# A GENDERED PERSPECTIVE ON ARTIFICIAL INTELLIGENCE

*Smriti Parsheera*[1]

[1]National Institute of Public Finance and Policy, New Delhi

## ABSTRACT

*Availability of vast amounts of data and corresponding advances in machine learning have brought about a new phase in the development of artificial intelligence (AI). While recognizing the field's tremendous potential we must also understand and question the process of knowledge-making in AI. Focusing on the role of gender in AI, this paper discusses the imbalanced power structures in AI processes and the consequences of that imbalance. We propose a three-stage pathway towards bridging this gap. The first, is to develop a set of publicly developed standards on AI, which should embed the concept of "fairness by design". Second, is to invest in research and development in formulating technological tools that can help translate the ethical principles into actual practice. The third, and perhaps most challenging, is to strive towards reducing gendered distortions in the underlying datasets to reduce biases and stereotypes in future AI projects.*

**Keywords** – Artificial intelligence, gender, ethics, fairness

## 1. INTRODUCTION

The term artificial intelligence (AI) was coined in a Dartmouth summer research proposal in 1955 that described itself as a "*2 month, 10 man study of artificial intelligence*". John McCarthy, Marvin Minsky and their fellow drafters explained it as a "*proposal to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves"* (McCarthy et al, 1955[1]). They highlighted these as problems that needed a carefully selected group of scientists to work on them and there seemed to be no doubt about the gender of those researchers.

Sixty years hence, AI is seen as one of the most promising fields of computer science. Its latest boom is fueled by the availability of vast amounts of data and corresponding advances in machine learning and neural technology. Self-driving vehicles, cancer detection technologies, image recognition tools, language translation and virtual assistants are some of the many AI applications that we encounter in everyday conversations. The field has, however, gone through its share of "AI winters", characterized by cutbacks in funding when research outcomes failed to keep up with the claimed progress.

Today, we are in a phase of AI boom. As per most accounts, AI based systems will play a much greater role in the coming decades, redefining business models, job markets and overall human development. In all the euphoria surrounding AI and its future, not enough was being said about the underlying processes that drive research in this field. This has begun to change in the last few years as countries begin to adopt national or regional AI strategies, many of which incorporate an inclusion and ethics dimension in them (Dutton, 2018[2]).

Like most human creations, AI artifacts tend to reflect the goals, knowledge and experience of their creators. They also draw from the strengths and weaknesses of the data that is used to train them. It is therefore natural to expect the limitations and biases of the creators and their datasets to be reflected in their results. This leads us to ask some basic questions. First, what is regarded as AI, who designs it and to what end? Second, what is the basis for determining the elements of intelligence that are found worth replicating in machines? Finally, to what extent do these decisions reflect the diverse experience and needs of human society?

These are complex questions, and the answers will necessarily vary based on the respondent's standpoint -- education, gender, race, class, religion, nationality and the intersectionality of these factors. Despite recent attempts to "diversify" AI research, and more generally research in the fields of science, technology, engineering and mathematics (STEM), the discipline has retained a male-oriented focus. It is telling that when the Institute of Electrical and Electronics Engineers (IEEE) instituted a Hall of Fame to acknowledge the leading contributors to AI, not one of the ten persons on the list was a woman (Wang, 2010[3]).

A research environment that fails to account for the worldview of one entire gender group is clearly lacking in many respects. In making this claim, we are cognizant of the fact that just as there is no universal "human knowledge", it is also not possible to classify "men's knowledge" and "women's knowledge" into distinct buckets. There exist a multiplicity of viewpoints within these groups. A more inclusive, and indeed more fruitful, research agenda should ultimately be able to overcome these binaries. Recognizing the existence of a gendered perspective on AI is, however, the starting point for this conversation. While this paper uses the role of gender in AI

research as its lens of enquiry, the issues that it poses and the solutions that it suggests are also relevant to broader pursuits of inclusiveness in AI-based systems.

## 2. DEFINING AI AND ITS "INTELLIGENCE"

John McCarty, one of the founders of this field, described AI as the "*the science and engineering of making intelligent machines*" where intelligence refers to "*the computational part of the ability to achieve goals in the world*" (McCarty, 2007[4]). Another suggestion is to look at intelligence as a "*quality that enables an entity to function appropriately and with foresight in its environment*" (Nilsson, 2010[5]). Both these definitions, forwarded by practitioners of AI, refer to intelligence in rather broad terms, as qualities which can be possessed by humans, animals and machines, albeit, at different levels.

Russell and Norvig (2010)[6] present a classification of the available definitions of AI along two lines -- (i) based on the function expected to be performed (*thought processes/ reasoning* of the machine versus the *outcome/ behaviour* that it exhibits); or (ii) the metrics used for assessing the success of AI (*human performance* versus an ideal standard of "*rationality*"). The Turing test, developed by British mathematician and cryptographer Alan Turing in 1950, reflects a combination of the behavioral element and human-like performance in the above classification. If upon the exchange of a series of questions with a person and machine, a human interrogator is unable to distinguish between the two, the Turing test would regard the machine to be an intelligent, thinking entity (Copeland, 2018[7]).

Despite its continued relevance over the years, the Turing test has also come under attack for its attempt to define the intelligence of machines by replicating human behaviour. Russell and Norvig (2010)[6] point to this as a limitation by saying, "*Aeronautical engineering texts do not define the goal of their field as making machines that fly so exactly like pigeons that they can fool even other pigeons*".

AI's claims of building intelligence in machines have also faced strong philosophical criticisms. These criticisms stem from arguments about the lack of a mind, of consciousness and intentionality in machines, features which some philosophers regard as essential for establishing true intelligence. John Searle illustrated this through his famous Chinese room thought experiment. As per this, person who does not know any Chinese can follow a set of rules on how to correlate Chinese symbols and produce a response to questions that may convince an outsider that the person is acting intelligently. Producing meaningful replies in Chinese would however not mean that the person has any actual understanding of the language.

In making the claim that similar behavior by a computer programme cannot be equated with intelligence, Searle draws a distinction between "weak AI", where the computer serves as a tool to study the mind and "strong AI" where the computer itself can be said to possess a mind. He focuses his criticisms on the latter by arguing that in order to constitute strong AI a machine would need to satisfy the tests of consciousness and intentionality or causal powers that are possessed by the human brain (Searle, 1980[8]). Similar debates on the "intelligence" of AI have also emerged from other fields like psychology, economics, biology, neuro-science, engineering and linguistics (Russell and Norvig, 2010 [6]).

Feminist epistemologist Alison Adam notes that these popular criticisms of are lacking in two major respects. First, they gauge the success or failure of AI based on philosophical tests of ideal intelligence, which for Adam is less relevant than understanding how AI is actually being put to use. For her, the success of AI lies in its widespread adoption in everyday life. Second, she notes that the traditional critiques of AI completely ignore how AI systems reinforce existing power structures. AI research has failed to represent the knowledge of certain social groups, such as women (Adam,2005[9]). This has worked to the disadvantage of society as well as the field itself.

## 3. GENDER OF AI DEVELOPERS AND THEIR ARTIFACTS

While the contours of what constitutes intelligence in AI has remained contested, a more operational understanding of AI has also emerged. As per some researchers, AI can simply be defined as "*what AI researchers do*" (Grosz et al, 2016[10]). This approach clearly gives the practitioners in this field immense power, not just in defining their own agenda but also the contours of the discipline that they represent. It therefore becomes pertinent to discuss who are these researchers and what is it that they do?

### 3.1 Early choices in AI research

Interestingly, even though we have seen significant advances in AI applications in recent years, the fundamental elements of what constitutes AI research have not changed very significantly. In 1955, the Dartmouth College proposal identified the following as some of the components of the AI problems that needed further research: programming a computer to use a language (*natural language processing*); self-improvement by machines (*machine learning*); and neuron nets (*neural networks* and *deep learning*) (McCarthy et al, 1955[1]). The text in parenthesis reflects the currently in vogue terminology for these processes. While these areas of research still remain relevant, newer sub-areas like computer vision and robotics have also been added along the way (Grosz, 2016[10]).

This leads us to ask – on what basis did AI researchers decide that certain elements of intelligence (versus others) were worth replicating in machines? In 1950, Alan Turing admitted that he did not know the right answer. He

proposed that it would be prudent to try both the approaches that were being suggested at that time. The first would be to choose an abstract activity like playing chess and teach machines to do it. The second would be to equip machines with sense organs and teach them the right answers, like teaching a child (Turing, 1950[11]).

Adam, 1998[12] uses the fascination with chess in early AI works to demonstrate how the interests and worldview of AI researchers influenced their conception of what amounted to intelligent behavior. She refers to the following quote from Rob Wilnesky, an AI researcher, to illustrate the point:

> "*They were interested in intelligence, and they needed somewhere to start. So they looked around at who the smartest people were, and they were themselves, of course. They were all essentially mathematicians by training, and mathematicians do two things - they prove theorems and play chess. And they said, hey, if it proves a theorem or plays chess, it must be smart.*"

The choice of chess and theorem proving, both being activities predominantly associated with men, therefore became a natural choice for early AI researchers (Adam, 1998[12]). The choice of chess as a metric for proving machine intelligence is particularly interesting given that the game still continues to suffer from a significant gender problem, resulting in the under-inclusion and under-performance of women (Maass et al, 2007[13]). Yet, it would be hard to claim that that the early choices of AI researchers stemmed from any malice against women or their role is society. Instead, these decisions reflected the researchers' own experiences, interests and social conditioning.

The "context" of AI researchers, which includes their gender, has therefore defined the directions in which the field has progressed. It is possible to imagine that if the group contemplating early ideas for testing machine intelligence included some women, an entirely different set of ideas may have emerged.

### 3.2 Different dimensions of gender bias

It has been over seven decades since AI first emerged as a discipline and yet the gender imbalance in AI, and more broadly in the fields of STEM, still remains significant. As per data released by the UNESCO Institute for Statistics, women constitute less than 29 percent of scientific researchers globally (UNESCO, 2017[14]). Further, there are many inter regional differences, with many developing countries showing a lower percentage of women in science. For instance, in India's case the figure of women in science was only about 14.3 percent (UNESCO, 2017[14]).

A study involving computer science PhD graduates in India found that 32 percent of the graduating PhD students in 2016 were women (Parkhi and Shroff, 2016[15]). This figure is closer to the world average of women in science although it is also worth noting that a majority of the PhD graduates opted for teaching jobs and only a small number went on to join research labs. Therefore the percentage of women from this pool who might have gone on to engage in applied research is likely to be much smaller.

The under-representation of women in AI research has the corresponding effect of under-representation of their ideas in setting AI agendas. This imbalance also manifests itself in other forms that go beyond issues of direct representation. Firstly, the few women who do manage to enter this field have reported systematic discrimination in terms of salaries, promotions and incidents of sexual harassment (Vasallo et al, 2015[16]). This contributes to the leaky pipe problem in STEM. Secondly, the AI industry is also replete with examples of gender based stereotypes being reflected in the identities of AI artifacts, their functions and outputs. To some extent this can be attributed to the lack of diverse perspectives in the designing and testing of these artifacts.

For instance, virtual assistants like Apple's Siri, Amazon's Alexa, Google's assistant and Microsoft's Cortana commonly come with female sounding voices (although in some cases like Apple's Siri users were later given the option to change the default voice). This is also the case with most GPS assistants. Several factors may contribute to this. On one hand, it could be a conscious business decision, based on physical and psychological reasons for preferring a woman's voice for such machines. On the other, it may be a case of unconscious reiteration of society's existing gender stereotypes -- a woman's voice being regarded as more suitable for roles that demand obedience (Glenn, 2017[17]. Similarly, the names and body shapes given to robots and other AI solutions have also been known to reflect the prevalent socio-cultural norms and gender identities (Bowick, 2009[18]).

Another dimension of the gender problem in AI comes from the perceptions and stereotypes of the real world, the data that emerges from there and its use in training algorithms. This can be illustrated with a few examples. When translation services, like the one offered by Google, translate text from gender neutral languages like Turkish and Finnish to a gendered one like English, the algorithm tends to attribute a gender to the subject. This classification may be based on the profession being described – engineers, doctors, soldiers are generally described as "he" while teachers, nurses and secretaries would be "she". It could also relate to the activities or emotions in question – happiness and hardwork are associated with "he" while terms like lazy and unhappy with "she" (Morse, 2017[19]).

Bolukbasi et al, 2016[20] explain that this problem can be attributed to the blind adoption of "word embedding" techniques. Word embedding enables the mapping of the affinity or relationship between different words, where a public resource like Google News serves as the training dataset. The researchers illustrate how this could influence the search results for a person looking for a computer science researcher in a particular university because the words "computer science" are more commonly associated with men -- "*between two pages that differ only in the names Mary and John, the word embedding would influence the search engine to rank John's web page higher than Mary"* (Bolukbasi et al, 2016[20]). Similar findings of gender biases been also been made in case of visual recognition tasks like captioning of images (Zhou et al., 2018[21]) and display of image search results based on occupations (Kay, 2015[22]).

These examples demonstrate that AI applications can often end up strengthening and reinforcing society's existing biases. For instance, Zhou et al., 2018[21] found that where training images for the activity of cooking contained 33% more females, the trained model for captioning images amplified the disparity to 68%. This seems to run contrary to Donna Haraway's vision of a cyborg universe where technology would offer a tool to break away from the dualities of human-machine and male-female identities (Haraway, 1991[23]). This is an inspiring idea and one that we still have an opportunity to fix. Concepts of equity, fairness and non-discrimination have been well entrenched in the human rights discourse for the past several decades. Yet, conscious and unconscious human biases often prevent these values from translating into actual outcomes. How then can we re-envision AI research in ways that could move us closer to this ideal?

## 4. RE-ENVISIONING AI FROM A GENDERED PERSPECTIVE

Improving the representation of women in AI research, both as researchers and as beneficiaries of the research is seen as a first step towards a gendered re-envisioning of AI. This has led to initiatives like having specialized programmes for women, funding support, mentorship initiatives, increased intake in educational institutions and promoting equal opportunities in the job market. However, even if such initiatives were to succeed, it is questionable whether merely increasing the number of women can bring the desired level of diversity in AI knowledge-making.

In her work on objectivity and diversity, Sandra Harding notes that although increasing the physical presence of excluded groups is an important first step, the real issue goes beyond that of participation. It involves questioning *whose agendas should be pursued by science*? (Harding, 2015[24]). A research agenda that is primarily funded through private resources will logically rely on market mechanisms to decide on the kind of problems that need to be solved and their optimum solutions. In the long run, this could very well lead to the development of breakthrough technologies, the benefits of which may ultimately trickle down the marginalized sections of society. However, there is a distinction between retrofitting newer objectives into available technologies versus a ground up approach of identifying specific problems and developing solutions for them.

The latter approach would require a more meaningful engagement by businesses, governments and the public in identifying AI research agendas and supplying resources to pursue them. These resources could be in the form of financial support, ethical frameworks, as well as making available open data resources that can feed into the design of AI solutions. For instance, the development of AI applications that are useful for addressing the health concerns of rural women in a developing country like India may not be an obvious interest area for many AI researchers. This may stem both from the lack of funding for sustained research in such areas and also the lack of access to the data that is necessary for enabling this research. Similarly, the ways in which algorithmic credit will work out in the Indian setting may be very different from what happens in other parts of the world. Agenda setting for future AI research must therefore be rooted in the social and cultural backdrop and institutional context of each society.

Having said that, there is also a case for evolving a robust set of ethical standards for AI research and the tools for translating those principles into tangible outcomes. Questions of bias and ethics have already found a place in many national AI strategies. For instance, the United Kingdom has noted that although it cannot match countries like the United States and China in terms of AI spending, it intends to play a greater role in AI's ethical development (House of Lords, 2018[25]). In India, a discussion paper issued by the Government think tank NITI Aayog (NITI Aayog, 2018[26]) as well as an AI Task Force set up by the Indian Government have spoken about the need for ethical standards, including auditing of AI to check that it is not contaminated by human biases (AI Task Force, 2018)[27]. Both these documents are, however, conspicuously silent on the gender dimensions of AI education and research in the country. Most large technology companies also have internal ethics policies to govern their research initiatives. Moving from these siloed structures to a collectively designed set of global minimum standards for AI development should be the next goal. These principles can then be applied based on each region's own context.

This above proposal comes with the worry that absent strict enforcement, producers would tend to interpret any ethical guidelines in a flexible manner. This could resulting in the under-production of "fairness" in the system. The opacity of AI algorithms and possibility of diverse interpretations on what constitutes fairness in any given situation only

compound the problem. But trying to solve this issue through heavy-handed regulation and strict ex-ante controls would present its own set of challenges. Such interventions may come at the cost of stifling efficiency and innovation. This also presumes a certain level of state capacity to effectuate the regulation, which is often not available in reality. How then can we strike a balance between these positions to make sure that AI research evolves in a socially and ethically responsible direction? We propose a three step approach towards this goal.

The first step would be to embed the concept of "***fairness by design***" in AI frameworks (Abbasi et al, 2018[28]). This draws from the concept of "privacy by design" that has evolved in the context of data protection debates (Cavoukian, 2011[29]). Fairness by design should compel developers to ensure that the very conception and design of AI systems is done in a manner that prioritizes fairness. Abbasi et al, 2018[28] propose that the components of such a framework would include:

(i) creating cross-disciplinary teams of data scientists and social scientists;
(ii) identifying and addressing the biases brought in by human annotators;
(iii) building fairness measures into the assessment metrics of the program;
(iv) ensuring that there is a critical mass of training samples so as to meet fairness measures; and
(v) adopting debiasing techniques.

A fair amount of research has been done on building solutions for gender biases in natural language processing. For instance, Bolukbasi et al, 2016[20] use debiased word embeddings for removing negative gender associations from word embeddings generated from a dataset. Another strategy is to use gender swap techniques to remove any correlation between gender and the classification decision made by an algorithm (Park et al, 2018[30]). A variation to this would be to conduct "stress tests" where certain parts of the data (such as the gender of some candidates in a selection process) can be randomly altered to check whether the randomization has an effect on the final outcome that is generated, i.e. the number of women being shortlisted (Economist, 2018)[31].

While encouraging further research of this nature, a lot more needs to be done in terms of mainstreaming these solutions and making them readily available to smaller developers. Google's "What-If" tool offers a useful example. It is an open source tool that allows users to analyze machine learning models against different parameters of fairness. For instance, the data can be sorted to make it "group unaware" or to ensure "demographic parity" (Weinberger, 2018[32]). Given the many positive externalities to be gained from the creation and opening up of such fairness enhancing tools, the second step of the re-envisioning AI project would be for governments and other agencies to invest in more research and development on this front.

Finally, we must remember that the datasets being used for training machine learning algorithms are created in the real-world, i.e. outside the AI ecosystem. Therefore, while building reactive use-case based solutions (NITI Aayog, 2018[26]) may solve some of our immediate needs, the larger agenda must be to correct the training dataset itself. To take an example, the outcomes of natural language processing can be made more inclusive if the persons generating the underlying text (writers, researchers, policymakers, journalists, publishers and other creators of digital content) work towards the feminization (using words like *she* and *her*) and neutralization (*chairperson* instead of *chairman*) of the language that they use (Sczesny et al, 2016[33]). Here again, there is a role for the State to use awareness, education and, if required, other policy tools to promote the use of gender fair language. Similar solutions need to considered for other fields of AI research, accompanied by the identification of the persons and processes needed to effectuate the desired changes.

## 5. CONCLUSION

From its very inception, the field of AI has largely remained the domain of men. This paper illustrates how the gender of its founders and subsequent researchers has played a role in determining the course of AI research. While efforts are now being made to fill this gap, including by promoting more women in STEM, the gender problem of AI is not just about the representation of women. It is also about understanding whose agendas are being pursued in AI research and what is the process through which that knowledge is being created.

Research has shown that AI's reliance on real-world data, which is fraught with gender stereotypes and biases, can result in solutions that end up reinforcing or even exacerbating existing biases. While fairness and non-discrimination are well recognized principles in the human rights discourse, these principles often fail to translate into practice, often on account of the conscious and unconscious biases. The challenge therefore is to find ways to bundle the technological progress of AI with the objectives of pursuing greater fairness in society -- for machines to eliminate rather than reinforce human biases.

We propose a three step process towards this end. First, we need to develop a set of publicly developed AI ethics that embed the concept of "*fairness by design*". To travel the distance from formulating ethical principles to their actual implementation is another challenge. We find that "fairness" as a concept is prone to diverse interpretations, which can result in its under-production in the system.

The second step would therefore be to invest in research and development in formulating technological tools to

implement AI ethics. This would, for instance, include further work on developing debiasing and fairness testing techniques. Open dissemination of such solutions to make them readily available for adoption by the AI community, will generate positive externalities for the system as a whole. This will require cooperation among a range of stakeholders, including governments, corporations, universities and researchers working in the fields of computer science, social science and data science.

Finally, we need to think about deeper solutions for cleaning up the gender biases and stereotypes in the underlying datasets that serve as fodder for training AI algorithms. For instance, feminization and neutralization of language have been suggested as solutions to enhance fairer outcomes in natural language processing. Similar solutions need to considered for other fields of AI research along with an identification of the persons and processes that are necessary to effectuate the desired changes.

## REFERENCES


[1] McCarthy et al, 1955: John McCarthy, Marvin Minsky, Nathaneil Rochester and Claude Shannon, *A proposal for the Dartmouth summer research project on artificial intelligence*, August 31, 1955, available at https://www.cs.swarthmore.edu/~meeden/cs63/f11/AIproposal.pdf

[2] Dutton, 2018: Tim Dutton, *Artificial Intelligence Strategies*, Medium, 28 Jun 2018, available at https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd

[3] Wang, 2010: Fei-Yue Wang, IEEE Intelligent Systems, Volume: 26, Issue: 4, July-Aug, 2011, available at https://ieeexplore.ieee.org/document/5968105/

[4] McCarty, 2007: John McCarthy, *What is artificial intelligence*?, 11 December, 2007, available at http://jmc.stanford.edu/artificial-intelligence/what-is-ai/index.html

[5] Nilsson, 2010: Nils J. Nilsson, *The quest for artificial intelligence - A history of ideas and achievements*, Cambridge University Press, 2010.

[6] Russell and Norvig (2010): Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 3rd Ed, Prentice Hall Series in Artificial Intelligence, 2010.

[7] Copeland, 2018: BJ Copeland, *Artificial Intelligence*, Britannica Encyclopedia, available at https://www.britannica.com/technology/artificial-intelligence

[8] Searle, 1980: John Searle, *Minds, brains, and programs*, Behavioral and Brain Sciences, 3 (3), 417-457.

[9] Adam 2005: Alison Adam, *Gender, Ethics and Information Technology*, Palgrave Macmillan, 2005.

[10] Grosz, 2016: Barbara Grosz et al, *Artificial intelligence and life in 2030*, One hundred year study on artificial intelligence, September 2016.

[11] Turing, 1950: Alan Turing, *Computing Machinery and Intelligence*, Mind, 49, 433-460.

[12] Adam, 1998: Alison Adam, *Artificial Knowledge - Gender and the Thinking Machine*, Routledge, 1998.

[13] Maass et al, 2007: Anne Maass, Claudio D'Ettole and Mara Cadinu, *Checkmate? The role of gender stereotypes in the ultimate intellectual sport*, European Journal of Social Psychology, Volume 38, Issue 2, March/April 2008, 231-245.

[14] UNESCO, 2017: UNESCO Institute of Statistics, Fact Sheet No. 43, March, 2017, available at http://uis.unesco.org/sites/default/files/documents/fs43-women-in-science-2017-en.pdf

[15] Parkhi and Shroff, 2016: Sachin Parkhi and Gautam Shroff, *ACM Survey on PhD Production in India for Computer Science and Information Technology*, 2015-16, available at http://india.acm.org/PhDProductionReport2015_16.pdf

[16] Glenn, 2017: Marie Glenn, 2017, *Few good men: Why is the growing population of AI voices predominantly female?*, IMB Blog, 2 Mar, 2017, available at https://www.ibm.com/blogs/insights-on-business/ibmix/good-men-growing-population-ai-voices-predominantly-female/

[17] Bowick, 2009: Micol Marchetti-Bowick, *Is Your Roomba Male or Female? The Role of Gender Stereotypes and Cultural Norms in Robot Design*, Intersect, Volume 2, Number 1, 2009.

[18] Vassallo et al, 2015: Trae Vassallo et al, *Elephant in the Valley Survey 2015*, available at https://www.elephantinthevalley.com/

[19] Morse, 2017: Jack Morse, *Google Translate might have a gender problem*, Mashable, 1 Dec 2017, available at https://mashable.com/2017/11/30/google-translate-sexism/

[20] Bolukbasi, 2016: Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama and Adam Kalai, *Man is to Computer Programmer as Woman is to Homemaker?*

*Debiasing Word Embeddings*, 2016, available at https://arxiv.org/pdf/1607.06520.pdf

[21] Zhao, 2017: Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez and Kai-Wei Chang, *Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints*, 29 Jul 2017, available at https://arxiv.org/abs/1707.09457

[22] Kay, 2015: Matthew Kay, Cynthia Matuszek, and Sean A Munson, *Unequal Representation and Gender Stereotypes in Image Search Results for Occupations*, ACM CHI Conference on Human Factors in Computing Systems, Apr 2015, available at https://www.researchgate.net/publication/271196763_Unequal_Representation_and_Gender_Stereotypes_in_Image_Search_Results_for_Occupations

[23] Haraway, 1991: Donna Haraway, *A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century*, in Simians, Cyborgs and Women: The Reinvention of Nature, Routledge, 1991.

[24] Harding, 2015: Sandra Harding, *Objectivity and Diversity: Another Logic of Scientific Research*, University of Chicago Press, 2015.

[25] House of Lords, 2018: House of Lords, Select Committee on Artificial Intelligence, *AI in the UK: ready, willing and able?*, 16 Apr 2018, available at https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf

[26] NITI Aayog, 2018: NITI Aayog, *Discussion Paper - National Strategy for Artificial Intelligence*, Jun 2018, available at http://niti.gov.in/writereaddata/files/document_publication/NationalStrategy-for-AI-Discussion-Paper.pdf

[27] AI Task Force, 2018: Artificial Intelligence Task Force, Constituted by the Ministry of Commerce and Industry, Government of India, available at https://www.aitf.org.in/

[28] Abbasi, 2018: Ahmed Abbasi, Jingjing Li, Gari Clifford and Herman Taylor, *Make "Fairness by Design" Part of Machine Learning*, Harvard Business Review, 1 Aug, 2018, available at https://hbr.org/2018/08/make-fairness-by-design-part-of-machine-learning

[29] Cavoukian, 2011: Ann Cavoukian, *Privacy by Design - The 7 Foundational Principles*, 2011, available at https://www.ipc.on.ca/wp-content/uploads/Resources/7foundationalprinciples.pdf

[30] Park et al, 2018: Ji Ho Park, Jamin Shin and Pascale Fung, *Reducing Gender Bias in Abusive Language Detection*, August, 2018, https://arxiv.org/abs/1808.07231

[31] Economist, 2018: *For artificial intelligence to thrive, it must explain itself*, 15 Feb 2018, available at https://www.economist.com/science-and-technology/2018/02/15/for-artificial-intelligence-to-thrive-it-must-explain-itself

[32] Weinberger, 2018: David Weinberger, P*laying with AI Fairness, People+AI Research Initiative*, available at https://pair-code.github.io/what-if-tool/ai-fairness.html

[33] Sczesny et al, 2016: Sabine Sczesny, Magda Formanowicz, and Franziska Moser, *Can Gender-Fair Language Reduce Gender Stereotyping and Discrimination?*, Frontiers in Psychology, 2016; 7: 25, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4735429/