Chapter 6
Diversity and Inclusion in Artificial Intelligence

Eduard Fosch-Villaronga and Adam Poulsen

Contents

Abstract Discrimination and bias are inherent problems of many AI applications, as seen in, for instance, face recognition systems not recognizing dark-skinned women and content moderator tools silencing drag queens online. These outcomes may derive from limited datasets that do not fully represent society as a whole or from the AI scientific community's western-male configuration bias. Although being a pressing issue, understanding how AI systems can replicate and amplify inequalities and injustice among underrepresented communities is still in its infancy in social science and technical communities. This chapter contributes to filling this gap by exploring the research question: what do diversity and inclusion mean in the context of AI? This chapter reviews the literature on diversity and inclusion in AI to unearth

E. Fosch-Villaronga (✉)
eLaw Center for Law and Digital Technologies, Leiden University, Leiden, The Netherlands
e-mail: e.fosch.villaronga@law.leidenuniv.nl

A. Poulsen
Charles Sturt University, Bathurst, New South Wales, Australia
e-mail: apoulsen@csu.edu.au

B. Custers and E. Fosch-Villaronga (eds.), *Law and Artificial Intelligence*,
Information Technology and Law Series 35,
https://doi.org/10.1007/978-94-6265-523-2_6

the underpinnings of the topic and identify key concepts, research gaps, and evidence sources to inform practice and policymaking in this area. Here, attention is directed to three different levels of the AI development process: the technical, the community, and the target user level. The latter is expanded upon, providing concrete examples of usually overlooked communities in the development of AI, such as women, the LGBTQ+ community, senior citizens, and disabled persons. Sex and gender diversity considerations emerge as the most at risk in AI applications and practices and thus are the focus here. To help mitigate the risks that missing sex and gender considerations in AI could pose for society, this chapter closes with proposing gendering algorithms, more diverse design teams, and more inclusive and explicit guiding policies. Overall, this chapter argues that by integrating diversity and inclusion considerations, AI systems can be created to be more attuned to all-inclusive societal needs, respect fundamental rights, and represent contemporary values in modern societies.

Keywords Artificial Intelligence · Gender · Diversity · Inclusion · LGBT · AI Act

6.1 Introduction

Artificial Intelligence (AI) technologies help automate industrial, retail, and farming sectors and, lately, healthcare, education, and public service. While AI can increase resource efficiency and productivity, automating parts of society reserved once only to humans is nonetheless straightforward and raises particular ethical, legal, and societal concerns.[1] A growing global concern is that AI systems may exacerbate and reinforce existing biases that different societies have with respect to gender, age, race, and sexual orientation.[2] For instance, face recognition systems having difficulty recognizing dark-skinned women and content moderator tools may automatically flag how drag queens use language as *toxic* and prevent them from freely communicating online.[3]

These outcomes may derive from limited datasets that do not fully represent the society[4] or from the AI scientific community's structural and systematic configuration biases.[5] Still, they are very influential.[6] For instance, there is an exponential growth of social robots and voice assistants that can socially interact with users. A common feature of these artefacts is that many of them are given female names, have female voices, and usually display a servile personality engineered to please users all the time.[7] The use of female voices for serviceable contexts reinforces

[1] Schönberger 2019; Wisskirchen et al. 2017; Righetti et al. 2019.

[2] Noble 2018; Raji and Buolamwin 2019; Fosch-Villaronga et al. 2021.

[3] Raji and Buolamwini 2019; Gomes et al. 2019.

[4] Zhao et al. 2017.

[5] Roopaei et al. 2021.

[6] Willson 2017; Noble 2018; Ito 2019.

[7] Liu 2021; Giger et al. 2019.

gender stereotypes about the role women should (or should not) play in society.[8] And these are usually biases rooted in oppressive gender inequalities that have existed throughout history and are typically exacerbated by the lack of diversity of the technical teams developing algorithms which usually work in companies with significant gender disparities in their board of directors.[9] Similar concerns are found in other AI applications, namely in algorithms for medical applications,[10] gender classifiers for marketing, social media platforms or recruiting practices, resulting in disparities in hiring.[11] Likewise, sex robotics often target straight males and objectify women's bodies.[12]

The scientific community broadly supports the narrative that integrating gender and sex factors in research makes better science.[13] However, many disciplines struggle to account for diversity. Authors continuously report that 'inequality and a lack of gender diversity still exist in medicine, especially in academia and leadership;'[14] and that 'when we look to the diversity in immunology research labs, overwhelmingly, women, people of color and LGBTQIA+ scientists are underrepresented among the laboratory head and top leadership roles.'[15] The AI community is no different in this respect, as highlighted by recent studies that explored gender biases in the community, i.e., 'our results indicate a huge gender disbalance among authors, a lack of geographical diversity (with no representation of the least developed countries and very low representation of African countries).'[16] Missing sex and gender considerations in the development of AI, however, can lead to adverse consequences for society that range from exacerbating existing biases and stereotypes (which are prohibited by law)[17] to safety concerns if misgendering a person in health-related applications.[18]

Although being a pressing issue, understanding how AI systems can replicate and amplify inequalities and injustice among underrepresented communities is still in its infancy in social science and technical communities. This chapter contributes to filling this gap by exploring the research question: what do diversity and inclusion mean in the context of AI? To address this question, this chapter reviews the literature on diversity and inclusion in AI to unearth the underpinnings of the topic. We identify key concepts, research gaps, and evidence sources to inform practice and

[8] Danielescu 2020.

[9] West et al. 2019; Rahman and Billionniere 2021.

[10] Cirillo et al. 2020.

[11] Park and Woo 2019.

[12] Richardson 2016.

[13] Schiebinger 2014; Tannenbaum et al. 2019.

[14] Ekmekcioglu 2021.

[15] Groom 2021.

[16] Freireç et al. 2020.

[17] See Article 5 of the Convention on the Elimination of All Forms of Discrimination against Women and Article 8(1)(b) of the Convention on the Rights of Persons with Disabilities.

[18] Cirillo et al. 2020.

policymaking in this area. As the most salient diversity and inclusion concerns in AI, sex and gender considerations are the focus here.

This chapter is structured as follows. First, three different levels of the AI development process where diversity and inclusion could be addressed are identified in Sect. 6.2: the technical, the community, and the target user level. Then, the implications of missing diversity and inclusion in AI affecting the target user level are expanded upon in Sect. 6.3, focusing on usually overlooked communities, namely women, the LGBTQ+ community, senior citizens, and disabled persons. This is done by examining four AI application case studies: social robots and gendered voices, algorithms for medical applications, gender classifiers for marketing, social media platforms, or recruiting practices, and sex robotics and gender-specific target market. In Sect. 6.4, mitigation strategies to account for missing sex and gender considerations in AI are proposed, including gendering algorithms, more diverse design teams, and more inclusive and explicit guiding policies. After that, this chapter closes with concluding remarks in Sect. 6.5.

6.2 Diversity and Inclusion in Artificial Intelligence

Like many concepts, such as intelligence, personality, or emotions, there are many ways to define, experience, and legalize diversity and inclusion. The dictionary defines diversity as 'the practice or quality of including or involving people from a range of different social and ethnic backgrounds and of different genders, sexual orientations.'[19]

In the context of AI, those at Google Research define diversity and inclusion as follows:[20]

- *Diversity*: Variety in the representation of individuals in an instance or set of instances, with respect to sociopolitical power differentials (gender, race, etc.). Greater diversity means a closer match to a target distribution over socially relevant characteristics.
- *Inclusion*: Representation of an individual user within an instance or a set of instances, where greater inclusion corresponds to better alignment between a user and the options relevant to them in an instance or set.

Given these definitions, diversity and inclusion in AI have ramifications at three different levels on which we expand here. The first one is the technical level, in which questions around the diversity of algorithms, techniques, and applications are centred around: are the algorithms taking into account all the necessary variables? Are these algorithms classifying users in discriminatory ways? The second level is the community surrounding the configuration, development, and deployment of such techniques and the questions revolving around their practices and how inclusive and

[19] *See* Lexico's definition at https://www.lexico.com/definition/diversity.

[20] Mitchell et al. 2020.

diverse they are: does the team have enough female representation? Are all the team members from the same background? The third level refers to the target user and focuses on questions about the person with whom the system will be interacting and affecting and often respond to questions around Responsible Research and Innovation (RRI): was the project conducted by taking all the stakeholders into account? Did the research include the users for feedback?

6.2.1 Technical Level

Algorithms are human-made and are likely to replicate human-like biases.[21] At the technical level, algorithms usually work in binary terms (e.g., yes/no, black/white, move/doesn't move) as if the world were a simple classification problem to be solved. However, the world is not black or white; it is not masculine or feminine. Think for instance the case of gender classifiers whose algorithms usually take *sex* as a primary point of reference when tasked with classifying users gender-wise: male or female. Gender Classification Systems (GCS) are trained using a training dataset (or corpus) of structured and labelled data. These labels categorize data, and the features within, as either masculine or feminine.

However, *sex*, *gender*, and *sexuality* are different concepts although they are often used in overlapping ways:[22]

- "*Sex*" usually refers to the assigned gender at birth based on sex characteristics (e.g. genitalia, chromosomes and hormones), usually 'male' or 'female.'—and in some cases 'indeterminate' for persons with intersex characteristics, in some places (e.g., New Zealand[23]). As one part of many gender-affirmation healthcare actions, medical transition can be engaged to accord sex characteristics with one's gender identity.[24]
- "*Gender*" is both a "person's internal, deeply held sense of their gender," also called *gender identity*[25]—also tied to social, cultural, and legal factors.
- "*Sexuality*" is taken to mean the 'physical, romantic, and/or emotional attraction to another person.'[26]

By using a binary understanding of *sex* as basis for algorithms, inferred data may lead to inaccuracies, e.g., systems can misclassify users whose *gender* differs from their *sex*.[27] Not classifying users correctly in gender terms can lead to bias and unfair decisions and may lead to self-fulfilling prophecies, a phenomenon well-known in

[21] O'Neil 2016; Caliskan et al. 2017.

[22] Fosch-Villaronga et al. 2021.

[23] *See* https://www.legislation.govt.nz/act/public/1995/0016/latest/DLM359369.html.

[24] *See* https://www.acon.org.au/wp-content/uploads/2019/07/TGD_Language-Guide.pdf.

[25] *See* https://www.glaad.org/reference/transgender.

[26] Ibidem.

[27] Fosch-Villaronga et al. 2021.

profiling.[28] These effects may amplify inequality, reinforce binarism, exacerbate gender stereotyping, and further push people into categories that are hard to break out.[29] This is particularly important because gender stereotyping is not only 'a generalized view or preconception about attributes or characteristics, or the roles that are or ought to be possessed by, or performed by, women and men.'[30] Gender stereotypes also affect members of the LGBTQ+ community, who often are subsumed under these roles too, e.g., gay men perceived to be feminine would map onto traditional 'warm but less competent' female stereotypes.[31] It also adversely affects the non-binary and transsexual communities, as it essentializes the body as the source of gender and cannot be accurately classified.[32]

6.2.2 Community Level

The AI community is not very diverse. As shown in a recent study reporting the lack of diversity amongst participants in top international AI conferences[33] or in the lack of gender balance ratio among the editors of AI journals (see Fig. 6.1), the AI community has been and continues to be male-based:

Historically, technological development seemed to refuse to acknowledge the existence of women and the LGBTQ+ community in science, as if science was only reserved to men.[34] Being queer was criminalized or devalued in many societies, and women were restricted to caring for the family and children upbringing. Not that long ago, in the 1950s, countries prosecuted homosexual scientists, as the United Kingdom famously did with Alan Turing. Elsewhere in the 1950s, Germany's opinion of the scientific community supported criminalizing homosexuality, defending the anti-gay paragraph 175 of the German criminal code.[35] Although Alan Turing was pardoned in 2013 and paragraph 175 of the German criminal code has since been abolished, the available data suggests it will take much more effort before diversity is a reality for this community in science.[36]

The same applies to the role women play in science. In the Netherlands, for instance, women accounted for only 24% of professors in 2018.[37] As a result, current research is shaped by heteronormative standards that tend to overlook essential elements that may affect women more negatively. For instance, not considering

[28] Custers 2013.

[29] O'Neil 2016; Hamidi et al. 2018.

[30] *See* https://www.ohchr.org/en/issues/women/wrgs/pages/genderstereotypes.aspx.

[31] Sink et al. 2018.

[32] Burdge 2007; Howansky et al. 2019.

[33] Freire et al. 2020.

[34] Cech and Waidzunas 2021; Tao 2018.

[35] Whisnant 2012.

[36] Gibney 2019.

[37] Rathenau Institute 2021.

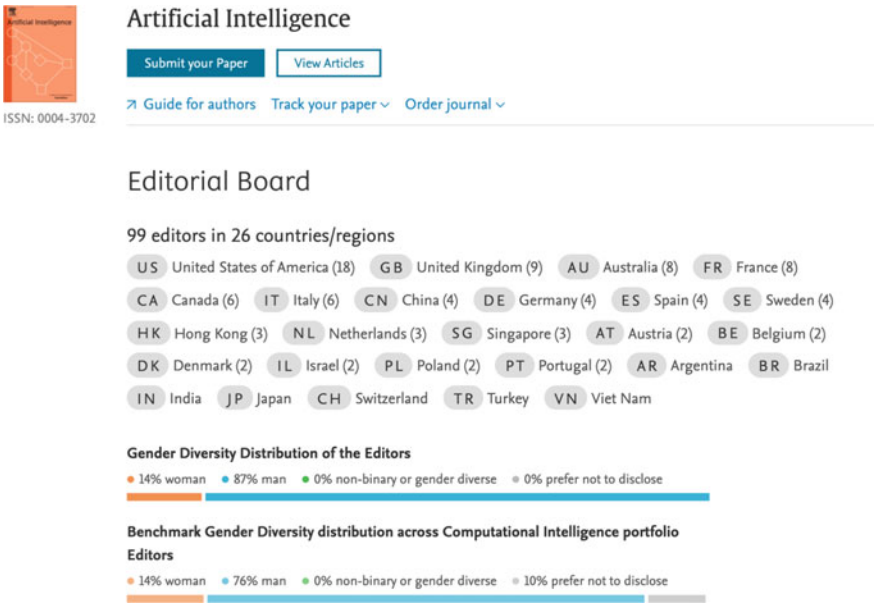

Fig. 6.1 Screenshot taken 15 September 2021, of the Artificial Intelligence Journal (AIJ) editorial board webpage[38]

gender and diversity issues in automotive engineering can lead to more significant injuries in accidents; or in biomedical research, failing to use female cells and tissues can pose more health risks to women.[39] The lack of diversity and inclusion in AI practices, ranging from datasets that represent only a portion of broader society, binary algorithms, and structural and systematic bias in the AI scientific community prevents the understanding of how these systems affect a big part of society and puts vulnerable communities at risk.[40] A more inclusive and diverse workforce could on the contrary promote the exploration of questions and the addressing of problems beyond the narrow slice of humanity that much science currently represents.[41]

6.2.3 Target User Level

Gender and power relations mediate the development of technology and technologies also impact our understanding of gender and human-technology relations,[42] which

[38] *See* https://www.journals.elsevier.com/artificial-intelligence/editorial-board.
[39] Schiebinger 2014.
[40] Poulsen et al. 2020.
[41] Nature Editorial 2018.
[42] Haraway 2006; Bray 2007; Wajcman 2007.

often goes beyond the male-female binary understanding.[43] Unfortunately, other attributes such as sexuality are often not taken into consideration in the development of technology (see, e.g., O'Riordan and Phillips[44]). Users of technology, however, constitute an extensive entanglement of social constructions, relations, and practices with technology because they "consume, modify, domesticate, design, reconfigure, and resist technological development".[45]

When framing technology in a traditional white straight male hegemony seen throughout science, technology, engineering and mathematics (STEM),[46] where inclusion reduces to binary mainstreaming strategies (e.g., the quantitative counting of some women/men),[47] one risks different forms of exclusion. Oudshoorn et al. 2004 warn how 'configuring the user as "everybody"' runs the risk of making it work for the majority, while effectively excluding minorities.[48] For instance, women, senior citizens, persons with disabilities, and the LGBTQ+ communities have not been traditionally considered in society as equal to straight men. As a consequence, there has been much technological development without taking into consideration these communities. For instance, bicycles are designed without taking into consideration women's bodies, resulting in back seat pain, and female sex toys are often created by men.[49]

Persons with disabilities have also been historically marginalized throughout technological progress, usually preventing them systematically from enjoying the same opportunities and resources as the abled population. For instance, consider the slow progress to remove some architectural barriers for people with disabilities, which remains a problem in some nations.[50] Still, much of the work investigating bias in AI centres on the racial and gender discriminatory power these systems have but does not consider how algorithmic systems may also affect the disabled communities.[51] Various are the examples in which technology has been developed for a specific community with a disability but without really engaging with it, e.g., in the case of deaf populations.[52] In this way, science and technology fields have failed to observe a key principle amongst the disabled community and its advocates about participation and co-design: 'nothing about us without us'.[53] In the case of sexual rights, although there have been international efforts towards realizing their sexual rights from institutions like the United Nations,[54] after nearly 30 years of discussion, this

[43] Faulkner 2001.
[44] O'Riordan and Phillips 2007.
[45] Oudshoorn and Pitch 2003.
[46] Page 2009.
[47] Vida 2020.
[48] Oudshoorn et al. 2004.
[49] MoMa 2021.
[50] Moscoso-Porras 2019.
[51] Whittaker et al. 2019.
[52] Bragg et al. 2019.
[53] Goggin and Newell 2003.
[54] United Nations 1993.

remains an unfinished agenda for the disabled[55] as if they failed to recognize people with disabilities as sexual beings.[56] In this respect, technology that could empower persons with disabilities to engage with their sexual rights is not mainstream and disregarded as an opportunity.[57]

If AI applications disregard the LGBTQ+ community, this is not necessarily a deliberate decision, but it could be due to lack of visibility of this community comparable to the inequitable experiences LGBTQ+ researchers in STEM fields.[58] One study points out significant biases that context-less online content moderation with AI-driven technology can have concerning online content produced by the LGBTQ+ community. The authors show that Perspective, an AI-driven tool used to measure the toxicity levels of text developed by Google's Jigsaw, could potentially impact drag queens and the LGBTQ+ community online.[59] After analysing several drag queen Twitter accounts, results show that the content produced by drag queens is flagged as having higher levels of toxicity than typically contentious profiles, such as Donald Trump and white supremacists. By failing to understand the LGBTQ+ perspective, the online moderator tool fails to discern that some members of the LGBTQ+ community reclaim derogatory language aimed at this community in a socially valuable way. Whether this is a result of sample bias, that is using a limited dataset to train the system which lacks LGBTQ+ voices, or due to exclusion bias, that is due to human-made decisions that tag certain content in the training dataset as derogatory at the time of data labelling, this leads to prejudicial and algorithmic bias. These biases unfairly alienate an already vulnerable community further and entrench rigid social expectations into systems. Advancing diversity and inclusion in AI could be a step towards creating practices and systems that are informed by the social context in which they occur, and not informed by context-blind training datasets.

6.3 Implications of Missing Diversity and Inclusion in AI

As the primary stakeholders and direct casualties of biased AI systems, such as drag queens using online social media being banned for language use in the example above on how AI-powered content moderator tools may silence the LGBT community, target users are most at risk of being affected by the lack of diversity and inclusion in AI. To show the broader implications of missing diversity and inclusion in AI, this section highlights the typically overlooked target user groups affected: women, the LGBTQ+ community, senior citizens, and disabled persons. Furthermore, this is done by framing and examining these implications in four AI application case studies:

[55] Temmerman et al. 2014.

[56] Maxwell et al. 2006; Roussel 2013.

[57] Fosch-Villaronga and Poulsen 2021.

[58] Cech and Waidzunas 2021.

[59] Gomes et al. 2019.

Sect. 6.3.1 gendered social robots: the mechanization of women, Sect. 6.3.2 Binary gender classifiers: guessing objectively what is subjective; Sect. 6.3.3 Algorithms for medical applications: gender as a safety parameter; and Sect. 6.3.4 Sex robotics: able-bodied and male-dominated markets.

6.3.1 Gendered Social Robots: The Mechanization of Women

Social robotics research does not fail to account for gender and sex considerations entirely. Instead, it is one-sided, with women being the primary target for objectification in social robotics. The ideal Stepford-wife image and social role, which presents women as subservient, dutiful, and pleasant, is commonplace in social robotics.[60] For instance, the digital social robot Azuma Hikari[61] presents a stereotypical image of women emerging from the Japanese social context, exacerbating existing prejudice towards women.[62] The developer's website describes Azuma Hikari as 'your personal bride' and in traditional, stereotypically feminine language, such as soothing and hard-working, closing with a quote from Azuma Hikari: 'I look forward to living with you, master!'[63] This approach to social robotics perpetuates a biased representation of women as having to be "young, sexy, soothing, and hard-working in housework" in service of a male master-like husband user.[64]

The stereotype of perfect womanhood in social robots[65] is observed globally. The service robot Sona 2.5,[66] developed in India in response to the COVID-19 pandemic and used in hospitals for food delivery, appears to have breasts and be wearing a skirt, neither of which fulfils any task (see Fig. 6.2). Instead, these aspects are an aesthetic design choice, ultimately reinforcing the biased view that caregiving is a woman's role. Other social robots, including Xiaoice,[67] Siri,[68] and Google Assistant,[69] all come with female voices out of the box.[70] Social robotics needs a 'feminist reboot'[71] at least and, at best, wider and fairer stakeholder engagement to ensure diversity and inclusion.

---

[60] Strengers and Kennedy 2020.

[61] *See* https://www.gatebox.ai/en/hikari.

[62] Liu 2021.

[63] *See* https://www.gatebox.ai/en/hikari.

[64] Liu 2021.

[65] Giger et al. 2019.

[66] *See* https://clubfirst.org/product/sona-2-5-covid-19-robot/.

[67] *See* http://www.xiaoice.com/.

[68] *See* https://www.apple.com/au/siri/.

[69] *See* https://assistant.google.com/.

[70] Liu 2021.

[71] Strengers and Kennedy 2020.

Fig. 6.2. Service robot Sona 2.5, with breasts and a skirt. Screenshot of a video uploaded to YouTube by India Times[72]

6.3.2 Binary Gender Classifiers: Guessing Objectively What is Subjective

Automated Sensitive Traits Recognition alludes to the use of inference classification systems that are, in part, trained to look for sensitive traits that stereotypically identify users as a certain type of person. One of the traits these systems infer is gender via GCS technology which attempts to identify and compare elements in novel input (e.g., word usage or images) to known data labelled by gender (e.g., stereotypical feminine or masculine words or imagery) and classify it by gender. These systems exacerbate existing stereotypes because they take 'sex' as a parameter. In this sense, these technologies usually build on 'male' and 'female' categories that exclude the intersex community.[73] For instance, a study by Park and Woo 2019 trained a system to identify women using a dataset that paired gender with the frequency of sentiment-driven words.[74] During training, the system learned that content produced by women tended to use the words 'thank', 'bless, 'scary,' and 'illness' about twice as often as men. At the same time, men used the words 'accurate,' 'important,' 'issue,' and 'aches' twice as often as women.[75] The algorithmically produced assumption that a person with sensitive traits, such as one who might more frequently use words like 'thank' and

[72] *See* India Times 2020 Covid-19: Jaipur Hospital Turns To Robots To Take Care Of Coronavirus Patients https://navbharattimes.indiatimes.com/video/news/covid-19-jaipur-hospital-turns-to-robots-to-take-care-of-coronavirus-patients/videoshow/74818092.cms.

[73] Fosch-Villaronga et al. 2021.

[74] Park and Woo 2019.

[75] Park and Woo 2019.

'bless,' is probably a woman, perpetuates stereotypical feminine-masculine societal roles which do not fully represent society.

Popular training datasets for GCS technology are significantly gender-biased, associating female names more often with family words than career words and with arts more than mathematics and science.[76] As a result, 'models trained to perform prediction on these datasets amplify the existing gender bias when evaluated on development data.'[77] For example, the verb 'cooking' is heavily biased towards women in a classifier trained using the imSitu dataset, amplifying existing gender stereotypes.[78] The same gender biases have been shown in natural language processing,[79] another method used to support gender classifiers.[80]

Algorithms perform poorly in recognizing objectively internal and subjective aspects tied to social and cultural factors, including gender and emotions.[81] Given that biases can propagate throughout AI models,[82] these systems may misclassify users. In the context of GCS, these systems can misgender users, which has adverse implications that go from reinforcing gender binarism to undermining autonomy. Also they can be a tool for surveillance that can threaten someone's safety.[83] To be misgendered reinforces the idea that society does not consider or recognize a person's gender as real, causing rejection, impacting self-esteem and confidence, feeling authenticity, and increasing one's perception of being socially stigmatized.[84]

However, the main problem is that gender identity is primarily subjective and internal, which completely opposes the idea that gender can be recognized automatically, at least with state-of-the-art GCS technology.[85] The same applies to emotional recognition systems aimed at recognizing user emotions: emotional AI follows a procrustean design, in which emotions are reduced to physiological parameters only.[86] In this line of thought, it is not hard to imagine that misclassifications can occur. If used to support ulterior decision-making processes, such misclassification may lead to adverse effects for the users, ranging from mere discomfort to a chilling effect or even harm.[87]

---

[76] Nosek et al. 2002a; Caliskan et al. 2017; Nosek et al. 2002b.

[77] Zhao et al. 2017.

[78] Zhao et al. 2017.

[79] Sun et al. 2019; Zhou et al. 2019.

[80] Campa et al. 2019.

[81] Dupré et al. 2020; Fosch-Villaronga et al. 2021.

[82] Buolamwini and Gebru 2018; Font and Costa-jussà 2019; McDuff et al. 2019; Torralba and Efros 2011.

[83] Hamidi et al. 2018.

[84] Keyes 2018; Fosch-Villaronga et al. 2021.

[85] Fosch-Villaronga et al. 2021.

[86] Fosch-Villaronga 2019a, b.

[87] Hamidi et al. 2018; Büchi et al. 2020; Nišević et al. 2021.

6.3.3 *Algorithms for Medical Applications: Gender as a Safety Parameter*

If failing to account for sex and gender considerations in algorithmic systems is a point of concern in AI-driven social media practices (e.g., using GCS technology), failing to do so in sensitive domain applications like healthcare, where these considerations are essential in determining patient safety and healthcare outcomes, is a salient concern. Despite clear evidence to the contrary, science holds on to the promise that these systems will help deliver safer care.[88,89]

The persistent phenomena of failing to support diversity and inclusion has especially gained ground in the context of rising inequities and bias in healthcare today, which does not provide adequate care for all, explicitly excluding minority groups in society like the transgender and the intersex communities. Intertwined with this concern of exacerbating pre-existing inequities, including gender inequalities, is embedded bias present in many algorithms due to the lack of inclusion of minorities in datasets.[90] For example, AI used in dermatology to diagnose melanoma lacks the inclusion of skin colour.[91] Another example is the corpus of genomic data, which so far has seriously underrepresented minorities.[92] In the context of AI for medicine, such crucial differences in sex and gender can be vital when it comes to critical conditions and directly impact patient safety.

These findings indicate that much work is still needed in the area of diversity in AI for medicine to eradicate embedded prejudice in AI and strive for medical research that provides a true representative cross-section of the population.[93] Algorithms should be designed to look at specific features from an intersectional point of view, like gender as a non-binary characteristic, which may prevent discrimination for this community. Also, developers should only use sensitive information relating to gender, sex, or race in specific and regulated applications where it is proven they matter.[94] On the contrary, and as far as possible, AI could also use gender-neutral biomarkers for decision-making, a practice that could be more in line with the data minimization principle enshrined in EU data protection law. Alternatively, developers could design discrimination-aware or privacy-preserving algorithms, also in the context of medicine.[95] In this way, biases could be eliminated from the data used to train the AI and ensure an equal representation of examples.

---

[88] Yu et al. 2018; Ahuja 2019.

[89] Cirillo et al. 2020.

[90] Topol 2019.

[91] Esteva et al. 2017.

[92] Wapner 2018.

[93] Topol 2019.

[94] Fosch-Villaronga et al. 2021.

[95] Kamiran et al. 2013; Cirillo et al. 2020.

6.3.4 Sex Robotics: Able-Bodied and Male-Dominated Markets

Sex robots are service robots that perform actions contributing directly to increase in the satisfaction of the sexual needs of a user.[96] These robots often target young, able-bodied, and typically straight men, both in the way they are marketed and designed.[97] Given the widespread views and narratives concerning sex and sexuality, sex robots are commonly not targeted to people with disabilities, a group which might benefit the most from sex robot intervention to help fulfil unmet sexual needs.[98]

Through a broader lens, the lack of wider inclusion of different user groups leads sex robotics to have intimate connections with misogyny, child sexual exploitation, male violence, and the idea that women are programmable.[99] The results of a systematic exploratory survey on public opinion on sex robots reveal that, in general, men find sex robots more acceptable than women.[100] On the expected capabilities of sex robots, the statistics also show that women, more than men, prefer robots to be instructed and obey orders.[101] This may suggest that sex robots increase the objectification of the person, regardless of gender, and that, more research is needed to understand how the interplay between diversity and inclusion could affect sex robot development.[102]

Engaging with diversity and inclusion could help open new avenues for the sex robot industry and potentially help create counternarratives that favour new developments in this area. For instance, sex robots have many sexual characteristics and capabilities that might prove helpful in fulfilling the sexual desires of those in disability care. However, for the most part, sex robotics research excludes persons with disabilities as crucial stakeholders.[103] Also, some research on sex robots addresses sex offenders as users, exploring these artefacts' use to reduce poor sexual behaviour.[104] Yet, studies show that sex offenders are less likely to perceive sex robots as adequate deterrents for sexual violence against persons.[105] Hence, not engaging with the right communities more inclusively may create wrong and inconsiderate technology and prevent parts of the population from enjoying the benefits technology offers.

---

[96] Fosch-Villaronga and Poulsen 2021.
[97] Fosch-Villaronga and Poulsen 2021.
[98] Jecker 2020.
[99] Richardson 2016.
[100] Scheutz and Arnold 2016.
[101] Scheutz and Arnold 2016.
[102] Fosch-Villaronga and Poulsen 2021.
[103] Fosch-Villaronga and Poulsen 2021.
[104] Behrendt 2018.
[105] Zara et al. 2021.

6.4 Addressing Diversity and Inclusion in AI

Mitigation strategies are needed to account for the implications of sex and gender considerations in AI, such as those explored above. This section proposes three holistic approaches to advance diversity and inclusion in AI that align with current legislation and are more attuned to societal needs. These are gendering algorithms, more diverse design teams, and more inclusive and explicit guiding policies.

6.4.1 Diversity in Algorithms: Gendering Algorithms

At the technical level, data collection practices could be more diverse and inclusive. For instance, consider the use of AI in medicine via clinical decision support algorithms, which are trained using large datasets of electronic health records. These datasets may contain an unbalanced representation of sex and gender factors, resulting in algorithmic bias emerging during training.[106] Ultimately, considering the impact of sex and gender on human health (e.g., through opportunities for therapeutic discovery and the frequency and magnitude of adverse health events), missing these considerations in AI-driven medicine is of concern.[107] In this case, a push towards more diverse and inclusive AI practices could be to make an effort towards reducing and eliminating biases from datasets by ensuring there is an equal representation of sex and gender differences.[108] The same practice could be used in GCS research to the same end. Furthermore, similar advances towards equal representation in the realisation of others systems could also reduce bias and improve diversity, particularly in social and sex robotics through fairer stakeholder engagement.

The exclusion of diverse gender and sex considerations in AI puts vulnerable communities at risk. Digital identity and participatory culture play a significant role in the sense of self in the modern world and there could be more efforts to realize diversity and inclusion in the online world[109] to not perpetuate the normative view that particular groups of people, such as trans or non-binary people, do not exist.[110] For instance, gender classifiers could be developed using a more accurate understanding of gender to represent contemporary society fully. For instance, algorithms can be designed to look at certain features from an intersectional point of view, like gender as a non-binary characteristic. As far as possible, gender-neutral biomarkers could also be used by AI for decision-making. In this way, biases can be eliminated from the data used to train the AI by ensuring there is an equal representation of examples, and diversity can be better accounted for[111] Having a GCS that accounts for diversity

[106] Cirillo et al. 2020.
[107] McGregor 2016.
[108] Cirillo et al. 2020.
[109] Jenkins et al. 2016.
[110] Keyes 2018.
[111] Kamiran et al. 2013.

and inclusion would help reduce bias in systems in which gender inferences flow, including search and recommendation systems, which similarly need to be fairness-aware (i.e., data handling is guided by ethical, social, and legal dimensions).[112]

6.4.2 Diverse Teams, Organizations, and Design

Accounting for stakeholder values, promoting positive value impact, and eliminating and mitigating adverse effects requires teams designing, developing, and implementing AI to have diverse configurations, administration, and design thinking. Diverse groups have more accurate discussions, cite more facts, make fewer mistakes, and are more willing to discuss sensitive topics such as racism.[113] Diverse teams also contribute to radical innovation processes[114] and although they are less confident and perceive group interactions as less effective, they perform better than more homogeneous groups.[115] In short, people from diverse backgrounds can help improve group thinking. Given that AI can affect individuals and society at large, thinking of ways to increase diversity in the teams building AI systems can prove beneficial in the long term.

To avoid replicating bias in AI, considering the values of vulnerable communities, such as people with disabilities, the LGBTQ+ community, or women, is crucial. Participatory, user-centred design methods that centre on diverse human values and include the voice of the user in the realization of an artefact, such as value sensitive design,[116] are the best way forward to account for diversity and inclusion in AI. Furthermore, adopting holistic inclusion strategies and diverse teams in robotics and AI could ease the understanding of the challenges around discrimination and bias experienced by vulnerable communities.[117] Noteworthy outlier initiatives which have embraced these approaches and are pushing for diversity and inclusion in robotics and AI include Pride@CSIRO[118] and Queer in AI.[119] Both these initiatives seek to foster inclusive environments in AI research, recruit diverse and talented people, and engage grand technology challenges with a diversity of the minds and lived experiences.

Digital identity and participatory culture play a determinant role in the sense of self in the modern world. In this sense, there could be more efforts towards realizing diversity and inclusion in the online world[120] to not perpetuate the normative

[112] Geyik et al. 2019.

[113] Sommers 2006; Rock and Grant 2016.

[114] Díaz-García et al. 2013.

[115] Phillips et al. 2009.

[116] Friedman and Hendry 2019; Friedman et al. 2006.

[117] Poulsen et al. 2020.

[118] CSIRO 2019.

[119] Queer in AI 2019.

[120] Jenkins 2016.

view that certain collectives such as trans or non-binary do not exist.[121] Holistic inclusion strategies on multiple levels, e.g., how these communities and the individuals can benefit from robot technology, could combat this issue. More research is needed to create knowledge about how different communities, such as women, LGBTQ+, and persons with discapacities, engage with and value technologies to identify how to better include them in all levels of the design, creation, and implementation process. An essential recommendation is that these users are included thoroughly in the design-implementation-use lifecycle of AI through participatory, user-centred design methods, such as value sensitive design,[122] as these could positively or adversely impact these user groups' lives.

6.4.3 More Inclusive Guidelines, Policies, and Regulation

Designers play a significant role in shaping technology to meet the needs of users and the goals of regulators.[123] However, robot developers are not always in a position to foresee the potential risks that their creations may have because they are usually too intent on solving a particular problem. Users may also be more concerned with the practical benefits that they gain from employing the technology than reflecting on whether it is beneficial for them or not.[124] Branching across and above the technical, community, and target user levels, policy operates on a meta-level that could help strengthen diversity and inclusion in AI throughout the other levels. AI designers, for instance, need to respect the EU Charter of Fundamental Rights (EU CFR), including its Articles 1 on dignity, 7–8 on private life and protection of personal data, 21 on non- discrimination, 23 on equality between women and men. There are also two international human rights treaties that include explicit obligations relating to harmful and wrongful stereotyping. These articles translate into direct obligations for AI designers to develop systems that are safe, respect user privacy, do not discriminate, and do not generate or reinforce stereotypes.

Still, AI developers may struggle to implement these human rights in their teams or their designs because the current legal framework is fragmented, lacks concrete guidance, and strives to account for diversity and inclusion.[125] For instance, sex and gender considerations have not been traditionally considered sensitive or essential aspects in related EU legal frameworks, such as the General Data Protection Directive (GDPR), the Medical Device Regulation, or the Safety Machinery Directive.[126]

---

[121] Keyes 2018.

[122] Friedman and Hendry 2019; Friedman et al. 2006.

[123] Fosch-Villaronga and Özcan 2019.

[124] Carr 2011.

[125] Jobin et al. 2019.

[126] Martinetti et al. 2021. *See* Regulation (EU) of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regards to the processing of personal data and on the free movement of such data; Regulation (EU) 2017/745 of the European Parliament and

In April 2021, the European institutions released a proposal for a regulation laying down harmonized rules on AI (also called draft AI Act).[127] The draft AI Act bases its wording on the idea that AI designers need to respect the principles of data protection, consumer protection, non-discrimination and gender equality. The draft AI Act complements existing EU law on non-discrimination that lays down specific requirements to minimize the risk of algorithmic discrimination, especially concerning the design and the quality of data sets used for AI systems and the obligations for testing, risk management, documentation, and human oversight throughout the AI systems' lifecycle.

The draft AI Act also identifies those 'AI systems that pose significant risks to the health and safety or fundamental rights of persons' as 'high-risk'. However, the AI Act does not have an intersectional approach for algorithmic discrimination of certain groups. Gender equality is only mentioned once, and although it is clear that the AI Act stresses that algorithms can discriminate against age groups, persons with disabilities, or persons of specific racial or ethnic origins or sexual orientation, this is in the context of work-related matters. However, failing to acknowledge that algorithms and AI can discriminate against society in general, including women, senior citizens, persons with disabilities, the LGBTQ+ community, or communities from different religions, is failing society. A more inclusive, diverse, and intersectional approach to AI regulation is deemed necessary if the EU expects to ensure that AI is of, by, and for the people.

Amidst this regulatory turmoil, the notion of responsible research and innovation (RRI) has emerged as an overarching concept that captures crucial aspects concerning what researchers can do to ensure that science, research, and innovation have positive, socially acceptable, and desirable outcomes.[128] The RRI approach provides a suitable framework to guide all the social actors involved in research and innovation (R&I) processes towards this aim. The European Commission defines RRI as "an approach that anticipates and assesses potential implications and societal expectations concerning research and innovation, intending to foster the design of inclusive and sustainable research and innovation."[129] Through the lens of RRI, the principles of inclusion, anticipation, reflection, and responsiveness typically guide the research and innovation (R&I) processes and could prove to be instrumental in achieving more inclusive and diverse AI—at least in transition times.

---

of the Council of 5 April 2017 on medical devices; and the Directive 2006/42/EC of the European Parliament and of the Council of 17 May 2006 on machinery.

[127] AI Act 2021.

[128] Stahl and Coeckelbergh 2016.

[129] European Commission 2012.

6.5 Conclusion

Social inequalities recreated as AI bias result from the lack of diversity and inclusion in AI practices. For instance, by failing to account for the socially valuable use of LGBTQ+ speech aiming to reclaim derogatory language within that community, AI-driven content moderator tools automatically flag online posts of drag queens using reclaimed language as *toxic* and prevent them from freely communicating online.[130] These kinds of biases emerge from a range of inequities preserved in AI practices, from limited datasets that do not fully represent society[131] to structural and systematic biased configurations of the AI scientific community.[132] At risk is the amplification of stereotypes, alienation of minority and silent communities, and entrenchment of rigid social expectations in systems.[133]

Although there is increasing attention from robotics, the Human-Robot Interaction and AI communities to address diversity, particularly biased and discriminatory algorithms,[134] biases persist, and vulnerable communities remain mainly invisible and at risk.[135] This calls for action toward the redefinition of inclusion and exclusion, the boundaries and limitations of diversity for the robotics and AI community.[136] Advancing diversity and inclusion in AI, therefore, could be a step towards creating practices and system output that are informed by the social context in which they occur, and not informed by a select few in a research laboratory or by context-blind trained systems.[137]

[130] Raji and Buolamwini 2019; Gomes et al. 2019.

[131] Zhao et al. 2017.

[132] Roopaei et al. 2021.

[133] Mitchell et al. 2020.

[134] Raji and Buolamwini 2019.

[135] Willson 2017; Noble 2018; Ito 2019.

[136] Some initiatives have started to explore these topics in the Netherlands. Check for instance the 'Gendering Algorithms' initiative started at Leiden University (*see* https://www.genderingalgorithms.org/) or the 'Diversity and Inclusion for Embodied AI' initiative started by the 4TU Federation and Leiden University (*see* https://www.dei4eai.com/).

[137] Mitchell et al. 2020.

References

Addlakha R et al (2017) Disability and sexuality: Claiming sexual and reproductive rights. Reproductive Health Matters https://doi.org/10.1080/09688080.2017.1336375

Ahuja A S (2019) The impact of artificial intelligence in medicine on the future role of the physician. Peer J, 7, e7702

Behrendt M (2018) Reflections on moral challenges posed by a therapeutic childlike sexbot. In: Cheok A, Levy D (eds) LSR 2017: Love and Sex with Robots. Springer, Cham, pp 96–113

Bragg D et al (2019) Sign language recognition, generation, and translation: An interdisciplinary perspective. In: Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility. ACM, New York, pp 16–31

Bray F (2007) Gender and technology. Annu. Rev. Anthropol. https://doi.org/10.1146/annurev.anthro.36.081406.094328

Büchi M, Fosch-Villaronga E, Lutz C, Tamò-Larrieux A, Velidi S, Viljoen S (2020) The chilling effects of algorithmic profiling: Mapping the issues. Computer law & security review 36, 105367

Burdge B J (2007) Bending gender, ending gender: Theoretical foundations for social work practice with the transgender community. Social work 52:243–250

Buolamwini J, Gebru T (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Proceedings of the First Conference on Fairness, Accountability and Transparency. PMLR, pp 77–91

Caliskan A et al (2017) Semantics derived automatically from language corpora contain humanlike biases. Science https://doi.org/10.1126/science.aal4230

Campa S et al (2019) Deep & machine learning approaches to analyzing gender representations in journalism. https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/reports/custom/15787612.pdf

Carr N (2011) The Shallows: What the Internet is doing to our brains

Cech E A, Waidzunas T J (2021) Systemic inequalities for LGBTQ professionals in STEM. Science Advanceshttps://doi.org/10.1126/sciadv.abe0933

Cirillo D et al (2020) Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. NPJ Digital Medicine https://doi.org/10.1038/s41746-020-0288-5

Commonwealth Scientific and Industrial Research Organisation (CSIRO) (2019) Diversity & inclusion at the robotics and autonomous systems group. https://research.csiro.au/robotics/diversity-inclusion-at-the-robotics-and-autonomous-systems-group/

Custers B (2013) Data dilemmas in the information society: Introduction and overview. In: Custers B et al (eds) Discrimination and Privacy in the Information Society. Springer, Berlin, pp 3–26

Danielescu A (2020) Eschewing gender stereotypes in voice assistants to promote inclusion. In: Torres M I et al (eds) Proceedings of the 2nd Conference on Conversational User Interfaces. ACM, New York, pp 1–3

Di Nucci E (2017) Sex robots and the rights of the disabled. In: Danaher J, McArthur N (eds) Robot Sex: Social and Ethical Implications. MIT Press, Cambridge, pp 73–88

Díaz-García C, González-Moreno A, Saez-Martinez FJ (2013) Gender diversity within R&D teams: Its impact on radicalness of innovation. Innovation, 15(2), pp. 149–160

Döring N et al (2020) Design, use, and effects of sex dolls and sex robots: Scoping review. Journal of Medical Internet Research https://doi.org/10.2196/18551

Dupré D, Krumhuber EG, Küster D, McKeown GJ (2020) A performance comparison of eight commercially available automatic classifiers for facial affect recognition. PloS one 15(4):e0231968

Ekmekçioğlu O et al (2021) Women in nuclear medicine. Eur. J. Nucl. Med. Mol. Imaging https://doi.org/10.1007/s00259-021-05418-9

European Commission (2012) Options for strengthening responsible research & innovation. Retrieved from https://ec.europa.eu/research/science-society/document_library/pdf_06/options-for-strengthening_en.pdf

Esteva A et al (2017) Dermatologist-level classification of skin cancer with deep neural networks. Naturehttps://doi.org/10.1038/nature21056
Faulkner W (2001) The technology question in feminism: A view from feminist technology studies. Women's Studies International Forum https://doi.org/10.1016/S0277-5395(00)00166-7
Font J E, Costa-jussà M R (2019) Equalizing gender bias in neural machine translation with word embeddings techniques. In: Costa-jussà M R et al (eds) Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing. Association for Computational Linguistics, Stroudsburg, pp 147–154
Fosch-Villaronga E (2019a) Robots, healthcare, and the law: Regulating automation in personal care. Routledge, Abingdon
Fosch-Villaronga E (2019b) "I love you," said the robot: Boundaries of the use of emotions in human-robot interactions. In: Ayanoğlu H, Duarte E (eds) Emotional design in human-robot interaction. Springer, Cham, pp 93–110
Fosch-Villaronga E, Özcan B (2020) The progressive intertwinement between design, human needs and the regulation of care technology: the case of lower-limb exoskeletons. International Journal of Social Robotics, 12(4), 959–972
Fosch-Villaronga E, Poulsen A (2020) Sex care robots. Paladyn, Journal of Behavioral Robotics https://doi.org/10.1515/pjbr-2020-0001
Fosch-Villaronga E, Poulsen A (2021) Sex robots in care: Setting the stage for a discussion on the potential use of sexual robot technologies for persons with disabilities. In: Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction. ACM, New York, pp 1–9
Fosch-Villaronga E et al (2021) A little bird told me your gender: Gender inferences in social media. Information Processing & Management https://doi.org/10.1016/j.ipm.2021.102541
Freire A et al (2020) Measuring diversity of artificial intelligence conferences. arXiv preprint. https://arxiv.org/abs/2001.07038
Friedman B, Hendry D G (2019) Value sensitive design: Shaping technology with moral imagination. MIT Press, Cambridge
Friedman B et al (2006) Value sensitive design and information systems. In: Zhang P, Galletta D (eds) Human-computer interaction and management information systems: Foundations. M. E. Sharpe, New York, pp 348–372
Gartrell A et al (2017) "We do not dare to love": Women with disabilities' sexual and reproductive health and rights in rural Cambodia. Reproductive Health Matters https://doi.org/10.1080/09688080.2017.1332447
Geyik S C et al (2019) Fairness-aware ranking in search & recommendation systems with application to LinkedIn talent search. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, New York, pp 2221–2231
Gibney E (2019) Discrimination drives LGBT+ scientists to think about quitting. Nature. https://www.nature.com/articles/d41586-019-02013-9
Giger J-C et al (2019) Humanization of robots: Is it really such a good idea? Hum. Behav. & Emerg. Tech. https://doi.org/10.1002/hbe2.147
Goggin G, Newell C (2003) Digital disability: The social construction of disability in new media. Rowman & Littlefield, Lanham
Groom J R (2021) Diversity in science requires mentoring for all, by all. Nat. Immunol. https://doi.org/10.1038/s41590-021-00999-x
Gomes A et al (2019) Drag queens and artificial intelligence: Should computers decide what is 'toxic' on the internet? Internet Lab. http://www.internetlab.org.br/en/freedom-of-expression/drag-queens-and-artificial-intelligence-should-computers-decide-what-is-toxic-on-the-internet/
Hamidi F et al (2018) Gender recognition or gender reductionism? The social implications of embedded gender recognition systems. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. ACM, New York, pp 1–3

Hao K (2019) Facebook's ad-serving algorithm discriminates by gender and race. MIT Technology Review. https://www.technologyreview.com/2019/04/05/1175/facebook-algorithm-discriminates-ai-bias/

Haraway D (2006) A cyborg manifesto: Science, technology, and socialist-feminism in the late 20th century. In: Weiss J et al (eds) The International Handbook of Virtual Learning Environments. Springer, Dordrecht, pp 118–158

Higgins A et al (2006) Sexual health education for people with mental health problems: What can we learn from the literature? Journal of Psychiatric and Mental Health Nursing https://doi.org/10.1111/j.1365-2850.2006.01016.x

Holder C et al (2016) Robotics and law: Key legal and regulatory implications of the robotics age (part II of II). Computer Law & Security Review https://doi.org/10.1016/j.clsr.2016.05.011

Howansky K et al (2021) (Trans)gender stereotypes and the self: Content and consequences of gender identity stereotypes. Self and Identity https://doi.org/10.1080/15298868.2019.1617191

International Federation of Robotics (2018) Executive summary world robotics 2018 service robots. https://ifr.org/downloads/press2018/Executive_Summary_WR_Service_Robots_2018.pdf

Ito J (2019) Supposedly 'fair' algorithms can perpetuate discrimination. MIT Media Lab. https://www.media.mit.edu/articles/supposedly-fair-algorithms-can-perpetuate-discrimination/

Jecker N S (2020) Nothing to be ashamed of: Sex robots for older adults with disabilities. Journal of Medical Ethics https://doi.org/10.1136/medethics-2020-106645

Jenkins H et al (2016) Participatory culture in a networked era: A conversation on youth, learning, commerce, and politics. Polity Press, Cambridge

Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. Nat Mach Intell 1(9):389–399

Kamiran F et al (2013) Techniques for discrimination-free predictive models. In: Custers B H M et al (eds) Discrimination and Privacy in the Information Society. Springer, Heidelberg, pp 223–239

Keyes O (2018) The misgendering machines: Trans/HCI implications of automatic gender recognition. Proceedings of the ACM on Human-Computer Interaction https://doi.org/10.1145/3274357

Liu J (2021) Social robots as the bride? Understanding the construction of gender in a Japanese social robot product. Human-Machine Communication https://doi.org/10.30658/hmc.2.5

Martinetti A, Chemweno PK, Nizamis K, Fosch-Villaronga E (2021) Redefining safety in light of human-robot interaction: A critical review of current standards and regulations. Front Chem Eng 32

Maxwell J et al (2006) A health handbook for women with disabilities. Hesperian, Berkeley

McCann E (2003) Exploring sexual and relationship possibilities for people with psychosis – A review of the literature. Journal of Psychiatric and Mental Health Nursing https://doi.org/10.1046/j.1365-2850.2003.00635.x

McDuff D et al (2019) Characterizing bias in classifiers using generative models. In: Wallach H et al (eds) Proceedings of the 33rd Conference on Neural Information Processing Systems. Curran Associates, New York, pp 1–12

McGregor A J et al (2016) How to study the impact of sex and gender in medical research: A review of resources. Biol. Sex Differ. https://doi.org/10.1186/s13293-016-0099-1

Mitchell M et al (2020) Diversity and inclusion metrics in subset selection. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. ACM, New York, pp 117–123

MoMa (2021) Design innovations for women. Design store. https://store.moma.org/design-innovations-for-women.html

Moscoso-Porras M et al (2019) Access barriers to medical facilities for people with physical disabilities: The case of Peru. Cadernos de Saúde Pública https://doi.org/10.1590/0102-311x00050417

Nature Editorial (2018) Science benefits from diversity. Nature, 558, 5–6, https://www.nature.com/articles/d41586-018-05326-3

Nišević M et al (2021) Understanding the legal bases for automated decision-making under the GDPR. In: Kostas E, Leenes R (eds) Research Handbook on EU Data Protection. Hart Publishing, Oxford [forthcoming]

Noble S U (2018) Algorithms of oppression: How search engines reinforce racism. NYU Press, New York

Nosek B A et al (2002a) Harvesting implicit group attitudes and beliefs from a demonstration web site. Group Dynamics: Theory, Research, and Practice https://doi.org/10.1037/1089-2699.6.1.101

Nosek B A et al (2002b) Math = male, me = female, therefore math ≠ me. Journal of Personality and Social Psychology https://doi.org/10.1037/0022-3514.83.1.44

Ntoutsi E et al (2020) Bias in data-driven artificial intelligence systems—An introductory survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery https://doi.org/10.1002/widm.1356

O'Neil C (2016) Weapons of math destruction: How big data increases inequality and threatens democracy. Crown, New York

O'Riordan K, Phillips D J (2007) Queer online: Media technology & sexuality. Peter Lang Publishing, Bern

Oudshoorn N, Pinch T (2003) How users matter: The co-construction of users and technology. MIT Press, Cambridge

Oudshoorn N et al (2004) Configuring the user as everybody: Gender and design cultures in information and communication technologies. Science, Technology, & Human Values https://doi.org/10.1177/0162243903259190

Page M et al (2009) The blue blazer club: masculine hegemony in science, technology, engineering, and math fields. Forum on Public Policy Online v2009:1–23

Park S, Woo J (2019) Gender classification using sentiment analysis and deep learning in a health web forum. Applied Sciences https://doi.org/10.3390/app9061249

Perry B L, Wright E R (2006) The sexual partnerships of people with serious mental illness. Journal of Sex Research https://doi.org/10.1080/00224490609552312

Phillips KW, Liljenquist KA, Neale MA (2009) Is the pain worth the gain? The advantages and liabilities of agreeing with socially distinct newcomers. Personality and Social Psychology Bulletin, 35(3), 336–350

Poulsen A et al (2020) Queering machines. Nature Machine Intelligence https://doi.org/10.1038/s42256-020-0157-6

Prince A E, Schwarcz D (2020) Proxy discrimination in the age of artificial intelligence and big data. Iowa Law Review 105:1257–1318

Quinn C, Browne G (2009) Sexuality of people living with a mental illness: A collaborative challenge for mental health nurses. International Journal of Mental Health Nursing https://doi.org/10.1111/j.1447-0349.2009.00598.x

Queer in AI (2019) Queer in AI. https://sites.google.com/view/queer-in-ai/

Rahman F, Billionniere E (2021) Re-entering computing through emerging technology: Current state and special issue introduction. ACM Trans. Comput. Educ. https://doi.org/10.1145/3446840

Raji I D, Buolamwini J (2019) Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. ACM, New York, pp 429–435

Rathenau Institute (2021) Women in Academia. https://www.rathenau.nl/en/science-figures/personnel/women-science/women-academia

Richardson K (2016) The asymmetrical 'relationship' parallels between prostitution and the development of sex robots. ACM SIGCAS Computers and Society https://doi.org/10.1145/2874239.2874281

Righetti L et al (2019) Unintended consequences of biased robotic and artificial intelligence systems [ethical, legal, and societal issues]. IEEE Robotics & Automation Magazine https://doi.org/10.1109/MRA.2019.2926996

Rock D, Grant H (2016) Why diverse teams are smarter. Harvard Business Review, 4(4), 2–5
Roopaei M et al (2021) Women in AI: barriers and solutions. In: Proceedings of the 2021 IEEE World AI IoT Congress (AIIoT). IEEE, New York, pp 0497-0503
Roussel S (2013) Seeking Sexual Surrogates. The New York Times. https://www.nytimes.com/video/world/europe/100000002304193/seeking-sexual-surrogates.html [video]
Schwalbe N, Wahl B (2020) Artificial intelligence and the future of global health. The Lancet https://doi.org/10.1016/S0140-6736(20)30226-9
Scheutz M, Arnold T (2016) Are we ready for sex robots? In: Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction. IEEE, New York, 351–358
Schiebinger L (2014) Scientific research must take gender into account. Nature 507, 9.https://doi.org/10.1038/507009a
Schönberger D (2019) Artificial intelligence in healthcare: A critical analysis of the legal and ethical implications. International Journal of Law and Information Technology https://doi.org/10.1093/ijlit/eaz004
Servais L (2006) Sexual health care in persons with intellectual disabilities. Mental Retardation and Developmental Disabilities Research Reviews https://doi.org/10.1002/mrdd.20093
Sink A, Mastro D, Dragojevic M (2018) Competent or warm? A stereotype content model approach to understanding perceptions of masculine and effeminate gay television characters. Journalism & Mass Communication Quarterly, 95(3), 588–606
Sommers SR (2006) On racial diversity and group decision making: identifying multiple effects of racial composition on jury deliberations. Journal of personality and social psychology, 90(4), 597
Søraa R A (2017) Mechanical genders: How do humans gender robots? Gender, Technology and Development https://doi.org/10.1080/09718524.2017.1385320
Sparrow R (2021) Sex robot fantasies. Journal of Medical Ethics https://doi.org/10.1136/medethics-2020-106932
Stahl BC, Coeckelbergh M (2016) Ethics of healthcare robotics: Towards responsible research and innovation. Robotics and Autonomous Systems, 86, 152–161
STOA (2018) Assistive technologies for people with disabilities. https://www.europarl.europa.eu/RegData/etudes/IDAN/2018/603218/EPRS_IDA(2018)603218_EN.pdf
Strengers Y, Kennedy J (2020) The smart wife: Why Siri, Alexa, and other smart home devices need a feminist reboot. MIT Press
Sun T et al (2019) Mitigating gender bias in natural language processing: Literature review. In: Korhonen A et al (eds) Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Stroudsburg, pp 1630–1640
Tannenbaum C, Ellis RP, Eyssel F, Zou J, Schiebinger L (2019) Sex and gender analysis improves science and engineering. Nature 575(7781):137–146
Tao Y (2018) Earnings of academic scientists and engineers: Intersectionality of gender and race/ethnicity effects. American Behavioral Scientist https://doi.org/10.1177/0002764218768870
Temmerman M et al (2014) Sexual and reproductive health and rights: A global development, health, and human rights priority. The Lancet https://doi.org/10.1016/S0140-6736(14)61190-9
Topol E J (2019) High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine https://doi.org/10.1038/s41591-018-0300-7
Torralba A, Efros A A (2011) Unbiased look at dataset bias. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, New York, pp 1521–1528
United Nations (1993) Standard rules on the equalization of opportunities for persons with disabilities. https://www.un.org/disabilities/documents/gadocs/standardrules.pdf
United Nations (2007) Convention on the Rights of Persons with Disabilities and Optional Protocol. https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities.html

Urry K, Chur-Hansen A (2020) Who decides when people can have sex? Australian mental health clinicians' perceptions of sexuality and autonomy. Journal of Health Psychology https://doi.org/10.1177/1359105318790026

Vaughan C et al (2015) W-DARE: A three-year program of participatory action research to improve the sexual and reproductive health of women with disabilities in the Philippines. BMC Public Health https://doi.org/10.1186/2Fs12889-015-2308-y

Vida B (2021) Policy framing and resistance: Gender mainstreaming in Horizon 2020. European Journal of Women's Studies https://doi.org/10.1177/1350506820935495

Wajcman J (2007) From women and technology to gendered technoscience. Information, Community and Society https://doi.org/10.1080/13691180701409770

Wapner J (2018) Cancer scientists have ignored African DNA in the search for cures. Newsweek. https://www.newsweek.com/2018/07/27/cancer-cure-genome-cancer-treatment-africa-genetic-charles-rotimi-dna-human-1024630.html

Weber J (2005) Helpless machines and true loving care givers: A feminist critique of recent trends in human-robot interaction. Journal of Information, Communication and Ethics in Society https://doi.org/10.1108/14779960580000274

West M et al (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

Willson M (2017) Algorithms (and the) everyday. Information, Communication & Society https://doi.org/10.1080/1369118X.2016.1200645

Wisskirchen G et al (2017) Artificial intelligence and robotics and their impact on the workplace. IBA Global Employment Institute

Wheeler A P, Steenbeek W (2021) Mapping the risk terrain for crime using machine learning. Journal of Quantitative Criminology https://doi.org/10.1007/s10940-020-09457-7

Whisnant C J (2012) Male homosexuality in West Germany. Palgrave Macmillan, London

Whittaker M et al (2019) Disability, bias, and AI. AI Now Institute. https://wecount.inclusivedesign.ca/uploads/Disability-bias-AI.pdf

World Health Organization (2015) Sexual health, human rights and the law report. https://apps.who.int/iris/bitstream/handle/10665/175556/9789241564984_eng.pdf

Yu KH, Beam AL, Kohane IS (2018) Artificial intelligence in healthcare. Nature biomedical engineering, 2(10), 719–731

Zara G et al (2021) Sexbots as synthetic companions: Comparing attitudes of official sex offenders and non-offenders. International Journal of Social Robotics https://doi.org/10.1007/s12369-021-00797-3

Zhao J et al (2017) Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In: Palmer M et al (eds) Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Stroudsburg, pp 2979–2989

Zhou P et al (2019) Examining gender bias in languages with grammatical gender. In: Padó S, Huang R (eds) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Stroudsburg, pp 5279–5287

Eduard Fosch-Villaronga is an Assistant Professor at the eLaw Center for Law and Digital Technologies at Leiden University (The Netherlands), where he investigates legal and regulatory aspects of robot and AI technologies, with a special focus on healthcare, diversity, governance, and transparency. Currently, he is the PI of PROPELLING, an FSTP from the H2020 Eurobench project, a project using robot testing zones to support evidence-based robot policies. Previously, Eduard served the European Commission in the Sub-Group on Artificial Intelligence (AI), connected products and other new challenges in product safety to the Consumer Safety Network (CSN) and was the PI of LIAISON, an FSTP from the H2020 COVR project that aimed to link robot development and policymaking to reduce the complexity in robot legal compliance.

In 2019, Eduard was awarded a Marie Skłodowska-Curie Postdoctoral Fellowship and published the book Robots, Healthcare, and the Law (Routledge). Eduard holds an Erasmus Mundus Joint Doctorate in Law, Science, & Technology, coordinated by the University of Bologna (Italy, 2017), an LL.M. from University of Toulouse (France, 2012), an M.A. from the Autonomous University of Madrid (Spain), and an LL.B. from the Autonomous University of Barcelona (Catalonia, Spain, 2011). Eduard is also a qualified lawyer in Spain.

Adam Poulsen is a computer scientist and researcher at Charles Sturt University in New South Wales, Australia. His research covers human-robot interaction, healthcare and social robotics, value sensitive design, computer ethics, care ethics, and LGBTQ+ aged care. At present, Adam primarily focuses on exploring the value sensitive design of robots to assist in, or enhance, the provision of care. Through his research, Adam has developed a novel value sensitive design approach, values in motion design, to model social robots for members of the older LGBTQ+ community experiencing loneliness. It is his hope that such robots can be helpful in the self-care of this vulnerable, under surveyed population and others in the future. Adam holds a Ph.D. and graduated 1st Class Honours of a Bachelor of Computer Science from Charles Sturt University (AU).
This work has been partly funded by the Global Transformation and Governance Challenges Seed Grant from Leiden University.