Interaction Design and Architecture(s) Journal - IxD&A, N.59, 2023-24, pp. 62 - 78
DOI: 10.55612/s-5002-059-002

A Remedy to the Unfair Use of AI in Educational Settings.

Johan Lundin[1], Marie Utterberg Modén[1], Tiina Leino Lindell[1], Gerhard Fischer[2]

[1] Department of Applied IT, University of Gothenburg, Sweden
[2] Center For Life Long Learning, Department of Computer Science and Institute of Cognitive Science, University of Colorado, Boulder, USA
{johan, lundin}@ait.gu.se

Abstract. This paper addresses concerns related to the ethical implications of artificial intelligence (AI) and its impact on human values, with a particular focus on fair outcomes. Existing design frameworks and regulations for ensuring fairness in AI are too general and impractical. Instead, we advocate for understanding fairness as situated in practice, shaped by practitioners' values, allowing stakeholders control in the situation. To accomplish this, the paper contributes by conceptually exploring a potential synergy by combining Cultural-Historical Activity Theory (CHAT) and Meta-Design. By doing so, human activities can be transformed to deal with challenges, in this case, those emerging from adaptive AI tools. While professional software developers are essential for making significant changes to the tool and providing solutions, users' involvement is equally important. Users are domain experts when it comes to determining practical solutions and aligning structures with their work practices. CHAT contributes through its emphasis on context, history, and mediation by tools. This enables a critical analysis of activity systems, helping to reveal underlying contradictions and identify areas where improvements or innovations are necessary. Meta-Design provides design concepts and perspectives that aim to empower participants, allowing them to actively shape the processes of tool design to align with their specific local needs and evolving conceptions of fairness in use-time. This offers an approach to empowering people and promoting more fair AI design.

Keywords: Fairness, Artificial intelligence, Education, Teachers, Educational technology, Cultural-historical activity theory, Meta-design

1. Introduction

The improved capabilities of artificial intelligence (AI)-based tools raise concerns about their potential impact on central human values. This paper explores ethical considerations related to the influence of AI on human values, with a specific focus on promoting equitable outcomes. Existing design frameworks and regulations intended to ensure fairness in AI are broad and challenging to implement effectively.

The ethical aspects of use, implementation, and consequences have generated an intense debate. In this paper, we particularly address the issue of fair outcomes. The pervasive development and integration of AI into our daily lives mean that we regularly interact with AI tools, often without conscious awareness. This integration spans various activities, from AI-assisted purchasing and customer support through chatbots to personalized recommendations in TV and music streaming services. Noteworthy developments in generative AI, where large language models serve as the foundation for automated production of text, images, and sound, also merit attention. The ability of ChatGPT and similar language models in generating text indistinguishable from human-generated text [1] poses a challenge across intellectual practices. However, despite significant advancements in AI technology, the ongoing question is how people can take control to make these tools align with the values of fairness.

In this paper, we argue that current design frameworks, legislation, and policies aimed at achieving fairness are overly broad and impractical in real-world application. Our main argument is that fairness must be understood as situated in practice. That is, it is shaped, interpreted, and upheld by the practitioners involved, grounded in their values. Consequently, the ethicality and fairness of a given technology may vary between contexts, being deemed fair and ethical in one context while deemed unfair and unethical in another, representing one important of many design trade-offs being typical for changes in the digital age [33]. Thus, the interpretation of fairness must emanate from human values, highlighting the importance of considering the social and cultural contexts in determining what is deemed fair. However, a significant challenge arises from the fact that many of the AI tools deployed today are opaque and difficult for users to understand or influence [2]. This lack of transparency and user control poses a potential threat to the principle of fairness and hinders AI tool users from intervening to ensure fair practices.

Our overarching aim is to explore how AI tools could produce fair and just outcomes, placing emphasis on the unique circumstances of each situation where these tools are used. Our focus is on promoting the ethical and social responsibility of AI tools, with a particular emphasis on preventing biased or discriminatory outcomes. We assert that users of AI tools should be able to confirm, challenge, or extend the functionalities of these tools to ensure fairness. To facilitate this, we explore the integration of Cultural-Historical Activity Theory (CHAT) [3] and Meta-Design [4], proposing that this fusion offers conceptual contributions for future design processes. We suggest that adopting such a perspective could enable the development of tools that empower users to make local adaptations.

Given the applied focus on contextualized understanding, our argumentation in this paper relies on a concrete setting to anchor our examples. Therefore, we use the lack of contextualized understanding as a starting point, employing the case of education as a narrative vehicle for our discussions.

2. AI and Fairness

In society, the potential risks and benefits of AI tools have been widely debated [5]. With significant investments in research and development, there is obviously a lot at stake right now when it comes to AI. The discussion has often been polarized, as is common in public debates, and it can be claimed to overemphasize possible benefits versus risks. While AI tools are often portrayed as value-neutral, emphasizing positive applications and potential, they align with a broader narrative of effectiveness, optimization, and cost-saving. This alignment is underscored by Birhane et al. [6, p. 182], who, in their review of high-impact conference papers in the machine learning field, noted a "*lack of consideration of potential negative impacts and the prioritization and operationalization of values such as performance, generalization, efficiency, and novelty*". Conversely, there have been various efforts to restrict and regulate the use of these tools due to anticipated significant risks. These efforts include legislation, policymaking, and guidelines for the implementation or design of AI tools.

One might argue that aspects of bias or unfairness must always be considered in the automation of work and public life. Any tool that produces support for decisions or, in practice, makes decisions on our behalf needs to align with our beliefs about what is considered fair or unfair. However, there are aspects of AI tools that make them particularly problematic in this respect [7]. Most significantly, many of the currently proposed tools are based on machine learning, which makes the decision-making process less transparent compared to other tools [8]. This, in turn, makes it difficult to audit and criticize them to ensure that they are making decisions that are accurate, fair, and unbiased. However, unpacking these "black boxes" may not necessarily facilitate fair use. It might be difficult for AI users to oversee and fully understand the workings of AI tools, even when efforts are made to increase transparency. Furthermore, making AI tools explainable does not necessarily guarantee fairness in all situations, especially in real-world contexts. This is echoed by Felzmann et al. [9, p. 1] as they state, "*The complexity of transparency for automated decision-making shows tension between transparency as a normative ideal and its translation to practical application*". Birhane [10] and Selbst et al. [11] argued that any context that surrounds an AI tool where it is deployed is abstracted away when the interest is narrowly bounded around the AI tool. Consequently, a lack of understanding of the social aspects in socio-technical systems, coupled with a focus on technical solutions, results in the generalized treatment of the social context without acknowledgment of local situated practices where benefits and harms are unevenly distributed.

3. Attempts to Deal with Fairness and Unfairness in Society

In society, ongoing discussions revolve around the impact of AI on human agency, as these tools increasingly shape our decision-making and actions. This prompts inquiries into accountability, specifically questioning whether humans or AI tools should bear responsibility for outcomes [2]. The evolving relationship between humans and machines has also sparked political debates on how to ensure fairness and justice [12]. The approach to addressing this issue in society can be categorized into three main areas: 1) ensuring fairness through the design of AI, 2) legislation regulating the use and implementation of AI to ensure fairness, and 3) ethical principles promoting and restricting the development and implementation of AI. While we will briefly outline these three categories, our main critique encompasses all of them. Even though generic models of ethics are useful, laws and ethical principles provide a foundation and serve to regulate actions. In situations where human judgment is involved, individuals are required to interpret, adapt, and apply these regulations in ways that are as appropriate as possible to each specific situation.

3.1 Design

Frequently, one central effort is the focus on engineering fairer and more just algorithms and models by using fairness itself as a property of the AI tool. In the field of AI, the concept of algorithmic fairness is commonly used to characterize technological solutions that are intended to mitigate systematic harm or benefits from AI tools [13]. This involves considering three stages: before, during, and after the computer processes information [14]. First, pre-algorithmic procedures focus on mitigating biases in datasets before using them to train algorithms. The goal is to prevent the algorithm from initially learning unfair biases. Second, in-algorithmic procedures concentrate on modifying the underlying rules of learning by adjusting algorithms' internal rules and processes to ensure fair treatment. The idea is to build fairness directly into how the algorithm works as an integral part of its functionality. Finally, post-algorithmic procedures include taking corrective actions after the algorithm generates solutions or predictions and adjusting these to reduce any potential unfairness that may have emerged during algorithmic processing. Taken together, from this perspective, fairness is about model accuracy and is assessed in mathematical terms, where the model is classified as (un)fair in terms of the measures of undesired bias that can potentially generate discrimination [11], [13], [15].

3.2 Legislation

Countries are intensifying their efforts to implement regulations to protect their citizens while simultaneously fostering innovation. In the wake of the rapid development of AI services, driven by large multinational corporations, who might not have fairness and democracy, but rather return on investments, as their main driver, there is a public response looking for policy and legislation. While this development triggers concern about negative impacts on democracy and human rights, it is also positioned as a way to gain international competitive advantage in export markets. In this regard, governments are introducing binding legislation to secure fair AI and foster long-term development, encompassing various domains such as privacy and data protection law, health law, and consumer law [16]. There is also cooperation around AI and governance and ongoing engagement to achieve consensus between countries and jointly agreed regulations. An illustrative example of such initiatives is the Artificial Intelligence Act, proposed by the European Commission [17], created to regulate high-risk tools. According to the proposal, certain AI tools are considered "high-risk" tools and, as such, obliged to undergo legally binding requirements based on their intended purpose. Certification of compliance must be ensured by the AI tool producer or provider. The AI act regulates the tool, at a particular action, and addresses product safety. Another example is the Convention on the Rights of the Child, ratified as law in many countries, which designates holders of rights and obligations. In this context, children are the holders of rights, and society assumes the holdership of obligations. Various institutions with child involvement have adults tasked with children's rights in different forms, ensuring their rights are respected. AI could potentially play a role in how these rights are addressed or impacted [18], [19].

3.3 Ethical Principles

Another response to the challenges posed by the introduction of AI tools and to prepare for their widespread use has been a turn to ethical principles, which aim to encourage voluntary actions that are aligned with ethical values rather than imposing strict regulations enforced by law. Several initiatives have been conducted to produce principles to sustain human rights and social values and to ensure fairness. These documents are being issued by a range of entities, both from the private sector, such as companies and non-governmental organizations, and from the public sector, including government agencies [20]. This trend highlights an expanding awareness and concern within the international community regarding ethical considerations in the development and application of AI. AlgorithmWatch has mapped the global landscape of guidelines and principles to make automated applications ethically developed and implemented. To date, over 160 documents have been included in the database. Jobin et al. [20] identified in their review of ethical guidelines for best practices emerging around five ethical principles: transparency, justice and fairness, non-maleficence, responsibility, and privacy.

3.4 Summary

Designing AI models, ethical principles, or legislation might be useful to society in many ways. Some of these experiences might even be generically valid to all or most citizens, serving as a sufficient basis for a new tool design. However, not all aspects are covered by these general solutions. It is widely known that the shifting of personal responsibility when using AI tools can lead to decision-making that contributes to both bias and discrimination against individuals. This poses a challenge because AI tools are machines that cannot be held accountable for their decisions under the law. When algorithms are not transparent or easily understood by the public, it leads to a situation in which people have less trust in these algorithms and the organizations or systems that use them. The lack of transparency in algorithms consequently results in a lack of social trust [12]. Therefore, we argue that it is necessary to allow for the local adaptation of AI functionality to address these issues and challenges.

4. AI and Fairness in Education

Education has several qualities that make it particularly relevant to our discussion. In education, there has been a long-term increase in the use of computational technologies in schools. Over time, this use has resulted in the generation of big datasets through educational platforms, digital textbooks, and free-online resources [21]. There is a large and growing interest from technology developers in building AI support for education [22]. Similarly, the EU AI Act proposal categorizes the use of AI in education as high risk, given its potential to determine access to education and shape an individual's future, such as through exam scoring [17]. As education is regarded as central to societal functioning and a human right, there is already a long discussion concerning the fair and equal treatment of students in the educational system. Given the state's commitment to democracy, education is emphasized as a means to implement respect for human rights and the fundamental democratic values on which society is built. Fair and just schools are understood as similar educational opportunities to all regardless of disabilities, gender, family background, or socioeconomic status [24]. In line with this, the role of teachers is to distribute justice and ensure fairness in their everyday work when planning and preparing classroom instruction. As this implies, AI tools must work in line with these underpinning values or, at the very least, not work against them. Children are obliged by law to go to school and cannot "opt out" from monitoring AI tools and unfair treatment [23]. Therefore, our examples will zero in on a concern related to AI in the realm of work and public authority: ensuring fairness in educational settings.

This calls for closer attention and points to a study in which researchers and teachers worked in close collaboration for several months to integrate an adaptive digital textbook with AI functionality in classroom practices. To effectively illustrate the challenges concerning fairness situated in practice, we will use the scenario presented below. The constraints of space here do not permit a more comprehensive empirical presentation, and we believe it is unnecessary to make the arguments presented in this paper. Thus, the following scenario is constructed by us, drawing from solid empirical work, and deriving from the findings of empirical research (see, for example, [25].

Teaching in a classroom with AI-supported digital textbooks: *Imagine a classroom with 25 students, all aged 14, equipped with laptops and using a digital textbook integrated with AI support. The math teacher relies on this digital resource to assign individual math exercises during part of each lesson. As with any classroom, the students possess varying levels of mathematical proficiency, and the teacher strives to treat them fairly and equally by assigning appropriate tasks based on their abilities. The teacher also wants to ensure that the class progresses through the curriculum in line with national standards and has specific goals planned for each lesson to facilitate a logical content progression.*

At the start of a lesson, the class engages in a collective activity focused on today's topic: the equation of a straight line. Afterward, the students work independently with the digital textbook, which automatically adjusts the difficulty level and provides personalized content based on each student's performance. However, the use of a digital textbook has introduced a challenge. Some students have already completed all of the content, while others are struggling to keep up. The high-achieving students find themselves working on tasks that were initially planned for much later, the *"middle" group is progressing as intended, and the low-performing students have barely begun.*

Despite this situation, the teacher, guided by AI functionality, decides to allow the high-achieving students to work ahead of the class. Additionally, the digital textbook's dashboard highlights the low-performing students in red, prompting the teacher to provide more support to this group. However, she is concerned that the high-achieving students are increasingly losing motivation due to the lack of attention they receive. The teacher initially planned to facilitate group discussions at the end of class but realizes that it becomes futile when almost half of the class is engaged with different content. Without knowing the exact possibilities for system design, the teacher desires more control over the textbook to ensure that all students work with the same content. Furthermore, she would like the dashboard to assist her in distributing her attention more fairly across the classroom.

This scenario offers a concrete example of the dilemmas teachers face when attempting to implement an AI tool in the classroom. A digital textbook is designed to self-adapt to individual students' levels of progression, with AI manifested as automated adaptivity in an intelligent tutoring system. The teacher encountered challenges aligning the content and instructions delivered by the AI tool, personalized for each student, with their own teaching tailored to class-level interests.

5. Cultural-Historical Activity Theory and Meta-Design

This section provides an explanation of CHAT and Meta-Design and the synergy from their integration.

5.1 Cultural-Historical Activity Theory

Technology-mediated activities can be investigated by Cultural-Historical Activity Theory (CHAT) as a conceptual lens to understand the historical and cultural context of human activity [3]. CHAT originates from the research conducted by Russian psychologists Lev Vygotsky and Aleksej Leontjev during the 1920s and 1930s [3], [26]. Vygotsky played a crucial role in shaping the theory by examining how individuals use tools and signs to engage in goal-oriented actions and transform challenging circumstances [27]. Later, Leontjev expanded Vygotsky's individual perspective to understand how collective actions are driven by a common motive within activities [26]. Engeström further developed the understanding of how activities are organized by emphasizing six different components [28].

Among these components, the *object* assumes a pivotal role within the activity system, as it serves to differentiate and distinguish one activity system from another. The object essentially contains the collective vision shared by the participants and sums up their aspirations and desired outcomes for the activity. Notably, the object is not rigidly predetermined but rather undergoes continuous transformation and refinement by the individuals involved as the activity unfolds. A fundamental driving force behind the object is the presence of a shared motive that resonates with the collective objective [29], [30]. The *subject* encompasses individuals or subgroups whose perspective and position serve as the chosen analytical lens. Within the activity system, the subject wields agency and plays a pivotal role [31]. *Tools*, on the other hand, are mediating artifacts that have a symbiotic relationship with the subject, wherein the tools serve to empower and assist humans in their endeavors [26]. *Rules*, encompassing both explicit and implicit regulations and norms, manifest themselves within the activity system, shaping its dynamics and interactions [27], [31]. Lastly, the *community* comprises individuals and subgroups who share a common object, forging a collective bond and driving collaborative efforts through a *division of labor* [27], [31]. These components in the system are interconnected, meaning that if any of the components change, the activity itself also undergoes a transformation. For example, if people in an organization decide to replace a computer with pens and paper, it will impact what is and is not possible within that context. This, in turn, alters the outcome of the activity.

In other words, the activity should be understood through its embedded dynamic relations encompassing mediated and collective human agency. CHAT focuses on system-wide transformation and identifies and resolves contradictions to achieve its objectives. Contradictions are shaped by tensions over time and can serve as a powerful force for people to bring about change [32]. Technology mediates human actions, and this provides a way to explain relations between humans and the sociotechnical context in which they participate [34]. By this means, fairness is relationally defined and comes into being as a consequence of interactions with tools within the activity. Fairness is a concept that is not static over time or uniform within a given activity, but it is defined by its everyday dynamic movements and is continually revised. Working with fairness involves engaging in an open-ended and long-term process.

Agency, in terms of CHAT, is described as “a transformative and relational process of breaking away from the given frame of action and taking the initiative to transform it” [35, p. 60]. The key instrument in this process is double stimulation, which includes first and second stimuli. The first stimuli are experienced by an individual (or a group) as a problem situation, triggering a conflict of motives between desirable alternatives requiring the courage of deliberate choice. Importantly, individuals are endowed with the power to act when they use an external artifact, a second stimulus, to find solutions to a problem. During this process, they formulate ideas and choose how they want to change the situation [30], [36], [37]. The second stimulus empowers subjects to extend the activity system to fit their needs. In CHAT research, several examples illustrate how participants in interventions can interpret theoretically identified contradictions and take action to change their situations [35], [38], [39]. Educational research, involving both teachers and students, shows a similar pattern [40]– [42].

The scenario outlined in Section 4 illustrates how the teacher experiences a conflict of motives between using the adaptive AI tool for controlling individual student tutoring and maintaining teacher-controlled instruction, including managing collective work. In turn, this can be understood as contradictions that emerge due to the dynamic and complex interactions that occur when an AI tool is used for teaching in a classroom setting. One identified contradiction arises between abstract notions of fairness and the practical application of fairness in specific situations—that is, between ***abstracted fairness and situated fairness***. Here, a conflict arises between two different views of what constitutes fairness in the context of an AI system. On the one hand, fairness is grounded in the design and programming of AI tools to be fair (abstracted fairness). On the other hand, fairness is an ongoing process that takes place within the environment where the AI tool is used, meaning that fairness depends on the interactions between the AI tool and the people or other systems with which it interacts (situated fairness). The other identified contradiction arises between ***self-adaptive tools and human-mediated adaptable tools***. This conflict originates from two different approaches to designing and managing complex digital systems. One approach is that AI tools are designed to be able to modify their own behavior and adapt to changing circumstances without direct human intervention. The other approach is AI tools that are based on flexible processes and workflows and designed to be able to be modified or adjusted by humans in response to changing conditions or new information. Overall, the main difference is the degree of autonomy and control.

Despite the advantages of using CHAT to reveal underlying tensions in terms of contradictions to transform activity systems, it currently needs strategies to empower participants in transforming activities through the re-design of tools, in use time, to ensure situated fairness. While we understand how subjects can form the activity by selecting different tools, we are now facing a new era in which AI systems possess their own agency. When subjects contribute to or more directly shape the activity, it changes how the AI tool is trained, which in turn changes the AI tool and thus the ongoing activity. This poses a significant challenge within the CHAT framework, as the agency of tools conflicts with the core principle that assigns agency to human subjects. Previous research has recognized this challenge and proposed the need for theoretical development to address it [43]– [45]. In this paper, we contribute by embracing the concept of Meta-Design within the CHAT framework. By doing so, we acknowledge that subjects have the agency to modify AI tools, providing a new perspective on the evolving relationship between humans and AI.

5.2 Meta-Design

Meta-Design is a participatory approach to design that empowers end-users to become active co-designers of technologies rather than just passive users [46], [47]. It offers a theoretical understanding of people's desire to take control of technology design in order to transform their activities [48]. Through Meta-Design, users are given the techniques and processes to shape and adapt the technology according to their specific needs and local contexts [46], [47]. Meta-Design's distinctive feature lies in its emphasis on empowering individuals to influence and control technologies, setting it apart from other design methods, including the widely used design-based research in education. Design-Based Research (DBR) and Meta-Design share a focus on the design process, iteration, collaboration with stakeholders, and practical application. However, while DBR primarily focuses on designing learning environments and may involve the development of technologies [49]– [51], Meta-Design places a greater emphasis on user empowerment and control [48]. In the words of Fischer et al. [46, p. 35], "*A fundamental objective of meta-design is to create socio-technical environments that empower users to engage actively in the continuous development of systems rather than being restricted to the use of existing systems*". Meta-Design engages users as continuous co-designers throughout a technology's lifecycle. To achieve effective Meta-Design, technologies should offer customization options. Unlike closed technologies, Meta-Design provides users with tools and structures for tailoring technologies to their needs and transferring control from designers to users. This empowers users to actively contribute to their local objectives, fostering a dynamic, user-centric development process [47]. Shifting from self-adaptive AI tools to adaptable approaches through end-user design empowers users, placing them in control and fostering human involvement based on "Intelligence Augmentation" (IA) [52].

5.2 Incorporating Meta-Design in Cultural-Historical Activity Theory

One of the key distinctions between Meta-Design and CHAT is their different focuses on the design process. Meta-Design stems from a tradition that highlights the active participation of end users in the design of technologies [46]. This dimension can be fruitfully integrated into CHAT as both frameworks share epistemological principles that emphasize the collective nature of activity and the transformative process that emerges from people's perceived challenges, without relying on predetermined solutions. Tensions in terms of contradictions (CHAT) and design trade-offs (Meta-Design) signify participants' need for change and act as catalysts for their motivation to reshape their situation [32], [32], [48]. The occurrence of tensions related to the use of technologies is, in other words, the relevant catalyst for integrating Meta-Design into CHAT, as they are linked to a need that stimulates people's agency and drives change. It is through the collective agency of individuals within their local contexts that enables them to change their situation. Thus, adaptions of technologies are negotiated and adjusted in action in relation to aspects that emerge in the dynamic activity. In both Meta-Design and CHAT, the process by which individuals collectively transform their situation is viewed as a form of learning. However, Meta-Design specifically focuses on the concrete actions of individuals and their involvement in the design process in use time. Therefore, this combination provides a valuable perspective for examining the dynamic interplay between AI-mediated systems and human agency. As Fischer mentions: "*Development and increased use of AI-systems have led to a growing importance of application domain knowledge held by domain experts rather than by software developers, who suffer from a thin spread of application domain knowledge. Another challenge is the need for open, evolvable systems that can adjust to fluctuating and conflicting requirements*" [52, p. 7].

Currently, engineers are designing AI tools for scale based on generic assumptions, requiring them to go beyond the needs of individual users and make choices that facilitate widespread adoption for future users. Along this line, Meta-Design supports the design and evolution of systems with an intended solution (at design time), while determining what will work for each individual user (at use time) [53].

6. Summary Statement and Comparative Overview of CHAT and Meta-Design

In this comparative overview, we highlight the dimensions of CHAT and Meta-Design, summarizing their scope, user involvement, analytical and design-oriented aspects, strengths, and challenges.

Table 1: Overview of CHAT and Meta-Design

Dimension	CHAT	Meta-Design
Scope	Understanding and transforming existing activity systems with their socio-cultural implications. Human-made development is not merely an afterthought but rather consciously integrated into the transformation of activity systems	Provide environments to support users as designers in designing the design process itself, tailoring design according to intended user needs and contexts
User involvement	Multi-voicedness and stakeholder inclusion	User involvement and co-design
Analysis vs. design	Focused on understanding and transforming activity systems	Focused on enriching and transcending existing design practices
Strengths	Contextual analysis for identifying problems and conflicts as sources of change	Architecture and frameworks for users to adapt and evolve the design
Challenge	Methodical strategies for re-designing tools in use time to adapt to local needs	Requiring substantial learning efforts for users

Applying this overview to the previous narrative on a teacher employing Meta-Design enables them to function as *users as designers*, collectively reshaping AI tools to adapt in real time. For example, we observed a conflict of motives between the adaptive AI tools, which are responsible for individual student tutoring, and teacher-controlled instruction, which oversees collective activities. Through the implementation of Meta-Design, educators can empower themselves to address this conflict by collaborating with developers to assume control of the AI tool design process, allowing for customization within their specific context. In practical terms, this might entail creating a teacher's interface within the digital textbook, granting teachers the ability to override system recommendations when necessary. This would enable educators to synchronize the class's progress and assign identical content to all students during group activities if the situation demands it. Consequently, these design modification processes have the potential to resolve identified contradictions within the educational environment and give end-users agency over the AI tools that respond to their needs. The integration of Meta-Design can thereby influence the entire educational activity system, helping stakeholders to collectively transform the situation to align with their unique requirements, thus promoting sustainable learning efforts.

7. Discussion

In this paper, we explore the possibility of integrating CHAT and Meta-Design to facilitate the customization of AI tools to meet people's needs in local contexts. These two frameworks share a common emphasis: the crucial transformation of AI-mediated systems to be responsive to people's needs. Both CHAT and Meta-Design underscore that systems can be adapted and modified over time as people's needs and desires evolve. Additionally, these frameworks emphasize the importance of collective development and local needs in this iterative process.

The growing interest and reliance on designing AI models, ethical guidelines, and regulations to ensure fairness have been questioned, highlighting concerns that this approach may lead to implementations without adequate safeguards to ensure fairness [54]. As Munn [54, p. 2] reasons, this results in "*a gulf between high-minded ideals and technological development on the ground – a gap between principles and practice*". In a review of empirical literature on AI in education, Baker and Hawn [55] identified and discussed causes of bias, revealing many unknowns about how certain groups of students are unfairly impacted by AI tools. Alongside this knowledge gap, Selwyn [56, p. 624] acknowledges the relational property of harm and the need to pay attention to local and specific experiences by individuals: "*Any instance of some people being disempowered and disadvantaged by the implementation of AI technologies in education is accompanied by others being empowered and advantaged. As such, any particular AI technology might appear to work perfectly well, and be of great advantage, for many teachers and students. Nevertheless, for many others, the same technology can simultaneously be experienced in harmful ways.*"

Thinking broadly about power and control and focusing specifically on how AI models affect certain individuals or communities, there is a need to engage with often overlooked (minority) groups affected by AI. This emphasizes a more inclusive and collaborative approach in the development and deployment of AI to address the diverse needs and concerns of affected communities [54]. In alignment with this perspective, there is a call for critical reflections on the implications of AI implementation for teachers and students in current classrooms. This call emanates from the observation that teachers and students have not been given the possibility to sufficiently influence the design or implementation of new technologies as active participants in a development process. Instead, they are often provided a space and role to act as facilitators for interaction [57]. A valuable lesson learned from this situation could be to integrate AI evaluations into discussions around pedagogy, curricula, the role of teachers working with automated tools, and agency. This integration can be achieved through interdisciplinary dialogues in collaboration with stakeholders such as teachers, students, and parents [58]. However, even if awareness of this might be a good starting point, it raises a nontrivial question of how this should be operationalized locally in schools and classrooms.

Combining CHAT and Meta-Design can address and complement different aspects of agency in dealing with adaptations to ensure fairness in local contexts. This theoretical extension offers an opportunity for people to adapt AI tools so that they can evolve alongside their changing perceptions of fairness in a given activity. By working together, people can identify their specific needs regarding fairness and design tools that are tailored to their local contexts. This approach enables people to modify and customize these technologies to better suit their needs, shifting from a culture where individuals can choose only from available AI tools to one where they actively participate in the design of these tools. As a result, people are not just consumers of technology but active co-creators. People can thus adapt these tools according to their context of use, providing new opportunities to address fairness with AI tools that cannot be easily resolved through law, design, and policy measures. This theoretical lens offers an approach to empowering people and promoting a more fair and contextually appropriate AI design.

It is essential to acknowledge that there are also recognized risks associated with end-user development. In safety-critical domains, where software reliability and accuracy are of utmost importance, end-user development can pose a danger. Also, rapid changes during the development process may lead to throwaway software and the development of unreliable and unmanageable software. Additional risks include the potential that end users lack sufficient knowledge about system development and an increased vulnerability to hacking attacks. These risks in end-user development can be mitigated through constructive support and end-user education that fosters a sense of responsibility [46].

To empower end-users to influence their circumstances, the design of AI systems must take their needs into consideration. This is particularly justified by end-users' unique insights into the challenges posed by technologies in their activities, a perspective that might differs significantly from that of software engineers. Involving end-users in the design process is therefore not only crucial but also deemed necessary for a comprehensive understanding and sustainable development [52]. Involving end-users in the design process can ensure that AI tools become sensitive to the needs and concerns of the local community and promote fairness. CHAT contributes through its emphasis on context, history, and mediation by tools. This enables a critical analysis of activity systems, helping to reveal underlying contradictions and identify areas where improvements or innovations are necessary. Meta-Design provides design concepts and perspectives that aim to empower participants, allowing them to actively shape the processes of tool design to align with their specific local needs and evolving conceptions of fairness in use-time. This offers an approach to empowering people and promoting more fair AI design.

Acknowledgments. This work was made possible through the support of Marianne and Marcus Wallenberg Foundation. Additionally, the authors acknowledge the support received from the program WASP-HS — The Wallenberg AI, Autonomous Systems and Software Program — Humanities and Society.

CRediT author statement. Johan Lundin: Conceptualization, Writing - Original Draft, Writing - Review & Editing, Supervision, Project administration, Funding acquisition. **Marie Utterberg Modén:** Conceptualization, Writing - Original Draft, Writing - Review & Editing, Investigation. **Tiina Leino Lindell:** Conceptualization, Writing - Original Draft, Writing - Review & Editing, Investigation. **Gerhard Fischer:** Conceptualization, Writing - Original Draft, Writing - Review & Editing

References

1. T. Susnjak, “ChatGPT: The end of online exam integrity?,” 2022, doi: 10.48550/ARXIV.2212.09292.
2. I. Tuomi, *The Impact of Artificial Intelligence on Learning, Teaching, and Education: Policies for the Future*. Luxembourg: Publications Office of the European Union, 2018.
3. Y. Engeström, “Activity theory as a framework for analyzing and redesigning work,” *Ergonomics*, vol. 43, no. 7, pp. 960–974, Jul. 2000, doi: 10.1080/001401300409143.
4. G. Fischer and E. Giaccardi, “Meta-Design: A framework for the future of end user development,” in *End User Development*, H. Lieberman, F. Paterno, and V. Wulf, Eds. Dordrecht, The Nederlands: Kluwer Academic Publishers, 2006, pp. 427–457.
5. S. Sun, Y. Zhai, B. Shen, and Y. Chen, “Newspaper coverage of artificial intelligence: A perspective of emerging technologies,” *Telematics and Informatics*, vol. 53, p. 101433, Oct. 2020, doi: 10.1016/j.tele.2020.101433.
6. A. Birhane, P. Kalluri, D. Card, W. Agnew, R. Dotan, and M. Bao, “The values encoded in machine learning research,” in *2022 ACM Conference on Fairness, Accountability, and Transparency*. Seoul, Republic of Korea: ACM, Jun. 2022, pp. 173–184, doi: 10.1145/3531146.3533083.
7. B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, “The ethics of algorithms: Mapping the debate,” *Big Data & Society*, vol. 3, no. 2, p. 205395171667967, Dec. 2016, doi: 10.1177/2053951716679679.
8. B. Schneiderman, *Human-Centered AI*. Oxford: Oxford University Press, 2022.
9. H. Felzmann, E. Fosch-Villaronga, C. Lutz, and A. Tamò-Larrieux, “Towards transparency by design for artificial intelligence,” *Science and Engineering Ethics*, vol. 26, no. 6, pp. 3333–3361, Dec. 2020, doi: 10.1007/s11948-020-00276-4.
10. A. Birhane, “Algorithmic injustice: A relational ethics approach,” *Patterns*, vol. 2, no. 2, p. 100205, Feb. 2021, doi: 10.1016/j.patter.2021.100205.
11. A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, “Fairness and abstraction in sociotechnical systems,” in *Proceedings of the Conference on Fairness, Accountability, and Transparency*. Atlanta, GA, USA: ACM, Jan. 2019, pp. 59–68, doi: 10.1145/3287560.3287598.
12. G. I. Zekos, *Political, Economic and Legal Effects of Artificial Intelligence: Governance, Digital Economy and Society*, in Contributions to Political Science. Cham: Springer International Publishing, 2022, doi: 10.1007/978-3-030-94736-1.
13. M. Dolata, S. Feuerriegel, and G. Schwabe, “A sociotechnical view of algorithmic fairness,” *Information Systems Journal*, vol. 32, no. 4, pp. 754–818, Jul. 2022, doi: 10.1111/isj.12370.
14. C. Haas, “The price of fairness – A framework to explore trade-offs in algorithmic fairness,” in *40th International Conference on Information Systems, ICIS 2019*. Munich, Germany: Association for Information Systems, 2019.
15. A. Aler Tubella, F. Barsotti, R. G. Koçer, and J. A. Mendez, “Ethical implications of fairness interventions: What might be hidden behind engineering choices?,” *Ethics and Information Technology*, vol. 24, no. 1, p. 12, Mar. 2022, doi: 10.1007/s10676-022-09636-z.
16. N. A. Smuha, “From a ‘race to AI’ to a ‘race to AI regulation’: Regulatory competition for artificial intelligence,” *Law, Innovation and Technology*, vol. 13, no. 1, pp. 57–84, Jan. 2021, doi: 10.1080/17579961.2021.1898300.
17. European Commission (EC), “Proposal for a regulation of the European Parliament and of the council laying down harmonized rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts,” 2022. [Online]. Available: https://artificialintelligenceact.eu/the-act/

18. V. Charisi *et al.*, “Artificial intelligence and the rights of the child: Towards an integrated agenda for research and policy,” Joint Research Centre, Seville, JRC Research Reports JRC127564, 2022. [Online]. Available: https://publications.jrc.ec.europa.eu/repository/handle/JRC127564
19. “AI and child rights policy,” UNICEF, New York, USA, Workshop report, Jun. 2019. [Online]. Available: https://www.unicef.org/globalinsight/media/661/file
20. A. Jobin, M. Ienca, and E. Vayena, “The global landscape of AI ethics guidelines,” *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, Sep. 2019, doi: 10.1038/s42256-019-0088-2.
21. M. Utterberg Modén, “Teaching with digital mathematics textbooks – Activity theoretical studies of data-driven technology in classroom practices,” Doctoral Dissertation, University of Gothenburg, 2021. [Online]. Available: https://gupea.ub.gu.se/bitstream/handle/2077/69472/gupea_2077_69472_1.pdf?sequence=1&isAllowed=y
22. W. Holmes and I. Tuomi, “State of the art and practice in AI in education,” *European Journal of Education*, vol. 57, no. 4, pp. 542–570, Dec. 2022, doi: 10.1111/ejed.12533.
23. B. Berendt, A. Littlejohn, and M. Blakemore, “AI in education: Learner choice and fundamental rights,” *Learning, Media and Technology*, vol. 45, no. 3, pp. 312–324, Jul. 2020, doi: 10.1080/17439884.2020.1786399.
24. “The Education Act,” Swedish Government Offices, SFS 2010:800, 2010. [Online]. Available: https://www.riksdagen.se/sv/dokumentlagar/dokument/svensk-forfattningssamling/skollag-2010800_sfs-2010-800
25. M. Utterberg Modén, M. Tallvid, J. Lundin, and B. Lindström, “Intelligent tutoring systems: Why teachers abandoned a technology aimed at automating teaching processes,” in *54th Hawaii International Conference on System Sciences*, 2021, pp. 1538–1547.
26. B. A. Nardi, “Studying context: A comparison of activity theory, situated action models, and distributed cognition,” in *Context and Consciousness: Activity Theory and Human-Computer Interaction*, B. A. Nardi, Ed. Cambridge, MA: MIT Press, 1996, pp. 69–102.
27. Y. Engeström, *Learning, Working and Imagining: Twelve Studies in Activity Theory*. Helsinki: Orienta-Konsultit Oy, 1990.
28. L. S. Vygotsky, *Mind in Society: Development of Higher Psychological Processes*. Cambridge: Harvard University Press, 1980, doi: 10.2307/j.ctvjf9vz4.
29. Y. Engeström, *Learning by Expanding: An Activity-Theoretical Approach to Developmental Research*. Helsinki: Orienta-Konsultit Oy, 1987.
30. A. Sannino, Y. Engeström, and M. Lemos, “Formative interventions for expansive learning and transformative agency,” *Journal of the Learning Sciences*, vol. 25, no. 4, pp. 599–633, Oct. 2016, doi: 10.1080/10508406.2016.1204547.
31. Y. Engeström and A. Sannino, “Studies of expansive learning: Foundations, findings and future challenges,” *Educational Research Review*, vol. 5, no. 1, pp. 1–24, Jan. 2010, doi: 10.1016/j.edurev.2009.12.002.
32. Y. Engeström, “Expansive learning at work: Toward an activity theoretical reconceptualization,” *Journal of Education and Work*, vol. 14, no. 1, pp. 133–156, Feb. 2001, doi: 10.1080/13639080020028747.
33. G. Fischer, J. Lundin, and O. Lindberg, "Rethinking and Reinventing Learning, Education, and Collaboration in the Digital Age — from Creating Technologies to Transforming Cultures," *International Journal of Information and Learning Technology,* vol. 37, no. 5, pp. 241-252, 2020, doi :10.1108/IJILT-04-2020-0051.
34. J. V. Wertsch, *Mind as Action*. New York: Oxford University Press, 1998.
35. A. Sannino and Y. Engeström, “Co-generation of societally impactful knowledge in Change Laboratories,” *Management Learning*, vol. 48, no. 1, pp. 80–96, Feb. 2017, doi: 10.1177/1350507616671285.

36. Y. Engeström, "Expansive visibilization of work: An activity-theoretical perspective," *Computer Supported Cooperative Work (CSCW)*, vol. 8, no. 1–2, pp. 63–93, Mar. 1999, doi: 10.1023/A:1008648532192.
37. Y. Engeström, "From design experiments to formative interventions," *Theory & Psychology*, vol. 21, no. 5, pp. 598–628, Oct. 2011, doi: 10.1177/0959354311419252.
38. Y. Engeström, "Activity theory and learning at work," in *Tätigkeit – Aneignung – Bildung*, U. Deinet and C. Reutlinger, Eds. Wiesbaden: Springer Fachmedien Wiesbaden, 2014, pp. 67–96, doi: 10.1007/978-3-658-02120-7_3.
39. Y. Engeström and A. Sannino, "From mediated actions to heterogenous coalitions: Four generations of activity-theoretical studies of work and learning," *Mind, Culture, and Activity*, vol. 28, no. 1, pp. 4–23, Jan. 2021, doi: 10.1080/10749039.2020.1806328.
40. T. Leino Lindell, "Teachers' challenges and school digitalization: Exploring how teachers learn about technology integration to meet local teaching needs.," Doctoral Dissertation, KTH Royal Institute of Technology, Stockholm. [Online]. Available: https://www.diva-portal.org/smash/get/diva2:1690709/FULLTEXT01.pdf
41. D. Nussbaumer, "An overview of cultural historical activity theory (CHAT) use in classroom research 2000 to 2009," *Educational Review*, vol. 64, no. 1, pp. 37–55, Feb. 2012, doi: 10.1080/00131911.2011.553947.
42. A. Sannino, "Teachers' talk of experiencing: Conflict, resistance and agency," *Teaching and Teacher Education*, vol. 26, no. 4, pp. 838–844, May 2010, doi: 10.1016/j.tate.2009.10.021.
43. R. A. Allen, G. R. T. White, C. E. Clement, P. Alexander, and A. Samuel, "Servants and masters: An activity theory investigation of human□AI roles in the performance of work," *Strategic Change*, vol. 31, no. 6, pp. 581–590, Nov. 2022, doi: 10.1002/jsc.2530.
44. S. Karanasios, "Toward a unified view of technology and activity: The contribution of activity theory to information systems research," *ITP*, vol. 31, no. 1, pp. 134–155, Feb. 2018, doi: 10.1108/ITP-04-2016-0074.
45. T. Tran, R. Valecha, and H. R. Rao, "Machine and human roles for mitigation of misinformation harms during crises: An activity theory conceptualization and validation," *International Journal of Information Management*, vol. 70, p. 102627, Jun. 2023, doi: 10.1016/j.ijinfomgt.2023.102627.
46. G. Fischer, E. Giaccardi, Y. Ye, A. G. Sutcliffe, and N. Mehandjiev, "Meta-Design: A manifesto for end-user development," *Communications of the ACM*, vol. 47, no. 9, pp. 33–37, Sep. 2004, doi: 10.1145/1015864.1015884.
47. G. Fischer and T. Herrmann, "Meta-Design: Transforming and enriching the design and use of socio-technical systems," in *Designing Socially Embedded Technologies in the Real-World*, 1st ed., D. Randall, K. Schmidt, and V. Wulf, Eds., in: Computer Supported Cooperative Work. , London: Springer, 2015, pp. 79–109, doi: 10.1007/978-1-4471-6720-4.
48. G. Fischer, "End-user development: Empowering stakeholders with artificial intelligence, meta-design, and cultures of participation," in *End-User Development*, vol. 12724, D. Fogli, D. Tetteroo, B. R. Barricelli, S. Borsci, P. Markopoulos, and G. A. Papadopoulos, Eds., in Lecture Notes in Computer Science, vol. 12724. Cham: Springer International Publishing, 2021, pp. 3–16, doi: 10.1007/978-3-030-79840-6_1.
49. T. Anderson and J. Shattuck, "Design-based research: A decade of progress in education research?," *Educational Researcher*, vol. 41, no. 1, pp. 16–25, Jan. 2012, doi: 10.3102/0013189X11428813.
50. S. Barab and K. Squire, "Design-based research: Putting a stake in the ground," in *Design-Based Research*, S. A. Barab and K. Squire, Eds. Psychology Press, 2016, pp. 1–14, doi: 10.4324/9780203764565.
51. F. Wang and M. J. Hannafin, "Design-based research and technology-enhanced learning environments," *ETR&D*, vol. 53, no. 4, pp. 5–23, Dec. 2005, doi: 10.1007/BF02504682.

52. G. Fischer, “Adaptive and adaptable systems: Differentiating and integrating AI and EUD,” in *End-User Development*, vol. 13917, L. D. Spano, A. Schmidt, C. Santoro, and S. Stumpf, Eds., in Lecture Notes in Computer Science, vol. 13917. Cham: Springer Nature Switzerland, 2023, pp. 3–18, doi: 10.1007/978-3-031-34433-6_1.
53. G. Fischer, “User modeling in human-computer interaction,” *User Modeling and User-Adapted Interaction*, vol. 11, no. 1/2, pp. 65–86, 2001, doi: 10.1023/A:1011145532042.
54. L. Munn, “The uselessness of AI ethics,” *AI Ethics*, vol. 3, no. 3, pp. 869–877, Aug. 2023, doi: 10.1007/s43681-022-00209-w.
55. R. S. Baker and A. Hawn, “Algorithmic bias in education,” *International Journal of Artificial Intelligence in Education*, vol. 32, no. 4, pp. 1052–1092, Dec. 2022, doi: 10.1007/s40593-021-00285-9.
56. N. Selwyn, *Education and Technology: Key Issues and Debates*, 3rd ed. London: Bloomsbury Academic, 2022.
57. I. Roll and R. Wylie, “Evolution and revolution in artificial intelligence in education,” *International Journal of Artificial Intelligence in Education*, vol. 26, no. 2, pp. 582–599, Jun. 2016, doi: 10.1007/s40593-016-0110-3.
58. D. Schiff, “Out of the laboratory and into the classroom: The future of artificial intelligence in education,” *AI & Society*, vol. 36, no. 1, pp. 331–348, Mar. 2021, doi: 10.1007/s00146-020-01033-8.