Re-Considering Bias: What Could Bringing Gender Studies and Computing Together Teach Us About Bias in Information Systems?

Claude Draude[1], Goda Klumbyte[2], Pat Treusch[3]

[1] University of Kassel, Pfannkuchstraβe 1, 3412 Kassel, Germany
claude.draude@uni-kassel.de
2 University of Kassel, Pfannkuchstraβe 1, 3412 Kassel, Germany
goda.klumbyte@uni-kassel.de
[3] TU Berlin, Marchstraβe 23, 10587 Berlin, Germany
p.treusch@tu-berlin.de

Abstract. This contribution focuses on interrogating the definition of bias in information systems. It serves as a discussion starter and asks what we could learn from approaches to bias analysis in the field of gender studies and how the methodologies developed in gender studies can be beneficial to understanding, analysing and managing bias in information systems. We look at two specific theories originating in gender and feminist science and technology studies – “situated knowledges” (Haraway) and “strong/weak objectivity” and “standpoint theory” (Harding). Specific parameters of gender-related bias are taken into account: androcentrism, over/underestimation of gender differences, stereotyping of gender traits and emphasizing dichotomies through research design. Through these we show how the above-mentioned theoretical framework can be applied to develop a better understanding of the workings of bias in information systems. The paper closes with pointing out the possibilities of a societally shared accountability for biased systems.

Keywords: Bias, Information Systems, Situated Knowledges, Standpoint Theory, Strong/Weak Objectivity, Gender.

As digitalization and the use of complex information systems as well as algorithmic tools gain speed, the question of bias in data, information and computational processes becomes ever more present both in academic and public discourse. Such bias can articulate in different forms, for instance, through the seemingly neutral prediction of crime rates as performed by ADM (algorithmic decision making [1]), or through software based on language processing and analysis, as for instance in the case of software trained on Google News that completed the sentence “Man is to computer programmer as woman is to X” with “homemaker.”

The knowledge produced through these systems and tools – be it “prediction,” analysis, pattern recognition, or other forms of output – is called into question especially when it turns out to favour certain social groups (such as favouring white against black populations in recidivism prediction) or when it reproduces stereotypes in social relations that are considered unfair or unethical (such as racial and economic discrimination patterns that are observable in predictive policing systems or sexual discrimination that until recently was observable in dominant search engines). The fundamental connection between algorithmic calculations and social relations has lately become more transparent through the work of critical math scholars such as O'Neil. She not only problematizes the role of computer-aided decision making, but also highlights the importance of choices that the developers make: "The math-powered applications powering the data economy were based on choices made by fallible human beings" [2]. While these choices became "opaque," the effects of the "encoded human prejudice, misunderstanding, and bias into the software systems" [2] result in an increase of inequality. In this regard, the effect of producing inequalities is grounded in the mathematical models themselves, making it critically important to investigate the biases, but also the neutrality and objectivity of mathematical models. Along these lines, however, it is rarely asked what exactly constitutes bias, how bias works and relatedly, what it would mean to generate non-biased or bias-free knowledge through computational systems. How does bias form? And: where to locate the bias – as part of the developer, as part of the information system design or as part of the algorithm? This also includes posing questions of accountability for a certain bias. What are the criteria for something to be considered biased? Is it even possible to produce un-biased knowledge through technological means?

Gender studies as an academic field deals not only with relations among genders, but critically reflects on systems of classification as such (man/woman, nature/culture, human/animal, – to name a few) and asks how these systems (re-)produce social inequalities. The approaches developed in gender studies highlight that social categories intersect. Furthermore, gender studies analyse how these intersections influence the way objectivity and knowledge production are understood and carried out. In this regard, to "[b]reak[...] with prejudices and reconstruct[...] the object of research requires a different way of seeing, in the light of which common-sense knowledge is reconstructed as a form of bias" [3]. As Agre [4] noted in his germinal work on critical technical praxis in building AI systems, in technical fields the concepts – such as "information," "intelligence," "knowledge" – are used both in specific mathematic, formalised terms, as well as in more colloquial terms. Thus, bias in information systems has a double connotation. It can be viewed in technical terms, but it is also important to interrogate in what ways the colloquial, everyday bias (and biased assumptions) play a role in information systems development. Gender studies deliver the analytical tools as well as the conceptual framework for acknowledging both – the double/simultaneous meaning of concepts (formalised and colloquial) as methods of reconstructing how social relations, including bias, are encoded into mathematical models.

To interrogate the intersection between different notions of bias and information systems we rely on two theories that originated in gender and feminist science and technology studies, namely "situated knowledges" (Haraway [5]) and "strong/weak objectivity” plus “standpoint theory” (Harding [6, 7]). In a nutshell, the theory of situated knowledges draws attention to how knowledge production cannot be cut off from the social and material positionality of a researcher, including their historical, conceptual, cultural, social, etc. context. According to Haraway, there is no “view from nowhere.” Thus, scientific claims are not universal. Instead, we need to re-think scientific knowledge production as valid from a specific perspective or position that operates always within certain figurations of time, space and artefacts; that is, situated knowledges. Harding’s theory of strong/weak objectivity and standpoint theory suggests that people involved in knowledge production must be attentive to relations of power that knowledge is always implicated in (whose perspective are we looking from? Who benefits from this perspective? Who loses?). We should, according to both theorists, aim not at “neutral” objectivity but rather at a kind of partial objectivity that acknowledges its perspective and is open about the benefits it produces to particular groups (while possibly excluding others).

These theoretical approaches can lead to, first, a better understanding of the workings of bias in information systems and, second, open up the possibilities of a societally shared accountability for biased systems. Specifically, we suggest that the perspective of situated knowledges points to the understanding of knowledge as a product of a complex network, where human researchers, data, data structure, algorithms, and broader social, political, historical and scientific context all contribute to the specific results that are produced (cf. [8]). This in turn invites to re-think bias as a complex phenomenon, distributed across the whole process of designing a particular information system. For instance, while researchers have accepted the possibility of data being biased, a situated knowledges perspective points to the importance of interrogating biases in the ways data is being classified (how are the categories formed? Which variables are being selected as important?) and processed (which kind of algorithms are built and used? How are different variables weighted in the process?), as well as to biases that occur in research design itself. In addition, this also means that bias is not a constant value, but rather that it can also change in relation to the categories formed, the variables selected and the ways in which data is processed.

As a starting point, gender studies provide specific parameters that can be used raising awareness towards gender-related bias in information system development, such as

- androcentrism (un/conscious focussing on masculine/male perspective);
- over/underestimation of gender differences (gender differences are either paid too much attention where they would not generally play a big role, or they are left unnoticed where such attention is due);
- stereotyping of gender traits (ascribing certain values/expectations/character to different genders, often reproducing prejudices that are well established in society);
- emphasizing dichotomies (focussing on showcasing differences between genders where such differences are not of major significance).

Since gender is always intersectional, a more exhaustive list of parameters needs to include stereotyping of racial traits and/or bias in regard to sexual orientation, religion, age, socioeconomic status and dis_ability.

Furthermore, relying on Harding's notion of strong/weak objectivity and standpoint theory, we argue that to develop better accountability standards and reduce biases occurring through the development process, attention should be paid to the purpose and the expected results of a particular information system. One way of doing that would be by developers and researchers paying a closer look at the concepts that are used to describe what information systems do, and how the meaning of those concepts shifts in different contexts. For example, the notion of "prediction" in algorithmic systems for developers might mean that the system analyses data and discovers patterns that express strong correlation. However, once such a system is used for the purposes of drawing policy suggestions (as in "predictive policing," for instance), the notion of "prediction" acquires a different, more colloquial meaning, thus affecting the expectations of the user and the (mis)understanding of what kind of knowledge the system generates.

A more "distributed" understanding of bias as occurring throughout the process affects also where the accountability for bias should be located. Nissenbaum [9] suggests that responsibility in computerised society needs to be re-defined in a more dispersed and nuanced way since the ownership of blame does not follow a clear, linear path but is instead more scattered through a network of actors. One recent example illustrates the need for such nuanced sensibility: in March 2016 Microsoft released the AI chatbot named Tay. However, "[t]he 19-year-old female chatbot was promptly coopted by a series of internet trolls and within 24 hours became a neo-Nazi mouthpiece for racist and sexist epithets" [10]. How much responsibility for this biased outcome of Tay's performance should be ascribed to software developers? How much are society and the users to blame? Tay is just one example that displays how strongly the question of bias in information systems relates to the question of accountability for how they are brought into use.

To sum up, this position paper is meant as a discussion starter and inspiration for further research into a more ethical and socially just information systems design by tapping into the interdisciplinary potential of gender studies. In short, we provided insight into gender studies approaches on knowledge production and asked how these approaches could be useful in accounting for bias in information systems. Understanding knowledge production – and the occurrence of bias – as a complex, embedded phenomenon and interrogating not only the processes but also the purposes and expectations of building information systems, helps understand accountability as a distributed process.

References

1. ADM Manifest, https://algorithmwatch.org/de/das-adm-manifest-the-adm-manifesto, last accessed 2018/01/19.
2. O'Neil, C.: Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy. Penguin Random House, London (2016): 3.
3. Oakley, A.: Experiments in Knowing: Gender and Method in the Social Sciences. Polity Press, Cambridge (2000).
4. Agre, P. E.: Toward a Critical Technical Practice: Lessons Learned in Trying to Reform AI. In: Bowker, G., Gasser, L., Star, L., Turner, B. (eds.): Bridging the Great Divide: Social Science, Technical Systems, and Cooperative Work. Erlbaum, New York (1997).
5. Haraway, D.: Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. Feminist Studies 14(3), 575–599 (1988).
6. Harding, S: Objectivity and Diversity: Another Logic of Scientific Research. The University of Chicago Press, Chicago (2015).
7. Harding, S.: Whose Science? Whose Knowledge? Thinking from Women's Lives. Cornell University Press, Ithaca (1991).
8. Akrich, M.: The De-Scription of Technical Objects. In: Bijker, W. E., Law, J. (eds.): Shaping Technology/Building Society. Studies in Sociotechnical Change, pp. 205-224. Cambridge, Mass: MIT Press (1992).
9. Nissenbaum, H.: Accountability in a Computerized Society. In: Friedman, B. (ed.): Human Values and the Design of Computer Technology, pp. 41-64. CSLI Publications/Cambridge University Press, New York, Melbourne (1997).
10. Montalvo, F. L.: Debugging Bias. Busting the Myth of Neutral Technology. bitch 17, 36-40 (2016).