Involving teachers in meta-design of AI to ensure situated fairness

Marie Utterberg Modén [1], Johan Lundin [1], Martin Tallvid [1] and Marisa Ponti [1]

[1] *Department of Applied IT, University of Gothenburg, Sweden*

Abstract

In this paper we propose a number of approaches to using formative design interventions to enable secondary school teachers to inform future design of artificial intelligence. The aim is to provide them with increased control, responsibility, and accountability for the deployment of AI-based applications in education to ensure fairness. The motivation for this project draws on results from our prior research on adaptive digital textbooks with AI-based technology. Participatory design has been recognized as a way of exploring workers knowledge and gaining knowledge of workplaces to improve system design when building new tools. However, we argue that applying PD methods in design of AI-based applications is somewhat different. Our intention with this paper is to introduce a discussion of methods and techniques for user involvement in design of AI, as well as to propose a possible remedy i.e., meta design.

Keywords

Artificial intelligence, AIEd, situated fairness, participatory design, activity theory, formative interventions, meta-design

1. Introduction

Participatory design (PD) has traditionally been a way to understand technology and how technology could be integrated in work activities, and also to empower future users participating in design and use of technology [14]. The use of PD in educational research addressing teaching and learning practices is still less extensive, although the interest is growing as PD includes methods to support members of a school community to engage in democratic processes in developmental interventions [15]. However, as discussed in literature, it is not clear how PD methods could be applied in design of artificial intelligence (AI) based applications to maintain PD values and desires. As pointed out by, for example Bratteteig and Verne [16, p. 3], "AI poses some new challenges to PD as the technology is different to other computing technologies by the fact that its behavior is unpredictable as it changes over time as it accumulates data presented to it – also from insufficient or biased data." Thus, researchers explore how to benefit from PD engagements that involve emerging AI technology (se for example [17, 18]).

We propose that using formative design interventions could enable users (in our case secondary school teachers) to inform future design with the aim of providing them with increased control, responsibility, and accountability for the deployment of AI-based applications. In our case we are particularly interested in AI systems in education (AIEd) in realtion to the concept of fairness. The work presented here is preparatory for future project, so to this point we have not yet started data collection, rather we intend to introduce a discussion of methods and techniques aiming for participants in a PD project to engage in telling of stories, making of things, and enacting possible futures [1] to generate successful design participation and result.

Our research project stems from the fact that intensified use of educational technologies in schools have led to a growing proportion of digital data being integrated in teachers' and students' everyday

Proceedings of CoPDA2022 - Sixth International Workshop on Cultures of Participation in the Digital Age: AI for Humans or Humans for AI? June 7, 2022, Frascati (RM), Italy

EMAIL: marie.utterberg@ait.gu.se (M. Utterberg Modén); johan.lundin@ait.gu.se (J. Lundin); martin.tallvid@.ait.gu.se (M. Tallvid); marisa.ponti@.ait.gu.se (M. Ponti)
ORCID: 0000-0002-3820-4063 (M. Utterberg Modén)

work [2, 3]. In this "data-driven" development, the generation and use of data means interpreting teachers and students work as quantifiable information, often referred to as datafication [2]. The flow of data has allowed for AIEd that could benefit teaching and learning. AI is already present in many aspects of peoples' life, and there is a growing interest in AI in educational environments. For example, digital textbooks with data-driven functionality have grown steadily in terms of use and are now an important part of many Swedish students' educational resources, including AI-applications in terms of intelligent tutoring systems, learning analytics and performance predictions [3].

There is an ongoing discussion about potential risks of harm and unfairness dependent on algorithmic discrimination, which has been demonstrated in areas such as criminal justice [4], recruitment [5] and face recognition [6]. The main critique concerning bias in AI has been on unfairness and unequal treatment based on race/ethnicity, gender, and nationality. It has also been noted that research evidence of bias and potential issues due to data collection and processing of data sets are "often implicit in the findings of prior work, rather than a primary focus of it" [7]. Baker and Hawn [8] state there is a lack of knowledge of how algorithmic bias affect, and potentially could affect, education. Importantly, there is a burgeoning interest in this topic, reflected in the increasing amount of literature discussing issues of bias and fairness (se for example [9, 10]). Taken together, literature call for a socio-technical view bringing together technology and the broader social context where AI operates to be able to create social and legal outcomes that account for human well-being. It has been highlighted that, "when an AIEd system fails to produce the desired outcome, the teachers are often unaware of how to proceed next" [11, p. 17]. Teachers act as gatekeepers when it comes to introducing AIEd in classrooms [12], and teachers should be able to expect that AIEd besides supporting them in their teaching should reduce, or at least not increase, discrimination and injustice, and must be able to trust the systems to use them [13].

The remainder of this paper is structured as follows. In the next section we briefly highlight aspects of fairness in AI since that is the motivation for our research project and what the participants are going to work with. The next section introduces formative interventions and meta-design in relation to our project, followed by a section where we argue for PD but also highlight challenges when involving AI technology. Finally, we present our suggested methods and techniques in the PD process.

2. Fairness

Choices during the design process reflect designers' and developers' ethical values, knowingly or not, and shape the technology and in the long-term shapes people's life [19]. People are usually unaware of when and how shared norms guide their ideas and behavior and that these norms are culturally specific rather than global. This commonly leads people to miss or ignore what are not typical for them, although this may well be the expression of other peoples' normality [20]. Thus, "when AIEd favors certain pedagogies, learning styles, and educational systems, it ultimately dis/ advantages certain students and their communities" [21, p. 339]. Nye [22] shows in a review that culture and language influenced the design and programming of AIEd systems and became a barrier when they were transferred to other contexts, which is a risk when algorithms are designed to be independent, abstract, and portable [10].

Further, the data set that are used to train algorithms can be endowed with societally prejudices encompassed with historical experiences and produce biases. Algorithms are based on machine learning that "find patterns within datasets that reflect our own implicit biases and, in so doing, emphasize and reinforce these biases as global truth" [23, p. 1524]. The language translation Google Translate offer one such example when training data reflected societal bias, reinforced gender roles, and maybe amplified them. Some languages gender nouns and some are gender neutral, and the translation to feminine and masculine forms reinforced gender bias and changed gender in a stereotypical way. Google translated for example presupposed doctors to be males, nurses to be females, and that he works, and she cooks.

An algorithm is designed as a statistical model of reality to predict potential outcomes and even a complex model is inevitably a simplification. A teaching and learning activity consist of a multitude of intertwined relations and the complexity in educational contexts cannot be underestimated [24]. It follows that AI technology needs to capture a messy, interactive, changeable and context bound world

[25]. It is important to consider how groups are represented, which people belong to a group, and not treat diverse groups as a single entity. Baker and Hawn [8] have found that when applying algorithmic models and group differences are ignored, or not accounted for, it has given rise to unfairness. For example, if an AIEd system is designed to predict which students bear the risk of falling behind and flag them to teachers, how this group should be defined is not trivial.

Kitchin [26] has addressed that data largely have come to be pre-analytical and pre-factual, meaning that data representing the truth. As such, algorithms are approached as inherently neutral and formal constructs. Similarly, Birhane [25] puts forward that in these times when social activities are becoming increasingly automated through algorithmic decision making and predictions, social activities are at the same time being transformed and aimed to be understood with a mathematical solution. The problem to be solved by algorithms are formulated as a mathematical model and unfair results are treated with a rational and logical solution. In line with this, Akgun and Greenhow [27] discuss AI in K-12 education and highlight that: "The ethical challenges and risks posed by AI systems seemingly run counter to marketing efforts that present algorithms to the public as if they are objective and value-neutral tools" (p. 4). For example, during the COVID-19 pandemic, students in the United Kingdom were awarded A-level grades through application of an algorithm. However, the algorithm was inconsistent and unfair, and in favor of private/independent school students whilst those from disadvantage backgrounds were negatively affected. Smith [28] stressed that it had to become uproar and public pressure for the government to make a U-turn and abandon the algorithm and issue correct grades. In summary, aspects of fair treatment of students are a huge challenge to AI in education. It is also the case that fairness must be understood as local and situated in particular educational practices. What might be fair to one group of students, in one set of activities, might be highly biased in another group. To address this issue, it is imperative that teachers are involved in the design process, as well as to investigate possibilities of adapting already deployed systems. We argue that meta-design [30] could be a possible form for allowing local adaptation of AIEd.

3. Formative interventions and meta-design

To identify relevant forms for meta-design, i.e., how it would be possible and relevant for teacher to adapt their AI-based applications to their classroom, we engage with teachers in PD. To formulate suitable interventions, we draw on Activity Theory [29]. Our unit of analysis will be the work activity (teaching) conceptualized as a collective activity system that participants can re-design and transform by identifying and solving problems and contradictions. Developing formative interventions for meta-design will allow teachers to rethink their role in their interaction with AIEd and will offer them methods they can use to find solutions to address evolving circumstances and pedagogical challenges. Therefore, we will engage with teachers, researchers, and other stakeholders (e.g., students, developers, and school leaders) in formative interventions, an approach that builds on and purposefully fosters participants' agency. As researchers, we will set up these interventions to investigate the possibilities for educational improvement by bringing in a new actor - namely, adaptive AI-based teaching materials - potentially seen as a problematic and contradictory object which can generate conflicts and resistance when used by teachers. This collective design effort is seen as part of an expansive learning process including participatory analyses and implementation phases. In fact, during these interventions, teachers and other participants will be involved in negotiations and debates to make their voices heard. Besides, the collaboration between teachers, researchers, and the other stakeholders within these formative interventions can result in the construction and implementation of a new organization of work by the collective. The result of these interventions is not known ahead of time to the researchers, as the outcome is determined by participants.

The purpose of our interventions is to develop new models for increasing the agency of teachers concerning the design and use of AI-based applications in education to ensure fairness. Based on previous work on algorithmic fairness we argue for an increased focus on situated fairness, i.e., fairness in practice. Such a perspective also addresses the difficulties and complexities of achieving fairness in algorithms and allows for teachers to compensate to ensure fairness in their practice. To reach the purpose of this project, we will develop formative interventions involving teachers and other stakeholders to aim to:

1. Identify the affordances and constraints of the AI-based application and how teachers would use it to appropriate and adapt it in their local practices.
2. Describe how and under what circumstances teachers perceive the AI-based application to benefit their local teaching practices.
3. Identify with teachers and other school stakeholders which features of the system can facilitate fairness and minimize bias for vulnerable groups.

We will focus our investigation on the potential of adaptive AI-based teaching materials. This type of application is intended to serve as a text for a course and integrates an intelligent tutoring system. We have chosen it because it is in use and is commercialized by a large established publisher in Sweden.

4. Method

We argue that it is necessary to identify how teachers interpret fairness in their local situations and to ensure that their interpretations underlie concrete system functionalities. Involving teachers is crucial as the interaction between them and AIEd should be one in which teachers become able not only to understand but also challenge algorithmic decisions and predictions.

4.1. Participatory design

Participatory design has been recognized as a way of exploring workers (tacit) knowledge and gaining knowledge of workplaces to improve system design when building new tools. To achieve this goal, researchers and designers and future users collectively participate in iterative re-design processes [14]. In this case, the goal is to make design changes on an AI-based application on ideas from teachers as future users, and other stakeholders, to meet their needs for the tool to be meaningful and sustainable. An extension is our ambition for the AI-based application to be designed for teachers to design after design, i.e., meta design [30]. Ideally, users are actively included in the pre-design process (ideation), during the tool design (development), and later in the process to test the tool in situ (implementation), although an evaluation of the tool in the use context often is difficult due to time limitations beyond the actual PD project [31]. A key aspect is mutual learning, that is facilitated by the collaborative nature of PD and methods ensuring that all participants have a say. Not only are researchers/designers learning about participants diverse skills, experiences, and work conditions. But the intention is also for the participants to learn about design, technology, and their own work.

AI uses data to recognize specific patterns or properties through statistical analyses. The data is labeled, and the performance of algorithms is calibrated against the correctness to the data, to make constant improvements, i.e., AI learns and changes over time. There are a wide range of AI features that can be designed, such as training data, algorithms, user interface and explanations of decision-making. However, AI-based applications is difficult to understand with its advanced technology, but also due to its black box properties making the relationship between input and output opaque and gives an unpredictable behavior. This means it becomes difficult to control actions and foresee effects on the social environment in the long term, which are valuable features in PD [1]. Thus, there are challenges needed to be addressed to make PD projects successful if they involve AI technology. Bratteteig and Verne [16] stress that participants need to understand the nature of AI and how it works to make design decisions in a PD project. The authors introduce a discussion of how future users can participate in a PD project and engage with AI technology. They start from three phases; First, bringing in a larger number of design ideas and imaging possible future activities can help to understand possibilities and limitations of AI-based applications. To do so, it will be necessary for participants (in this case teachers and other stakeholders) to consider different consequences of use and to be able to shift perspectives. Second, selecting one design idea among many can contribute to enhanced understanding of AI. The design idea can be made concrete by exploring possible futures and trying to understand how the idea will be experienced in relation to existing values. Finally, evaluating design decisions by participants can be facilitated if they use every day AI experiences. One challenge is the long-term evaluation of AI applications as they develop over time, which is usually not reasonable within the duration of a PD project. Also, similar AI applications develop differently dependent on the use.

An overview of our suggested research method is provided and summaries the details of data collection and data resources (Table 1).

Table 1: Methods

Aim	Activity	Comment
Identify affordances and constraints of the AI-based application and in which ways it would be relevant to teachers to change the system in their contexts to meet their local needs	Future Workshops with different groups: students-teachers-administration - developers. Focus will be on future workshops where each group will be introduced to AI, envision future practice, and discuss how fairness and inclusion in the development could be set up, as well as for which purposes AI are seen as beneficial in relation to their practice.	The purpose here is to allow a group of users, that are quite unaware of the practical consequences of AI technology to think about and envision what such use might mean to their everyday work.
Identify the main benefits that teachers perceive with the AI-based application.	Design studios. The different groups participate in design studios where they create applications on paper/computer with desirable applications. AI and fairness are discussed based on these created systems.	Here we want the teachers to engage in actual design work. And through involvement in design explore the possibilities, but also limitations of such systems. In particular we want to address solutions including meta-design and local adaptation.
Elicit how teachers interpret **situated fairness**. Find their views on how AI-based applications should be designed to facilitate fairness and minimize bias for vulnerable groups. Identify dilemmas and discuss what **values** they think should be **prioritized**. Identify how school regulations, norms, and expectations influence teachers' interpretations of fairness.	Semi-structured interviews with students-teachers-administrators and school leader-developers. Selected parts from videos from the workshops will be used as triggers for discussion	After a number of practical involvements with the AI application we now finally want to explore the experiences that teachers made in the previous activities. The broad purpose here is to identify what would be relevant dimensions for teachers to be able to locally adapt their use of AIEd-systems – i.e., in what ways should the systems cater for meta-design?

5. Acknowledgements

This work is funded by Marianne and Marcus Wallenberg Foundation.

6. References

[1] Brandt, E., Binder, T., Sanders, E.: Tools and techniques. Ways to engage telling, making and enacting. In: Simonsen, J. and Robertson, T. (eds.) Routledge international handbook of participatory design, pp. 145-181. Routledge (2012).

[2] Williamson, B.: Big data in education: The digital future of learning, policy and practice. Sage, London, (2017).

[3] Utterberg Modén, M.: Teaching with Digital Mathematics Textbooks-Activity Theoretical Studies of Data-Driven Technology in Classroom Practices, Ph.D. thesis, Gothenburg University, (2021). http://hdl.handle.net/2077/69472

[4] Angwin, J. Larson, J., Mattu, S., Kirchner, L.: Machine Bias: there's software used across the country to predict future criminals. And it's biased against blacks. ProPublica, (2016, May 23). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[5] Meyer, D.: Amazon Reportedly Killed an AI Recruitment System Because It Couldn't Stop the Tool from Discriminating Against Women. Fortune, (2016). https://fortune.com/2018/10/10/amazon-ai-recruitment-bias-women-sexist/

[6] Cheng, S.: An algorithm rejected an Asian man's passport photo for having ′closed eyes′. Quarts, (2016). https://qz.com/857122/an-algorithm-rejected-an-asian-mans-passport-photo-for-having-closed-eyes/

[7] Olteanu, A., Castillo, C., Diaz, F., Kiciman, E.: Social data: Biases, methodological pitfalls, and ethical boundaries. Frontiers in Big Data 2:13, 2019. https://doi.org/10.3389/fdata.2019.00013

[8] Baker, R. S., Hawn, A.: Algorithmic bias in education. International Journal of Artificial Intelligence in Education, (2021). https://doi.org/10.1007/s40593-021-00285-9

[9] Dignum, V.: The role and challenges of education for responsible AI. London Review of Education 19(1), 1-11 (2021).

[10] Selbst, A., Boyd, D., Vertesi, J.: Fairness and abstraction in sociotechnical systems. In: Proceedings of the 19th Conference on Fairness, Accountability, and Transparency (FAT), 2019. https://dl.acm.org/doi/pdf/10.1145/3287560.3287598

[11] Bhimdiwala, A., Neri, R. C., Gomez, L.M.: Advancing the Design and Implementation of Artificial Intelligence in Education through Continuous Improvement. International Journal of Artificial Intelligence in Education, (2021). https://doi.org/10.1007/s40593-021-00278-8

[12] Utterberg Modén, M., Tallvid, M., Lundin, J., Berner, L.: Intelligent tutoring systems: Why teachers abandoned a technology aimed at automating teaching processes. In: Proceedings of the 54th Hawaii International Conference on System Sciences, (2021). http://hdl.handle.net/10125/70798

[13] Qin, F., Li, K., Yan, J.: Understanding user trust in artificial intelligence-based educational systems: Evidence from China. British Journal of Educational Technology 51(5), 1693–1710 (2020).

[14] Bratteteig, T., Bødker, K., Dittrich, Y., Mogensen, P. H., Simonsen, J.: Methods. Organising principles and general guidelines for participatory design projects. In: Simonsen, J. and Robertson, T. (eds.) Routledge international handbook of participatory design, pp. 117-144. Routledge (2012).

[15] Cumbo, B., Selwyn, N.: Using Participatory Design Approaches in Educational Research. International Journal of Research & Method in Education 45(1), 60-72 (2022).

[16] Bratteteig, T., Verne, G.: Does AI make PD obsolete? Exploring challenges from artificial intelligence to participatory design. In: Proceedings of the 15th Participatory Design Conference (2018). https://doi.org/10.1145/3210604.3210646

[17] Choi, J. H., Forlano, L., Kera, D.: Situated automation. Algorithmic creatures in participatory design. In: Proceedings of the 16th Participatory Design Conference (2020). https://doi.org/10.1145/3384772.3385153

[18] Razak, T. R., Ismail, M. H., Fauzi, S. S. M., Gining, R. A. J. M., Maskat, R.: A framework to shape the recommender system features based on participatory design and artificial intelligence approaches. International Journal of Artificial Intelligence 10(3), 727-734 (2021).

[19] Borenstein, J., Howard, A.: Emerging challenges in AI and the need for AI ethics in education. AI and Ethics 1, 61-65 (2021). https://doi.org/10.1007/s43681-020-00002-7

[20] Blanchard, E.G.: Socio-Cultural imbalances in AIED research: investigations, implications and opportunities. International Journal of Artificial Intelligence in Education 25(2), 204–228 (2015).
[21] Schiff, D.: Out of the laboratory and into the classroom: the future of artificial intelligence in education. AI & society 36(1), 331–348 (2021).
[22] Nye, B.D.: Intelligent tutoring systems by and for the developing world: A review of trends and approaches for educational technology in a global context. International Journal of Artificial Intelligence in Education 25(2), 177-203 (2014).
[23] Howard, A., Borenstein, J.: The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. Science and engineering ethics 24(5), 1521-1536 (2018).
[24] Selwyn, N.: Education and technology. Key issues and debates. Bloomsbury Academic, London (2017).
[25] Birhane, A.: Algorithmic injustice: a relational ethics approach. Patterns 2(2), 1–9 (2021).
[26] Kitchin, R.: The data revolution: big data, open data, data infrastructures and their consequences. Sage, London (2014).
[27] Akgun, S., Greenhow, C.: Artificial intelligence in education: Addressing ethical challenges in K-12 settings. AI and Ethics (2021). https://doi.org/10.1007/s43681-021-00096-7
[28] Smith, H.: Algorithmic bias: should students pay the price? AI & society 35(4), 1077–1078 (2020).
[29] Engeström, Y.: Learning by expanding. An activity-theoretical approach to developmental research. Cambridge university press (2015).
[30] Fischer, G., Fogli, D., Piccinno, A.: Revisiting and broadening the meta-design framework for end-user development. In: Paterno, F. and Wulf, V. (eds.) New perspectives in end-user development, pp. 61-97. Springer (2017).
[31] Bratteteig, T., Wagner, I.: What is a participatory design result? In: Proceedings of the 14th Participatory Design Conference (2016). http://dx.doi.org/10.1145/2940299.2940316