crucial issue for AI development. A UK document states that 'if UK AI cannot improve the diversity of its workforce, the capability and credibility of the sector will be undermined' (Hall & Pesenti, 2017: 56).

Bias in AI can also occur due to more technical reasons such as using datasets poorly representative of the wider population for training AI (House of Lords, 2018) or using AI systems that 'are untested and poorly designed for their tasks' (Whittaker et al., 2018: 8) and 'not robust against malicious attacks' (European Commission, 2018a: 92). Moreover, the IEEE document highlights that secrecy of AI development can contribute to biases:

> Software engineers should employ "black-box" (opaque) software services or components only with extraordinary caution and ethical care, as they tend to produce results that cannot be fully inspected, validated, or justified by ordinary means, and thus increase the risk of undetected or unforeseen errors, biases, and harms.
>
> (IEEE, 2017: 71).

This is closely related to the lack of proper safety measures, auditing, oversight, transparency and regulation. Moreover, the AI Now 2018 report highlights that when AI systems 'make errors and bad decisions, the ability to question, contest, and remedy these is often difficult or impossible' (Whittaker et al., 2018: 8). Different reasons for bias discussed above can interact and reinforce each other exacerbating problems:

> The technology scandals of 2018 have shown that the gap between those who develop and profit from AI—and those most likely to suffer the consequences of its negative effects—is growing larger, not smaller. There are several reasons for this, including a lack of government regulation, a highly concentrated AI sector,

> insufficient governance structures within technology companies, power asymmetries between companies and the people they serve, and stark cultural divide between the engineering cohort responsible for technical research, and the vastly diverse populations where AI systems are deployed. These gaps are producing growing concern about bias, discrimination, due process, liability, and overall responsibility for harm.
>
> (Whittaker et al., 2018: 7).

To sum up, AI policy documents mention a number of reasons for bias in AI including pre-existing bias in society, lack of diversity in AI workforce, technical problems and lack of regulation and oversight. Data and design of algorithms reflect long-standing biases in society. In AI systems, that can be perceived as more objective, such biases tend to be less obvious. Furthermore, lack of diversity in AI workforce is seen as a major reason for bias because AI systems reflect views and perspectives of their designers. While recently AI workforce has been dominated by highly paid and technically educated male developers, policy documents remind that historically at early days of programming the field was female-dominated. Bias in AI can also be due to technical problems such as using datasets which are not representative of the wider population or untested and poorly designed AI systems that lack oversight and regulation.

Consequences of bias in AI

Policy documents highlight that bias in AI can have major negative consequences leading to discriminatory outcomes, disadvantaging certain groups and reinforcing stereotypes (IEEE, 2017). This becomes increasingly problematic as AI is used to make decisions over a broad range of issues in the fields such as finance, health, education, security and employment (BIC/APPGAI, 2017b). The AI Now 2016 report emphasize potential impact on AI to produce unfair outcomes:

> As AI systems take on a more important role in high-stakes decision-making – from offers of credit and insurance, to hiring decisions and parole – they will begin to affect who gets offered crucial opportunities, and who is left behind. This brings questions of rights, liberties, and basic fairness to the forefront.
>
> (Crawford & Whittaker, 2016: 6).

Harmful impacts of bias in AI applications in a number of major areas including recruitment, judicial systems and healthcare are discussed. One area of concern is impact of algorithms on employment prospects. The IEEE document highlights that uncritical use of AI in the workplace 'is of utmost concern due to high chance for error and biased outcome' (IEEE, 2017: 201). Another area of concern is use of algorithms in judicial system. Several documents (BIC/APPGAI, 2017b; Campolo et al., 2017; ICO, 2017) refer to the ProPublica study about the machine bias in the judicial risk assessment (Angwin et al., 2016) as an example of harmful outcomes. Furthermore, the dangers of bias embedded in AI health applications that can have an incredibly high cost leading to misdiagnosis and improper treatment are emphasized (Campolo et al., 2017). When bias in AI can have such a wide range of potentially damaging impacts that can further exacerbate structural inequalities and power imbalances, what are the recommendations for tackling it?

Recommendations for dealing with bias in AI

Recommendations outlined in the policy documents for tackling bias in AI focus on legal, technical, and educational measures as well as on increasing diversity of AI workforce and developing AI through transdisciplinary collaborations.

Regulation, guidelines and policy is presented as one way to deal with bias in AI. Documents emphasize the urgent need for regulation (Whittaker et al., 2018), updating of current law (Crawford & Whittaker, 2016), and strengthening of existing rights (CNIL, 2017). Related suggestions include developing and adopting tools and systems for testing and auditing to explain the rationale behind algorithmic decisions and check for bias, discrimination, and errors (House of Lords, 2018; ICO, 2017).

More technical suggestions for tackling bias include ensuring that data used is truly representative (House of Lords, 2018) and running pre-release trials (Campolo et al., 2017). According to the Royal Society report:

> Technological solutions can also help ensure machine learning systems handle data fairly, and in ways that are in accordance with anti-discrimination legislation. For example, machine learning systems can be coded in a way that restricts how they use different inputs.
>
> (The Royal Society, 2017: 114).

Education, awareness raising, and provision of information is recommended to address bias in AI. This includes embedding unconscious bias training in higher education and industry (Hall & Pesenti, 2017), raising awareness among developers about ethical issues as well as providing information to citizens about the decisions that affect them (CNIL, 2017).

Many AI policy documents suggest that increasing diversity of AI workforce is crucial because 'a diverse group of programmers reduces the risk of bias embedding into the algorithm and enables a fairer and higher quality output' (Hall & Pesenti, 2017: 56). Recommendations for achieving that include demonstrating the advantages of diversity for AI development and breaking down stereotypes (Hall & Pesenti, 2017), broadening participation, hiring developers from diverse gender, ethnic and socio-economic backgrounds and offering additional support for women (Hall & Pesenti, 2017; House of Lords, 2018). It is also suggested to request companies, universities, conferences and other stakeholders to 'release data on the participation of women, minorities and other marginalized groups within AI research and development' (Campolo et al., 2017: 2), address the cultures of exclusion and discrimination in the workplace (Whittaker et al., 2018) and build more inclusive workplaces (Campolo et al., 2017).

Further recommendations to increase diversity in AI include focus on fostering transdisciplinary approaches and perspectives involving diverse disciplines and stakeholder groups:

> There is an urgent need to expand cultural, disciplinary and ethnic diversity within the AI field in order to diminish groupthink, mitigate bias and broaden intellectual frames of reference beyond the purely technical.
>
> (Campolo et al., 2017: 20).

It is suggested that more diverse academic disciplines should be involved including philosophers, social scientists, legal theorists and political scientists who could bring their long-standing expertise

dealing with questions around bias and prejudice (House of Lords, 2018). Adding social perspectives can improve understanding of bias:

> The current focus on discrete technical fixes to systems should expand to draw on socially-engaged disciplines, histories, and strategies capable of providing a deeper understanding of the various social contexts that shape the development and use of AI systems.
>
> (Whittaker et al., 2018: 43).

The IEEE report suggests not only to include behavioral scientists but also target populations especially from potentially disadvantaged groups to diagnose and correct biases and assess whether AI applies norms in discriminatory ways to different races, ethnicities, genders, ages, body shapes, and so on (IEEE, 2017).

As the discussion above shows, policy documents suggest a range of ways to address bias in AI. However, they provide hardly any recommendations for tackling the root cause, namely, long-standing structural injustice and inequalities in society. The AI Now 2018 report points to the need to address issues of power and hierarchy and suggests that the shift in the balance of power in the public's favor

> will require significant structural change that goes well beyond a focus on technical systems, including a willingness to alter the standard operational assumptions that govern the modern AI industry players.
>
> (Whittaker et al., 2018: 43).

To sum up, documents provide a range of recommendations for dealing with bias, including calls for new and updated regulation, guidelines and policies. Recommendations also highlight the importance of education and awareness raising including in relation to ethical issues. To increase diversity of AI workforce, it is suggested not only to hire more women and minorities but also to build more inclusive workplace culture. Furthermore, recommendations for more diversity in AI focus on transdisciplinary collaborations bringing in a broader range of disciplines including social sciences and humanities as well as affected social groups. Some documents outline technological fixes and solutions to bias, while others argue for the need to expand focus to gain a deeper understanding of social contexts and structural change necessary in the way AI is developed. Overall, more explicit political recommendations are little discussed in the documents.

CONCLUSIONS: TECHNICAL AND SOCIAL FRAMINGS OF BIAS IN AI

In this contribution we have demonstrated that bias is part of broader policy debates about AI. Framing of bias highlights various intersectional characteristics, diverse causes and important negative consequences. A range of recommendations for tackling bias are outlined. Among various approaches of diagnosing problems and suggesting prescriptions, we can distinguish two stylized framings of bias in AI policy—one more technical, another more social. Technical framing suggests that AI can be solution to human bias and can help to detect and eliminate it. According to this frame, bias in AI is largely a technical problem that can be solved by technical means such as improving data quality and reliability and designing better algorithms. This frame resonates

with approaches such as technological fixes (Johnston, 2018), solutionism (Morozov, 2013) and technochauvinism (Broussard, 2023).

This popular technical framing is challenged by a more social (or socio-technical) narrative. According to this social framing, bias is not just a technical issue but has important historical, political and power aspects. This frame calls for going beyond technical fixes and having a deeper understanding of social contexts that shape development and use of AI systems. It recognizes that addressing bias in AI as a social issue would require structural change and a shift in the balance of power in the public's favor. This social framing of bias in AI reflects more political understanding of technology including AI as being co-shaped by pre-existing power structures. In AI policy debates this more social and critical approach has been shaped by the influential AI Now reports (Crawford & Whittaker, 2016; Campolo et al., 2017; Whittaker et al., 2018, 2019; West et al., 2019), which have influenced discussions on this topic in professional associations (IEEE, 2017), national and European institutions (CNIL, 2017; European Commission, 2018a) and other contexts (e.g., Collett & Dillon, 2019), expanding discussion of bias in AI by bringing in social aspects in addition to technical ones.

If prescription for addressing bias in the technical frame is a simple technological fix, then prescriptions in the case of social frame are less straightforward. We suggest that a productive way to think about bias would be to approach it as a complex social 'wicked problem' (Rittel & Webber, 1973) that can be addressed by a broader strategy involving diverse stakeholders and actions (Head, 2022; Ulnicane, 2022).

Examination of rhetorical frames matters for revealing competing understandings of legitimate problems and actions to address them. Existing hierarchies and structures behind the powerful technical framing of bias in AI might have strong vested interests to present easily available technocratic solutions, while avoiding discussions about political, social and economic roots of the problem. Therefore, it is of great importance that alternative social framing is challenging this technocratic approach and is pointing towards more complex social diagnosis of the problem and ways to tackle it. This social frame considerably broadens the legitimate understanding of bias in AI and the scope of possible actions beyond technological fix.

This contribution is the first step in examining policy framing of bias as a key issue in AI. Future studies would benefit from expanding the range of materials analyzed including other documents such as media reports and funding programmes as well as following up on if and how the rhetorical frames of technical and social understandings of bias are implemented through concrete actions and policy instruments. Such future studies would be important for deepening our understanding of how politics and power are co-shaping the ways in which bias is understood and tackled in AI policy and vice versa.

AI policy documents analyzed in this study have contributed to setting the scene for many policy, legislative, investment and educational initiatives around the world (Broussard, 2023; UNESCO, 2020). Having bias discussed in a number of AI policy documents has played an important role in shaping overall discourse of AI policy. While often this discourse focuses on economic benefits of AI, discussion of bias has contributed to highlighting problematic aspects and concerns about AI and its governance. In recent years, a number of dedicated reports have discussed gender and equality issues in AI (Collett & Dillon, 2019; European Commission, 2020b; UNESCO, 2019; UNESCO, 2020; West et al., 2019; Young et al., 2021). While such specialized reports provide an opportunity to go into depth in equality and AI issues, they can also contribute to perception of these topics as niche issues, which are primarily of concern for women and minorities and are thus marginal to overall AI policy and governance agenda. Rather than treating intersectional bias as a siloed issue, it is fruitful to see it as a unique opportunity to fundamentally

rethink major problems and concerns of AI governance and policy discussed in this special issue such as creating new social divides and prioritizing economic issues over social and political ones (e.g., Giest & Samuels, 2022; Kim, 2023; Rönnblom et al., 2023; Schiff, 2023). Intersectional and feminist approaches to AI can help to radically reimagine AI governance, policy, power and politics in more inclusive and participatory ways bringing in diverse groups, perspectives and voices.

ACKNOWLEDGMENTS

This contribution has benefited from feedback on earlier versions presented at several virtual events including the Science and Technology Studies Conference Graz 2021, International Conference on Public Policy 2021, the 2021 General Conference of the European Consortium for Political Research, and the special issue workshop in early 2022. We are grateful for support from the Frontrunners internship programme at De Montfort University. Early ideas for this contribution were discussed with colleagues in the Human Brain Project. In March 2023, discussions at the Shaping AI workshop in London and at the Algorithms for Her 2 conference in Sheffield provided much needed inspiration for final revision.

FUNDING INFORMATION

The research reported in this article has received funding from the EU's Horizon 2020 Research and Innovation Programme under grant agreements no. 720270 (HBP SGA1), no. 785907 (HBP SGA2), and 945539 (HBP SGA3).

ORCID

Inga Ulnicane https://orcid.org/0000-0003-2051-1265

ENDNOTES

[1] Abbreviation 'AGI' stands for 'Artificial General Intelligence', namely envisaged future stage of AI development when AI is expected to be able to perform all the tasks that humans do (Mitchell, 2019).

[2] Use of male pseudonyms in tech industry has a longer history. One well-known case comes from the 1960s when entrepreneur Stephanie Shirley in the UK signed her letters as Steve (Winterson, 2021:211). Wiener (2019) shows that it is still necessary in the 21st century Silicon Valley.

[3] Exceptions here are discussions of gender issues in AI policy in the European Union and Spain by Ariana Guevara-Gómez et al. (2021) and in Sweden by Malin Rönnblom et al. (2023).

[4] Details on the methodology and creation of this dataset of AI policy documents are elaborated in Ulnicane, Knight, et al., 2021.

[5] The CNIL, Commission Nationale Informatique & Libertés, is the French Data Protection Agency.

REFERENCES

af Malmborg, F. (2022). Narrative dynamics in European Commission AI policy—Sensemaking, agency construction, and anchoring. *Review of Policy Research*, 1–25. https://doi.org/10.1111/ropr.12529

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks*. ProPublica.

Bacchi, C. (2000). Policy as discourse: What does it mean? Where does it get us? *Discourse: Studies in the Cultural Politics of Education*, *21*(1), 45–57.

Bender, E. M., Gebru, T., McMillan-Major, A. and Shmitchell, S. (2021) On the dangers of stochastic parrots: Can language models Be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610-623.

Benjamin, R. (2019). *Race after technology. Abolitionist tools for the New Jim Code*. Polity.

Broussard, M. (2023). *More than a Glitch. Confronting Race, Gender, and Ability Bias in Tech*. The MIT Press.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, *81*, 1–15.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Cave, S., & Dihal, K. (2020). The whiteness of AI. *Philosophy & Technology*, *33*(4), 685–703. https://doi.org/10.1007/s13347-020-00415-6

Ciston, S. (2019). Intersectional AI is essential: Polyvocal, multimodal, experimental methods to save artificial intelligence. *Journal of Science and Technology of the Arts*, *11*(2), 3–8. https://doi.org/10.7559/citarj.v11i2.665

Coad, A., Nightingale, P., Stilgoe, J., & Vezzani, A. (2021). The dark side of innovation. *Industry and Innovation*, *28*(1), 102–112. https://doi.org/10.1080/13662716.2020.1818555

Coeckelbergh, M. (2022). *The Political Philosophy of AI. An Introduction*. Polity.

Collett, C., & Dillon, S. (2019). *AI and Gender: Four Proposals for Future Research*. The Leverhulme Centre for the Future of Intelligence. doi:10.17863/CAM.41459

Courtland, R. (2018). The bias detectives. *Nature*, *558*(7710), 357–360.

Crenshaw, K. (1991). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, *43*(6), 1241–1299.

Criado Perez, C. (2019). *Invisible Women: Data Bias in A World Designed For Men*. Random House.

D'Ignazio, C., & Klein, L. F. (2020). *Data Feminism*. MIT Press.

Eubanks, V. (2019). *Automating Inequality. How High-Tech Tools Profile, Police, and Punish the Poor*. Picador.

European Commission (2019a) *A Definition of Artificial Intelligence: Main Capabilities and Scientific Disciplines. High-Level Expert Group on Artificial Intelligence*. European Commission. https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines (Accessed 9 June 2023).

European Commission (2019b) *Ethics Guidelines for Trustworthy AI. High-level expert group on artificial intelligence*. European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (Accessed 9 June 2023).

European Commission (2020a) *On Artificial Intelligence – A European Approach to Excellence and Trust*. European Commission. White Paper. COM(65) 19.2.2020.

European Commission (2020b) *Gender & Intersectional Bias in Artificial Intelligence. Dialogue Factsheet*. European Commission. https://op.europa.eu/en/publication-detail/-/publication/286e1432-021a-11eb-836a-01aa75ed71a1 (Accessed 9 June 2023).

Fothergill, B. T., Knight, W., Stahl, B. C., & Ulnicane, I. (2019). Intersectional observations of the human brain Project's approach to sex and gender. *Journal of Information, Communication and Ethics in Society*, *17*(2), 128–144. https://doi.org/10.1108/JICES-11-2018-0091

Freeman, R., & Maybin, J. (2011). Documents, practices and policy. *Evidence & Policy*, *7*(2), 155–170.

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, *14*(3), 330–347.

Garvey, S. C. (2021). Unsavory medicine for technological civilization: Introducing 'Artificial Intelligence & its Discontents'. *Interdisciplinary Science Reviews*, *46*(1–2), 1–18. https://doi.org/10.1080/03080188.2020.1840820

Giest, S., & Samuels, A. (2022). Administrative burden in digital public service delivery: The social infrastructure of library programs for e-inclusion. *Review of Policy Research*, 1–20. https://doi.org/10.1111/ropr.12516

Gruber, M., & Benedikter, R. (2021). The role of women in contemporary technology and the feminization of artificial intelligence and its devices. In T. Keskin & R. D. Kiggins (Eds.), *Towards an International Political Economy of Artificial Intelligence* (pp. 17–38). Palgrave.

Guevara-Gómez, A., de Zárate-Alcarazo, L. O., & Criado, J. I. (2021). Feminist perspectives to artificial intelligence: Comparing the policy frames of the European Union and Spain. *Information Polity*, *26*(2), 173–192.

Head, B. W. (2022). *Wicked Problems in Public Policy. Understanding and Responding to Complex Challenges*. Palgrave Macmillan.

Hicks, M. (2018). *Programmed Inequality: How Britain Discharged Women Technologists and Lost Its Edge In Computing*. The MIT Press.

Jasanoff, S. (2003). Technologies of Humility: Citizen participation in governing science. *Minerva*, *41*(3), 223–244. https://doi.org/10.1023/A:1025557512320

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

Johnston, S. F. (2018). The technological fix as social cure-all: Origins and implications. *IEEE Technology and Society Magazine*, *37*(1), 47–54.

Kim, J. (2023). Traveling AI-essentialism and national AI strategies: A comparison between South Korea and France. *Review of Policy Research*, 1–24. https://doi.org/10.1111/ropr.12552

Koene, A., Dowthwaite, L. and Seth, S. (2018) IEEE P7003™ standard for algorithmic bias considerations: Work in progress paper. In Proceedings of the International Workshop on Software Fairness, pp. 38–41.

Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: Review, synthesis, and future research directions. *European Journal of Information Systems*, *31*(3), 388–409.

Little, B., & Winch, A. (2021). *The New Patriarchs of Digital Capitalism: Celebrity Tech Founders and Networks of Power*. Routledge.

Liu, W. (2020). *Abolish Silicon Valley: How to Liberate Technology from Capitalism*. Repeater.

Mitchell, M. (2019). *Artificial Intelligence: A Guide for Thinking Humans*. Penguin Books.

Morozov, E. (2013). *To Save Everything, Click Here. Technology, Solutionism and the Urge to Fix Problems That Don't Exist*. Penguin Books.

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press.

O'Neil, C. (2016). *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy*. Penguin Books.

Radu, R. (2021). Steering the governance of artificial intelligence: National strategies in perspective. *Policy and Society*, *40*(2), 178–193. https://doi.org/10.1080/14494035.2021.1929728

Rein, M., & Schon, D. (1993). Reframing policy discourse. In F. Fischer & J. Forester (Eds.), *The Argumentative Turn in Policy Analysis and Planning* (pp. 145–166). UCL Press.

Rein, M., & Schon, D. (1996). Frame-critical policy analysis and frame-reflective policy practice. *Knowledge and Policy*, *9*(1), 85–104. https://doi.org/10.1007/BF02832235

Rittel, H. W., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy Sciences*, *4*(2), 155–169. https://doi.org/10.1007/BF01405730

Rönnblom, M., Carlsson, V., & Öjehag-Pettersson, A. (2023). Gender equality in Swedish AI policies. What's the problem represented to be? *Review of Policy Research*, 1–17. https://doi.org/10.1111/ropr.12547

Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology? *Research Policy*, *44*(10), 1827–1843. https://doi.org/10.1016/j.respol.2015.06.006

Sadowski, N., & Phan, T. (2022). 'Open secrets': An interview with Meredith Whittaker. In T. Phan, J. Goldenfein, D. Kuch, & M. Mann (Eds.), *Economies of Virtue: The Circulation of 'Ethics' in AI* (pp. 140–152). Institute of Network Cultures.

Schiff, D. (2023). Looking through a policy window with tinted glasses: Setting the agenda for U.S. AI Policy. *Review of Policy Research*, 1–28. https://doi.org/10.1111/ropr.12535

Schiff, D., Borenstein, J., Biddle, J., & Laas, K. (2021). AI ethics in the public, private, and NGO sectors: A review of a global document collection. *IEEE Transactions on Technology and Society*, *2*(1), 31–42. https://doi.org/10.1109/TTS.2021.3052127

Schon, D., & Rein, M. (1994). *Frame Reflection: Toward the Resolution Of Intractable Policy Controversies*. Basic Books.

Schopmans, H., & Cupac, J. (2021). Engines of patriarchy: Ethical artificial intelligence in times of illiberal backlash politics. *Ethics & International Affairs*, *35*(3), 329–342.

Simon, J., Wong, P. H., & Rieder, G. (2020). Algorithmic bias and the value sensitive design approach. *Internet Policy Review*, *9*(4), 1–16. https://doi.org/10.14763/2020.4.1534

Søraa, R. (2023). *AI for Diversity*. CRC Press.

Toupin, S. (2023). Shaping feminist artificial intelligence. *New Media & Society*, 146144482211507.

Ulnicane, I. (2022). Emerging technology for economic competitiveness or societal challenges? Framing purpose in artificial intelligence policy. *Global Public Policy and Governance*, *2*(3), 326–345. https://doi.org/10.1007/s43508-022-00049-8

Ulnicane, I., Eke, D. O., Knight, W., Ogoh, G., & Stahl, B. C. (2021). Good governance as a response to discontents? Déjà vu, or lessons for AI from other emerging technologies. *Interdisciplinary Science Reviews*, *46*(1–2), 71–93. https://doi.org/10.1080/03080188.2020.1840220

Ulnicane, I., Knight, W., Leach, T., Stahl, B. C., & Wanjiku, W.-G. (2021). Framing governance for a contested emerging technology: Insights from AI policy. *Policy and Society*, *40*(2), 158–177. https://doi.org/10.1080/14494035.2020.1855800

Ulnicane, I., Knight, W., Leach, T., Stahl, B. C., & Wanjiku, W.-G. (2022). Governance of artificial intelligence: Emerging international trends and policy frames. In M. Tinnirello (Ed.), *The Global Politics of Artificial Intelligence* (pp. 29–55). CRC Press. https://doi.org/10.1201/9780429446726-2

UNESCO (2019) *I'd blush if I could: Closing gender divides in digital skills through education*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=1 (Accessed 9 June 2023).

UNESCO (2020) *Artificial Intelligence and Gender Equality. Key findings of UNESCO's Global Dialogue*. https://en.unesco.org/AI-and-GE-2020 (Accessed 9 June 2023).

Van Lente, H., Spitters, C., & Peine, A. (2013). Comparing technological hype cycles: Towards a theory. *Technological Forecasting and Social Change*, *80*(8), 1615–1628. https://doi.org/10.1016/j.techfore.2012.12.004

Wellner, G., & Rothman, T. (2020). Feminist AI: Can we expect our AI systems to become feminist? *Philosophy & Technology*, *33*(2), 191–205.

West, S. M. (2020). Redistribution and Rekognition: A feminist critique of algorithmic fairness. *Catalyst: Feminism, Theory, Technoscience*, *6*(2), 1–24.

West, S. M., Whittaker, M., & Crawford, K. (2019). *Discriminating Systems: Gender, Race and Power in AI*. AI Now Institute. https://ainowinstitute.org/publication/discriminating-systems-gender-race-and-power-in-ai-2 Accessed 9 June 2023.

Whittaker, M., Alper, M., Bennett, C. L., Hendren, S., Kaziunas, L., Mills, M., Morris, M. R., Rankin, J., Rogers, E., Salas, M., & West, S. M. (2019). *Disability, bias, and AI*. AI Now Institute. https://ainowinstitute.org/publication/disabilitybiasai-2019 Accessed 9 June 2023.

Wiener, A. (2020) *Uncanny Valley. Seduction and Disillusionment in San Francisco's Startup Scene*. 4th Estate.

Winner, L. (1980). Do artifacts have politics? *Daedalus*, *109*(1), 121–136.

Winterson, J. (2021). *12 Bytes: How We Got Here. Where We Might Go Next*. Jonathan Cape.

Wong, P. H. (2020). Democratizing algorithmic fairness. *Philosophy & Technology*, *33*(2), 225–244.

Young, E., Wajcman, J., & Sprejer, L. (2021). *Where Are the Women? Mapping the Gender Job Gap in AI*. The Alan Turing Institute. https://www.turing.ac.uk/news/publications/report-where-are-women-mapping-gender-job-gap-ai Accessed 9 June 2023.

Zou, J., & Schiebinger, L. (2018). AI can be sexist and racist—it's time to make it fair. *Nature*, *559*, 324–326. https://doi.org/10.1038/d41586-018-05707-8

ANNEX 1

Dataset of AI policy documents analyzed (in alphabetical order):

Accenture (2017) *Embracing Artificial Intelligence. Enabling Strong and Inclusive AI Driven Growth*. Accenture.

BIC/APPGAI (Big Innovation Centre/All-Party Parliamentary Group on Artificial Intelligence) (2017a) *APPG AI Findings 2017*. Big Innovation Centre.

BIC/APPGAI (Big Innovation Centre/All-Party Parliamentary Group on Artificial Intelligence) (2017b) *Governance, Social and Organizational Perspective for AI*. Big Innovation Centre.

BIC/APPGAI (Big Innovation Centre/All-Party Parliamentary Group on Artificial Intelligence) (2017c) *Inequality, Education, Skills, and Jobs*. Big Innovation Centre.

BIC/APPGAI (Big Innovation Centre/All-Party Parliamentary Group on Artificial Intelligence) (2017d) *International Perspective and Exemplars*. Big Innovation Centre.

BIC/APPGAI (Big Innovation Centre/All-Party Parliamentary Group on Artificial Intelligence) (2017e) *What is AI? A theme report based on the 1st meeting of the All-Party Parliamentary Group on Artificial Intelligence*. Big Innovation Centre.

Bowser, A., M. Sloan, P. Michelucci and E. Pauwels (2017) *Artificial Intelligence: A Policy-Oriented Introduction. Wilson Briefs*. Wilson Center. Accessed June 1, 2023. https://www.wilsoncenter.org/publication/artificial-intelligence-policy-oriented-introduction

Campolo, A, M. Sanfilippo, M. Whittaker and K. Crawford (2017) *AI Now 2017 Report*. AI Now Institute, New York University. Accessed June 1, 2023. https://ainowinstitute.org/publication/ai-now-2017-report-2

CNIL (Commission Nationale Informatique & Libertes) (2017) *How can humans keep the upper hand? The ethical matters raised by algorithms and artificial intelligence*. CNIL. Accessed June 1, 2023. https://www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_ai_gb_web.pdf

Crawford, K. and M. Whittaker (2016) *The AI Now Report. The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term*. AI Now Institute, New York University. Accessed June 1, 2023. https://artificialintelligencenow.com/media/documents/AINowSummaryReport_3_RpmwKHu.pdf

EDPS (European Data Protection Supervisor). (2016) Artificial Intelligence, Robotics, Privacy and Data Protection. Room document for the 38th International Conference of Data Protection and Privacy Commissioners. Accessed June 1, 2023. https://edps.europa.eu/sites/edp/files/publication/16-10-19_marrakesh_ai_paper_en.pdf.

European Commission. (2017). *AI Policy Seminar: Towards and EU strategic plan for AI*. Digital Transformation Monitor.

European Commission. (2018a). *Artificial Intelligence: A European Perspective*. Publications Office of the European Union. https://data.europa.eu/doi/10.2760/11251 Accessed 1 June 2023.

European Commission (2018b) *Artificial Intelligence for Europe. COM/2018/237 final*. European Commission. Accessed June 1, 2023. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2018:237:FIN

European Commission (2018c) *Coordinated Plan on Artificial Intelligence. COM/2018/795 final*. European Commission. Accessed June 1, 2023. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52018DC0795

European Economic and Social Committee (2017) Artificial Intelligence - The consequences of Artificial intelligence on the (digital) single market, production, consumption, employment and society. Opinion. OJ C 288, 31.8.2017, p. 1–9. Accessed June 1, 2023. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52016IE5369

European Group on Ethics in Science and New Technologies (2018) *Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems*. European Commission. Accessed June 1, 2023. https://data.europa.eu/doi/10.2777/531856

European Parliament (2016) *European Civil Law Rules in Robotics. Study for the JURI Committee*. European Parliament. Accessed June 1, 2023. https://www.europarl.europa.eu/RegData/etudes/STUD/2016/571379/IPOL_STU(2016)571379_EN.pdf

European Parliament (2017) *Resolution with recommendations to the Commission on Civil Law Rules on Robotics*. European Parliament. Accessed June 1, 2023. https://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_EN.html

European Parliament (2018) *Understanding Artificial Intelligence. Briefing EPRS*. European Parliament. Accessed June 1, 2023. https://www.europarl.europa.eu/RegData/etudes/BRIE/2018/614654/EPRS_BRI%282018%29614654_EN.pdf

Executive Office of the President (2016a) *Artificial Intelligence, Automation, and Economy, Report*. Executive Office of the President. Accessed June 1, 2023. https://obamawhitehouse.archives.gov/sites/whitehouse.gov/files/documents/Artificial-Intelligence-Automation-Economy.PDF

Executive Office of the President. (2016b). *Preparing for the future of artificial intelligence*. National Science and Technology Council Committee on Technology. Accessed June 1, 2023. https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf

Executive Office of the President. (2016c). *The National Artificial Intelligence research and development Strategic Plan*. National Science and Technology Council. Networking and Information Technology Research and Development Subcommittee. Accessed June 1, 2023. https://www.nitrd.gov/pubs/national_ai_rd_strategic_plan.pdf

Future of Humanity Institute et al (2018) *The Malicious Use of Artificial Intelligence: Forecasting, Prevention and Mitigation*. Future of Humanity Institute. Accessed June 1, 2023. https://maliciousaireport.com/

Government Office for Science (2016) *Artificial Intelligence: opportunities and implications for the future of decision making*. Government Office for Science. Accessed June 1, 2023. https://www.gov.uk/government/publications/artificial-intelligence-an-overview-for-policy-makers

Hall, W., & Pesenti, J. (2017). *Growing the Artificial Intelligence Industry in the UK*. Independent review. Accessed June 1, 2023. https://www.gov.uk/government/publications/growing-the-artificial-intelligence-industry-in-the-uk

HM (Her Majesty's) Government (2018) *Artificial Intelligence Sector Deal*. HM Majesty. Accessed June 1, 2023. https://www.gov.uk/government/publications/artificial-intelligence-sector-deal

House of Commons Science and Technology Committee (2016) *Robotics and Artificial Intelligence*. House of Commons.

House of Lords (2018) *AI in the UK: Ready, Willing and Able?* House of Lords. Accessed June 1, 2023. https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf

ICO (Information Commissioner's Office) (2017) *Big data, artificial intelligence, machine learning and data protection. Data Protection Act and General Data Protection Regulation*. Information Commissioner's Office). Accessed June 1, 2023. https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf

IEEE (Institute of Electrical and Electronics Engineers) (2017) *Ethically aligned design. A vision for prioritizing human well-being with autonomous and intelligent systems. Version 2*. IEEE. Accessed June 1, 2023. https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

IEEE European Public Policy Initiative (2017) *Artificial Intelligence: Calling on Policy Makers to Take a Leading Role in Setting a Long-Term AI Strategy. Position Statement*. IEEE. Accessed June 1, 2023. https://www.ieee.org/content/dam/ieee-org/ieee/web/org/about/european-public-policy/artificial_intelligence.pdf

IEEE-USA (2017) *Artificial Intelligence Research, Development & Regulation. Position Statement*. IEEE. Accessed June 1, 2023. http://globalpolicy.ieee.org/wp-content/uploads/2017/10/IEEE17003.pdf

International Telecommunication Union (2017) *AI for Good Global Summit Report 2017, Geneva*. International Telecommunication Union. Accessed June 1, 2023. https://www.itu.int/en/ITU-T/AI/Documents/Report/AI_for_Good_Global_Summit_Report_2017.pdf

IPPR (Institute for Public Policy Research). (2017). *Managing Automation: Employment, Inequality and Ethics in the Digital Age*. Discussion Paper. Accessed June 1, 2023. https://www.ippr.org/research/publications/managing-automation

Ministry of Economic Affairs and Employment (2017) *Finland's Age of Artificial Intelligence*. Ministry of Economic Affairs and Employment. Accessed June 1, 2023. https://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/160391/TEMrap_47_2017_verkkojulkaisu.pdf

Ponce Del Castillo, A. (2017). *A Law on Robotics and Artificial Intelligence in the EU? Foresight Brief*. European Trade Union Institute ETUI. Accessed June 1, 2023. https://www.etui.org/publications/foresight-briefs/a-law-on-robotics-and-artificial-intelligence-in-the-eu

Rathenau Institute (2017) *Human Rights in the Robot Age. Challenges arising from the use of robotics, artificial intelligence, and virtual and augmented reality. Report for the Parliamentary Assembly of the Council of Europe*. Rathenau Institute. Accessed June 1, 2023. https://www.rathenau.nl/en/digitalisering/human-rights-robot-age

SGPAC. (2017). *Governance, Risk & Control: Artificial Intelligence. Effective Deployment, Management and Oversight of Artificial Intelligence (AI). Version 1.0*. SGPAC Consulting & Advisory. Accessed June 1, 2023. https://www.sgpac.co.uk/wp-content/uploads/2017/02/GRC-Framework-for-AI-v1.0-Published-22-March-2017.pdf

Tata Leading the Way with Artificial Intelligence: The Next Big Opportunity for Europe. TCS Global Trend Study – Europe. (2017). Tata Consultancy Services.

The 2015 panel. (2016). *Artificial Intelligence and life in 2030. One hundred year study on artificial intelligence. Report of the 2015 study panel*. Stanford University. Accessed June 1, 2023. https://ai100.stanford.edu/2016-report

The Federal Government (2018) *Artificial Intelligence Strategy*. The Federal Government. Accessed June 1, 2023. https://www.ki-strategie-deutschland.de/home.html

The Royal Society (2017) *Machine Learning: The Power and Promise of Computers That Learn By Example*. The Royal Society. Accessed June 1, 2023. https://royalsociety.org/~/media/policy/projects/machine-learning/publications/machine-learning-report.pdf

Thierer, A., Castillo O'Sullivan, A., & Russell, R. (2017). *Artificial Intelligence and Public Policy Report*. Mercatus Center, George Mason University. Accessed June 1, 2023. https://www.mercatus.org/research/research-papers/artificial-intelligence-and-public-policy

UNI (Union Network International) Global Union (2017) *Top 10 Principles for Ethical Artificial Intelligence. The Future World of Work*. UNI Global Union. Accessed June 1, 2023. http://www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf

Villani, C. (2018) *For a Meaningful Artificial Intelligence. Towards a French and European Strategy*. The President of French Republic. Accessed June 1, 2023. https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf

Vinnova. (2018) *Artificial Intelligence in Swedish Business and Society*. Vinnova. Accessed June 1, 2023. https://www.vinnova.se/contentassets/29cd313d690e4be3a8d861ad05a4ee48/vr_18_09.pdf

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., Myers West, S., Richardson, R., Schultz, J., & Schwartz, O. (2018). *AI Now Report 2018*. AI Now Institute, New York University. Accessed June 1, 2023. https://ainowinstitute.org/publication/ai-now-2018-report-2

World Economic Forum. (2018). Artificial Intelligence for the Common Good. Sustainable, Inclusive and Trustworthy. White Paper for attendees of the WEF 2018 Annual Meeting.

AUTHOR BIOGRAPHIES

Dr. Inga Ulnicane is a Senior Research Fellow at De Montfort University, Leicester, UK. Her research focusses on politics and policy of science, technology and innovation, governance of emerging technologies and societal challenges. Her publications have appeared in journals such as Policy and Society, Journal of Responsible Innovation and Science and Public Policy. She has also prepared commissioned reports for the European Parliament and European Commission. She has PhD from Science, Technology and Policy Studies Department at University of Twente, Netherlands.

Aini Aden contributed to this study during her DMU Frontrunners internship October 2020–March 2021. At that time, she was a third-year psychology student at De Montfort University, Leicester, UK.

How to cite this article: Ulnicane, I., & Aden, A. (2023). Power and politics in framing bias in Artificial Intelligence policy. *Review of Policy Research*, *00*, 1–23. https://doi.org/10.1111/ropr.12567