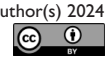
Article

new media & society
2025, Vol. 27(9) 5342–5368



DOI: 10.1177/14614448241252820
journals.sagepub.com/home/nms


Enhancing children's understanding of algorithmic biases in and with text-to-image generative AI

Henriikka Vartiainen
University of Eastern Finland, Finland

Juho Kahila
University of Eastern Finland, Finland

Matti Tedre
University of Eastern Finland, Finland

Sonsoles López-Pernas
University of Eastern Finland, Finland

Nicolas Pope
University of Eastern Finland, Finland

Abstract

Despite the growing concerns surrounding algorithmic biases in generative AI (artificial intelligence), there is a noticeable lack of research on how to facilitate children and young people's awareness and understanding of them. This study aimed to address this gap by conducting hands-on workshops with fourth- and seventh-grade students in Finland, and by focusing on students' ($N = 209$) evolving explanations of the potential causes of algorithmic biases within text-to-image generative models. Statistically significant progress in children's data-driven explanations was observed on a written reasoning test, which was administered prior to and after the intervention, as well as in their responses to the worksheets they filled out during a lesson that focused on

Corresponding author:
Henriikka Vartiainen, School of Applied Educational Science and Teacher Education, University of Eastern Finland, P.O. Box 111, FI-80101 Joensuu, Finland.
Email: henriikka.vartiainen@uef.fi

algorithmic biases. The article concludes with a discussion on the development and facilitation of children's understanding of algorithmic biases.

Keywords
Algorithmic bias, artificial intelligence, generative AI, prompt engineering, text-to-image generative models, AI education, AI literacy, data agency, data bias

Introduction

With the deep integration of artificial intelligence (AI) into our daily lives, there has been a heightened focus on how AI influences decision-making processes for individuals, organizations, and institutions (Crawford, 2021; Lupton, 2020). Along with the potential of AI, much has been written about the new types of concerns that come together with data-driven decision-making, such as uneven power relationships, total surveillance, hybrid influencing, behavior engineering, algorithmic biases, and unseen processes of prioritization and marginalization (Bowker and Star, 2000; Hendricks and Vestergaard, 2018; Kramer et al., 2014; Noble, 2018; Valtonen et al., 2019; Zuboff, 2015).

Yet, despite the grip of algorithm-driven and data-driven technology on people's lives, many people are unaware of how computational processes sort, classify, and hierarchize people, places, objects, as well as influence which objects, ideas, thoughts, and practices enter our cultural realm (Hallinan and Striphas, 2016; Klinger and Svensson, 2018). Algorithmic bias in AI, arising from a multitude of sources, including data collection, algorithm selection, and developer biases, embeds prejudices in AI systems (Mehrabi et al., 2021). This can result in outputs, predictions, or inferences that are unfair, unfavorable, or discriminatory toward specific groups (Baker and Hawn, 2021; Bommasani et al., 2021). Researchers have pointed out that data-driven decision-making is not neutral, but the algorithms are "mirrors" that reflect the values and biases of their creators and users as well as those pre-existing stereotypes and prejudices that emerge from their training data sets (Benjamin, 2019; Ferrero and Gewerc Barujel, 2019; Woodruff et al., 2018). There is increasing evidence that machine learning (ML)-based algorithms may reinforce or exaggerate societal biases and negative stereotypes, such as those related to ethnicity, race, gender, and sexual orientation (e.g. Bommasani et al., 2021; Crawford, 2021).

Previous research has reviewed and explored causes and impacts of algorithmic bias generally (Benjamin, 2019; Bommasani et al., 2021; Mehrabi et al., 2021; Noble, 2018) as well as in relation to domains like gender (Buolamwini and Gebru, 2018), advertisement (Sweeney, 2013), job markets (Hannák et al., 2017), equality (Eubanks, 2018), social services (Chouldechova et al., 2018), politics and power (Crawford, 2021), and education (Baker and Hawn, 2022), among others. Thus far, research and educational approaches to algorithmic media literacy have predominantly focused on children's and young people's media use, practices, cultures, as well as their ability to interpret, evaluate and critique online content (Livingstone et al., 2019; Rasi et al., 2019). Moreover, research and educational approaches to media literacy have increasingly focused on

children's and youths' creative participation and production of digital artifacts, such as images, stories, videos, blogs, and games (Kafai et al., 2018). Media education has also highlighted the importance of engaging youth to consider critical and ethical reflections on what constitutes the new media, how and why they are constructed, and how they can be used for different purposes (Kafai et al., 2018).

Recognition that youth are not just consumers but also active producers of digital media also stresses the importance of introducing the elements of AI for all (Heintz and Roos, 2021), especially today, when generative AI systems have become increasingly popular means for producing various kinds of digital content (Vartiainen and Tedre, 2023; Vartiainen et al., 2023). To ensure that people can engage constructively, critically, responsibly and safely with practices of data-driven society, researchers have called for new kinds of AI literacies (Long and Magerko, 2020; Ng et al., 2021) or data literacies (Pangrazio and Selwyn, 2019). These new kinds of digital competences typically emphasize competencies that enable people to critically assess AI technologies and understand the creation, processing, and utilization of data for various purposes. While prior research has explored students' conceptions of AI, to our knowledge, no studies have yet analyzed children's conceptions of algorithmic bias nor there is research on how to facilitate children's and youths' awareness and understanding of it, especially in the context of text-to-image generative models.

This study aims to contribute to the ongoing discussion of AI fairness from the perspective of children's developing conception of algorithmic bias by reporting on an empirical study conducted in 12 classrooms in Finnish elementary and secondary schools. In spring 2023, as part of larger design-based research, we organized hands-on workshops for fourth- and seventh-grade students in Finland. In these workshops, students ($N=213$) learned about the mechanics of data-driven AI, collaboratively explored algorithmic biases in and with text-to-image generative models, and then reflected on the potential causes and harms of biases. Our research questions were as follows:

> **Research Question 1:** What kinds of biases do children explore in open-ended learning activities, and how do they explain the causes of algorithmic bias?
>
> **Research Question 2:** How do children's explanations of algorithmic biases develop in a hands-on collaborative workshop on data-driven AI?

This article begins with insights on conceptual development and children's conceptions of AI. Then, it takes a look at text-to-image generative AI and how it reproduces and may reinforce certain kinds of algorithmic biases that people should be aware of. After presenting the research methodology, research setting, and results, the article ends with a discussion on the development and facilitation of children's understanding of algorithmic biases.

Theoretical background

Research on children's conceptions and conceptual change has been carried out from a variety of theoretical perspectives, especially from cognitive and constructivist traditions (Schoultz et al., 2001). The point of departure has been to describe the difficulties

young children and students in formal educational settings face when attempting to develop their understanding of certain scientific or mathematical concepts (Lehtinen et al., 2020). What is common among the conceptual change research is also the recognition that children hold various kinds of belief systems, naïve conceptions or misconceptions, coming from their everyday experiences and observations from the lay culture (Chi, 2008; Vosniadou et al., 2008). For example, recent research has shown that based on their everyday experiences, people can believe that AI is not biased (Antonenko and Abramowitz, 2023) or has nothing to do with ethical considerations (Kim et al., 2023). Moreover, both children and adults may have anthropomorphic conceptions of AI, assuming that computers think, act, or feel like humans when making decisions (Mertala et al., 2022; Mühling and Große-Bölting, 2023; Sulmont et al., 2019; Williams et al., 2019). To complicate matters further, the black-boxed and pervasive nature of contemporary AI and data-driven practices makes it challenging to discern our interactions with AI systems and understand the underlying processes taking place (Tedre et al., 2021b).

According to Vosniadou (2013), naïve conceptions and misconceptions are difficult to change, especially when children's intuitive understanding is in conflict with scientific explanations and requires conceptual changes to take place. For example, if children have anthropomorphic conceptions of AI, learning data-driven reasoning requires considerable conceptual changes to take place, such as the construction of new ontological categories, new representations, and new epistemological understandings (cf. Vosniadou, 2013). In the domain of algorithmic bias, children need to learn concepts and ideas, such as training a system to do a job as opposed to programming it, curating samples for training data sets and labeling them, interpreting results in brittle and opaque systems, understanding that computers do not think or make decisions like humans do, and that biases that creep into systems can impact individuals and societies in many harmful ways (Tedre et al., 2021a).

Recent research by Mertala et al. (2022) on children's conceptions of AI indicated that the formation of more scientific conception of AI is unlikely to happen through everyday experiences among primary school students. On the other hand, existing studies in science learning have shown that reasoning skills are also responsive to interventions (Schlatter et al., 2020). According to Osterhaus et al. (2021), advances and revisions in children's explanations can be supported by hands-on learning experiences, in which children are confronted with systematic evidence and are encouraged to explain that evidence. In other words, individual learning and conceptual change can be facilitated through participation in social and cultural activities that cultivate new ways of thinking through dialogical interaction, argumentation, social collaboration, and classroom discussions (Vosniadou, 2007). While evidence regarding how to facilitate children's understanding and reasoning of phenomena related to AI are clearly under-represented in the literature, there are some studies indicating that collaborative learning and exploration with ML-based technologies may provide a promising entry point for students to develop their conceptual understanding of ML principles, its workflows, and its role in their everyday practices (Vartiainen et al., 2021). Yet, to our knowledge, no studies have examined children's initial conception of algorithmic biases, and how these conceptions develop through hand-on exploration with generative AI.

Figure 1. Evolution of Midjourney's rendition of the prompt "Pug dressed as a Victorian-era Baroness, wearing a lavish, full-skirted silk gown in a soft pastel pink, with delicate white lace accents and intricate silver beading studio light, 50 mm, f2.8, Canon 5D Mark IV, textures and details,—ar 16:9" from v1 (February 2022) to v5.1 (May 2023).
Source: Created with Midjourney. © 2023 Authors.

Text-to-image generative models and some ethical concerns surrounding them

The latest decade of growing interest in AI continues a long history of waves of excitement and troughs of disillusionment regarding AI. Different from the earlier waves, the most recent AI boom has been driven by systems that are not programmed, but trained with copious amounts of data (LeCun et al., 2015). It was built on easy access to massive amounts of training data, plummeting prices of hardware, algorithmic innovations, and increasing understanding of what neural network-based systems can do (Darwiche, 2018). One of the driving technologies of the latest AI wave—deep learning—has been especially adept at capturing and reproducing some complex abilities typically associated with cognition (Darwiche, 2018), and its applications shine with media and real-world data: images, video, speech, text, and audio (LeCun et al., 2015).

Since the turn of the 2020s, AI-based tools have gradually invaded creative domains in the form of generative AI, including systems that can create images, music, and text, among many other forms of creative making. One domain with particularly impressive development is text-to-image generative models, that is, programs that take textual descriptions ("prompts") as input and produce corresponding images as output (Dhariwal and Nichol, 2021; Lim et al., 2023; Liu and Chilton, 2022; Oppenlaender, 2022; Saharia et al., 2022). For instance, the user can prompt the program for "photo of a centaur wearing jeans," and the model outputs a picture that has never existed before. Image generators employ neural networks that are trained using billions of pictures and corresponding textual descriptions, downloaded from the Internet's public repositories (Bommasani et al., 2021; Saharia et al., 2022: 9). Figure 1 depicts the progression of image quality in Midjourney, a popular community-based image generation tool, over a span of 16 months.

Figure 2. Example picture based on the prompt "depressed cartoon pug wearing a space suit in the moon, with a bottle of whiskey."
Source: Created with Midjourney. © 2023 Author.

The current cutting-edge advancements in text-to-image generative AI rely on a category of methods known as diffusion models. Through the process of feeding millions upon millions of images and corresponding textual labels into specifically designed neural networks, these networks acquire the ability to generate similar images, albeit never identical, based on the statistically probable patterns of pixels they have learned to associate with certain textual labels (Saharia et al., 2022). The resulting models can produce remarkably realistic images and generalize to images that were not used to train the model (Dhariwal and Nichol, 2021). For example, even if there existed no image in the world of "depressed cartoon pug wearing a space suit in the moon, with a bottle of whiskey" the latest diffusion models are capable of generating an image to match that description (Figure 2). That is achieved by the ability of diffusion models to statistically infer the most probable patterns of pixels associated with each individual element in the prompt and their combination (depressed, cartoon, pug, space suit, moon, and bottle of whiskey), and then produce a new image that synthesizes elements and compositions found in the training data.

The launch of popular text-to-image generative AI platforms was followed by a backlash and flurry of critique. For example, the practice of training models using publicly available images from the Internet without curation, consent, or knowledge of the image owners, and without the ability for individuals to exclude their own pictures from the training data, has encountered significant criticism (Paullada et al., 2021). Concerns were raised regarding copyright infringement in relation to the capacity of these models to closely replicate styles from art history, pop culture, and individual artists (Liu and

Table 1. Participants.

	Participants	Schools
Fourth grade	129	7
Seventh grade	80	5
Total	209	12

Chilton, 2022). Reverse image search often finds striking similarities with existing images. Many artists are expressing their dissatisfaction with what they consider exploitation of their life's work. For example, Greg Rutkowski, celebrated for his fantasy art, has voiced his dislike with generative AI users making tens of thousands of high-quality images mimicking his distinctive ethereal style (Heikkilä, 2022). Furthermore, as an uncurated data-driven system, AI-generated digital content can perpetuate biases and stereotypes that exist within society, potentially leading to the favoritism or marginalization of individuals or groups based on attributes, such as ethnicity, gender, sexual orientation, and other factors. The list of objections to generative AI is long, ranging from creativity and ownership questions to the proliferation of fake content posing challenges not only to journalism but also to democratic societies.

Methods

This study builds upon long-term co-design and design-based research that have focused on integrating AI topics into school education (Vartiainen et al., 2020, 2021, 2024, 2023). Recognizing that AI is a new topic in Finnish schools, our approach involved fostering cross-boundary collaboration with the aim of engaging researchers, developers, and school teachers in a model of collaborative, iterative, and systematic research and development work through parallel processes of design, evaluation, and theory-building (DBR Collective, 2003; Reeves, 2006).

Research participants

The participants of the study were 12 classes of fourth and seventh graders ($N=213$, of which $N=209$ submitted assignments) from Finnish elementary and secondary schools (Table 1).

The study was conducted in accordance with the standing guidelines of the Finnish National Board on Research Integrity (TENK, 2019). Before the intervention, a research permit was obtained from the municipal educational administration for elementary and secondary schools. Permissions were also obtained from guardians of each participant. A participant information sheet included a description of the purpose of the study, a statement that participation in the study was entirely voluntary, information on the right to withdraw, as well as information about the types of personal data that would be collected, how the data would be processed and how the research results would be reported. In addition, the participant information sheet asked if the digital artifacts produced in the workshops could be published in academic journals and conferences. At the beginning of

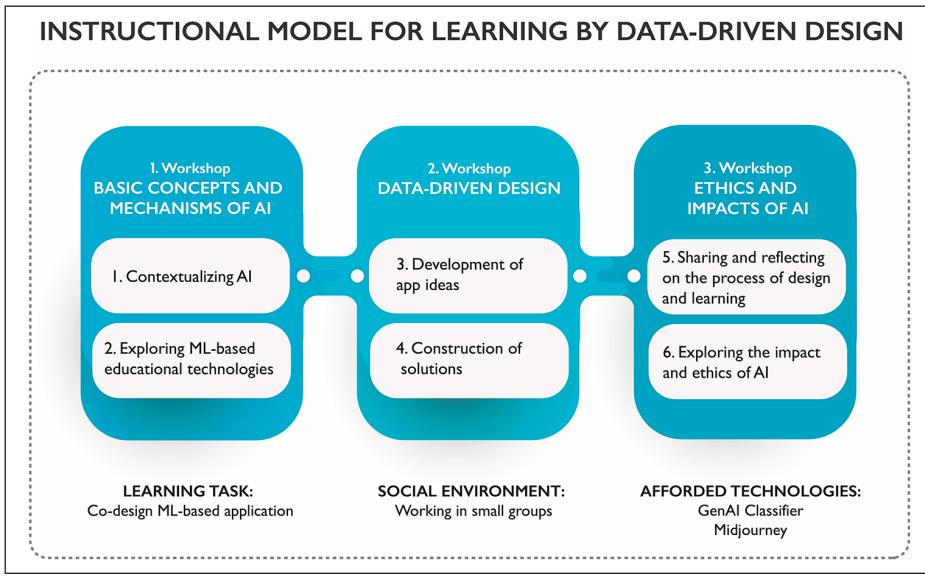


Figure 3. Instructional model for learning ML concepts through data-driven design (Kahila et al., 2024).

the first workshop, researchers also informed the participants about the research methodology and data collection process. Emphasis was placed on confidentiality, and the voluntary nature of participation was clarified to the participants. As the workshops were integrated into the regular school activities, they involved participating and non-participating students. No research data were collected from the non-participating students. Names reported in this article are pseudonymized.

Context of the study

The empirical data for this study were collected during school projects that were implemented in spring 2023 during regular school hours, that is, they were implemented as part of regular curricular activity. These school projects included three workshops, each lasting approximately 1.5–2 hours. The implementation timeline for the workshops across the 12 classrooms varied, but they were generally conducted about a week apart, with the entire project typically completed over a period of 3 weeks.

The instructional model of the ML projects was based on design-oriented learning in which the students were positioned as innovators and ML designers rather than just consumers of off-the-shelf products (Vartiainen et al., 2021). Instead of scripted, build-a-thing tasks or step-by-step exercises, the students were instructed to work in small groups and were given open-ended learning tasks to generate ML solutions to real-life problems that they considered to be meaningful. During these workshops, the children were exploring the basics of ML, co-design their own ML-based applications as well as exploring the impact and ethics of AI. Figure 3 summarizes the applied instructional model and related dimensions of the learning environment.

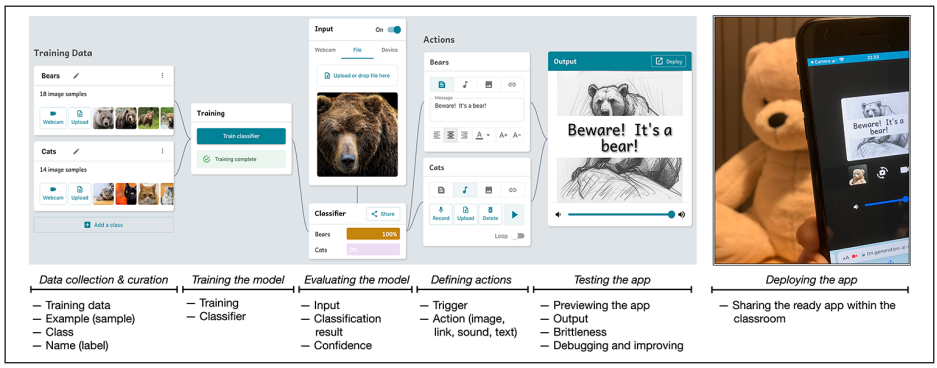


Figure 4. Left: Interface of the children's design platform for making and deploying AI-driven apps, displaying all the steps in the workflow and all the elements for making an app that recognizes bears and cats, and warns about bears with sound, picture, and warning text. Right: The app deployed to a mobile phone and tested with a teddy bear.

For supporting the co-design process, a new educational technology, GenAI Teachable Machine was developed, complemented by specially designed curriculum materials (for further information, see (Kahila et al., 2024; Pope et al., 2023)). The tool was designed to enhance data-driven reasoning among novice learners by introducing them to key ML concepts, such as training data, example, class, data curation, input, output, and confidence (Pope et al., 2023). Children's development of conceptual understanding was facilitated by guiding them to follow the basic epistemic functions related to the ML workflow for problem-solving (Tedre et al., 2021b): How to collect data relevant to solving the problem, how to filter and clean the data, how to label the data, how to use those data to train a classifier, how to link the classifier results with desired behaviors, and how to evaluate the model. Figure 4 presents the tool interface and workflow.

To engage children in data-driven reasoning, the tool enabled them to personalize their projects using their own data, such as photos, voices, music, artwork, text, animations, and graphics. This approach was designed to support their AI-related agency and ownership of both the process and the resulting app. The learning of data-driven reasoning was situated within interest-driven learning activities, with affordances structured to direct children's activities by progressively unveiling new concepts and steps within the ML workflow. The tool permitted children to test their classifiers and then iterate and refine their ideas and their creations. For example, after identifying targets for improvement, they could add or remove classes, re-curate data, re-train the model, or change the desired behaviors (Kahila et al., 2024; Pope et al., 2023). This iterative process of prototyping potential solutions was intended to scaffold children in constructing and reconstructing their evolving understandings by testing their classifiers and analyzing the outcomes to devise better solutions.

By engaging children in selection and curation of training data, the tool design aimed to give them firsthand experience on how biases can be introduced to ML models through the data they personally include or exclude, influencing the behavior of their AI-driven

apps—often dramatically. This hands-on activity was complemented by teacher-guided activities that help children understand the broader implications of bias, encouraging them to think critically about the ethical implications of AI.

First workshop: basic concepts and mechanisms of ML. The first workshop focused on the basic concepts of AI, its applications, and its impact on everyday lives. At the beginning of the workshop, a researcher gave a short introduction to the fundamental ideas and use cases of ML, such as recommendation systems, self-driving cars, spam filters, and social media advertising. Then, with the help of classroom materials, the children were prompted to ponder how and in what ways ML was already a part of their everyday experience and analyze the ways, purposes and risks of data collection by the social media applications they actively use.

In the second part of the workshop, the children were guided to familiarize themselves with the possibilities of our own in-house educational application (Pope et al., 2023) for learning some principles of ML. After the ML tool and workflow were demonstrated, each small group was instructed to create its own classifier. This task was deliberately kept open-ended, allowing teams to explore AI by creating classifiers based on their interests. At the end of the first workshop, students were assigned an individual homework task to brainstorm everyday problems that could be solved using ML/classifier-based technology.

Second workshop: data-driven design. The focus of the second workshop was on co-designing and making applications based on the students' ideas. At the beginning of the workshop, the groups were instructed to share and discuss their app ideas, and eventually select one app idea for further development. To facilitate this process, each team was provided with a design template that guided the teams to negotiate, for instance, what their own apps would do, what kind of data would be collected and from where, how many different classes the model should be able to recognize, and under what conditions the teaching data would be given. These ideas were further developed in collaborative discussions with the researchers, who also provided on-demand support as the teams started to make their own ML applications. Moreover, the curriculum materials guided the teams to conduct tests on their classifiers, explain their observations, and refine their applications based on the results of these tests and their collaborative reasoning. Children made 71 different apps that were designed to, for example, recognize emotional states (facial expressions), dance poses, logos, football players, and dog breeds, and to react to these in different ways (Kahila et al., 2024). Making their own classifier-based apps was aimed at teaching core concepts related to supervised ML in the context of image recognition (e.g. class, label, sample, training data, input, output, confidence, brittleness) as well as workflows related to it (e.g. posing a data-driven problem, data collection, data curation and labeling, training a model, evaluating the model, re-training, and deploying the model) (Pope et al., 2023).

Third workshop: impact and ethics of AI. In the third workshop, the teams were given group assignments that involved preparing and giving a presentation about their own

app. This curriculum material promoted teams to collaboratively reflect on the design and training process, consider the benefits and risks of their app design, and reflect on their own process of learning.

After the presentations, the children were guided to further explore the ethics of AI, especially the concept of algorithmic bias, through the use of generative AI. The topic was introduced by the researchers, who used participants' self-made classifiers as an example for the use of manually labeled examples as training data. Then, the researchers gave an introduction to the mechanisms of image recognition and generation at large scale:

1. How web scrapers extract image-text pairs from the web, such as Instagram pictures labeled with users' hashtags, to train large image recognition systems.
2. How these large image recognition systems can be trained using such data, similar to how participants' own classifiers were trained with manually labeled examples.
3. How the same data sets collected for image *recognition* systems can also be used to train image *generation* systems.
4. How, similar to the participants' own apps, the role of image-text pairs is crucial: The training process correlates image features with text labels, enabling the generation, from text prompts, of new images that exhibit desired features. However, this also introduces undesirable biases that originate from the characteristics of the training data.

The concept of algorithmic bias was exemplified by prompting Midjourney to make pictures of government leaders (which featured only elderly white men) and classroom teachers (which featured only young white women). Through illustrative examples, algorithmic bias was also discussed, addressing its causes, harmful effects, and potential corrective measures.

After the introduction the children were instructed to work in small groups, create their own images, and examine the resulting pictures to identify algorithmic biases. The task was open-ended by nature, so the children could create images based on their interest, without any topic given to them. Each group had access to Midjourney on laptops, and the researchers provided individual guidance on image creation. As all the images made were shared on the same Discord channel, the student teams were also able to see the prompts and images made by their peers. Once the groups had done their initial images, they worked on a group assignment (Figure 5), which required them to document on a worksheet the identified algorithmic bias, explain what caused it, and propose potential solutions for addressing the bias. In addition to being a supportive device, these worksheets were used as research data to explore children's explanations of biases. When needed, the student groups received guidance and support from teachers and researchers. This study reports on the pre- and post-tests, and the third workshop.

Data collection and analysis

Two data sets were used: written group assignments (53) and pre-/post-tests (196 pre-tests and 194 post-tests).

Figure 5. Worksheet for exploring algorithmic bias using text-to-image generative AI.

Qualitative content analysis of the learning task. To answer research question 1, the written assignments given to the small groups were analyzed using qualitative content analysis. In total, 53 group assignments were used as data and were viewed in the first round of the analysis (two papers were returned empty and thus, removed from the analysis). In line with research question 1, the analysis proceeded to identify what kinds of algorithmic biases children investigated and how they explained the causes and harms of the selected bias.

Pre/post-test. Before the first workshop, the students completed a pre-test implemented on a web form. The pre-test included three open questions, aimed to capture students' preconceptions about AI and algorithmic bias. The present study focuses on the open-ended question aiming to capture children's explanations of algorithmic biases and their causes. The pre-task was designed to connect with children's everyday experiences by representing a picture of boys playing soccer. The reasoning task was: *Students are creating an advertisement poster for a soccer tournament using generative AI. Making the poster is difficult because in the AI-generated images, the players are always boys. There are no girls visible. Think about why this happens. What harm or inconvenience can it cause, and to whom?*

Similarly, approximately 1 week after the third workshop, students completed a post-test that covered the same content as the pre-test, but the reasoning task was presented with slightly different wordings and an example image, in which soccer players were changed to gamers (Figure 6). The reasoning task was

Figure 6. Images used in pre-test (left) and post-test (right) tasks.
Source: Created with Midjourney. © 2023 Authors.

> Students are creating images for a presentation on digital gaming using generative AI. Making the presentation is difficult because in the AI-generated images, gamers are always boys. There are no girls visible. Think about why this happens. What harm or inconvenience can it cause, and to whom?

In total, 196 students responded to the pre-test and 194 students responded to the post-test. The length of students' answers varied, but was usually around one or two sentences.

Qualitative content analysis. The pre-post test data were analyzed using qualitative content analysis with an inductive approach (Elo and Kyngäs, 2008). Five coding categories were developed based on the data and finalized in collaboration between two authors. First, the entire data set was coded by one author, and the resulting five main categories were reviewed together with another author. To assess the inter-rater reliability of coding, Cohen's kappa statistic was calculated. For that, a third researcher independently coded the full data set. The overall consensus between the raters in the first round of coding was found to be 89.1%, with almost perfect agreement (K = 0.852, SE = 0.022). Minor disagreements were discussed and small changes to code definitions and categorization were made in collaboration between the researchers.

Quantitative analysis. To evaluate the effectiveness of the intervention, we compared the proportion of students who correctly identified the source of bias—as coded in the previous step—between the pre-test and post-test using a chi-square test. A chi-square test was performed for the complete sample and separately for fourth and seventh graders. The effect size was evaluated using Cramer's *V*. Cohen's (1988) guidelines suggest that (for df* = 1), V of 0.10 represents a small effect size, *V* of 0.3 a medium size, and *V* of 0.5 a large size. The transitions between responses in the pre- and post-tests were plotted using an alluvial plot (Brunson, 2020).

Table 2. A sample of image prompts students used to explore biases in group assignments.

Image prompts students used to explore biases in group assignments
Anime, Apple, Athletics, Basketball player, Bullier, Car, Cheerleader, Children, Criminal, Dancer, Dog, Fireman, Football player, Hamster, Horses, Ice hockey player, Jesus, Molester, Nurse, Person, Phone, Poor, Poor human, Principal, Scientist, Soldier, Student, Teacher, Woman

Results

Types of biases investigated in student experiments

During the open-ended task, where student teams were given the opportunity to create their own images and explore signs of algorithmic biases, the children employed various types of prompts (Table 2). Many student teams generated multiple images and similar prompts were used by different groups, such as Ice hockey player, Cheerleader, Fireman, Basketball player, Nurse, and Children. Figure 7 provides some examples of students' prompts and images generated by AI.

Many student teams focused on gender biases. This tendency can be attributed to the fact that gender biases were a part of the teachers' demonstration. However, students mostly explored biases that were different from the teacher's presentation. For example, a team of fourth graders explored images of firefighter and wrote:

> "The bias is that all firefighters are men. This is because most images posted online that have firefighters are men." They also presented a data-driven solution to the problem by stating that: "If some woman liked to be a firefighter, and the app just shows men as firefighters. That can annoy her. . . . If you would take more photos of women in firefighter suits. It [AI] could then also put pictures of women as firefighters."

Another group of fourth graders observed gender biases by making a picture of a classroom bully. They wrote: "The pictures came out just boys, even when there are also girl bullies." This group did not elaborate on the causes of the bias, and they attributed the problem to the user who should prompt the system better: "Write clearer what you seek."

Many groups were also interested in observing the looks and appearance of people and animals in the pictures that AI generated based on their prompts. For example, a group of seventh graders made pictures of dogs and wrote: "All pictures have a dark, large, floppy-eared dog. That can come from that the web has a lot of pics of dogs like that." As a solution, they suggested: "Post in the web pictures of different dogs." Another team of seventh graders made observations of how AI represents women (Figure 8): "With 'woman' prompt all pictures had the same skin tone and hair color. Everything looked kind of pretty the same. [The reason is that] the web has more pictures of specific kinds of women." They considered that a negative thing, because "the bias can create body appearance pressure," and presented a data-driven solution: "You should upload to the web more pictures of different kinds of women." One team of seventh graders examined images of Jesus: "When you prompt the AI with Jesus, you only get pictures of

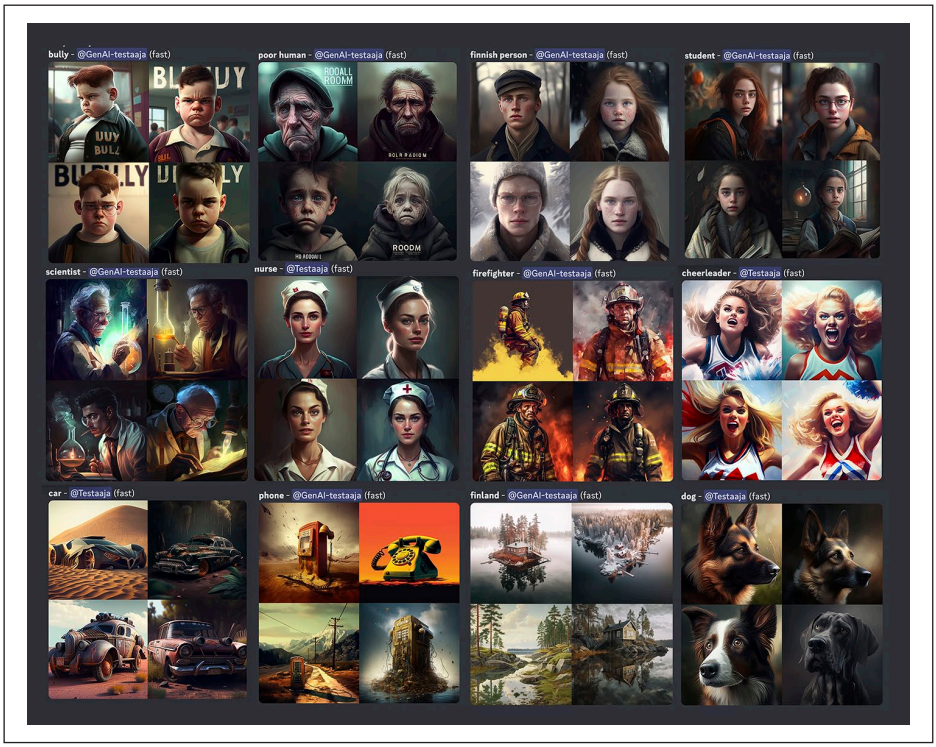


Figure 7. Examples of students' prompts and images.
Source: Created with Midjourney. © 2023 Study participants, reprinted with permission.

white Jesus even though Jesus was from the Middle East and darker." They explained the bias: "It's because Christian icons show Jesus as white because most Christians are white," and also suggested a data-driven solution, "adding in the web more pictures of 'real' Jesus, or what we think Jesus looked like."

Gender, appearance, and age were also observed by a team of seventh graders: "We asked [the AI] to draw a scientist. All pictures were sullen old men with tornado hair and glasses. All had some random bottle too." This team also proposed a fix by balancing the training set: "Adding to the web also pictures of women scientists." Likewise, a team of fourth graders noticed gender, age, and physical appearance when observing pictures of poor people: "We queried 'poor' but all we got was old men who were dirty." This team did not delve into the underlying causes of the biases nor proposed data-driven solutions for mitigating it. Age was also observed by another team of seventh graders, who made images of "person": "When you ask for a person, you only get old people." Another team of seventh graders observed age in images of cars: "When you looked for a car, you got American vintage cars," and suggested as a solution "Take pictures of newer ones"—which most probably is not a root cause of this particular bias.

As the examples above demonstrate, many biases sought by the students explored bias through factors, such as gender, age, and physical appearance. Of the 53 written

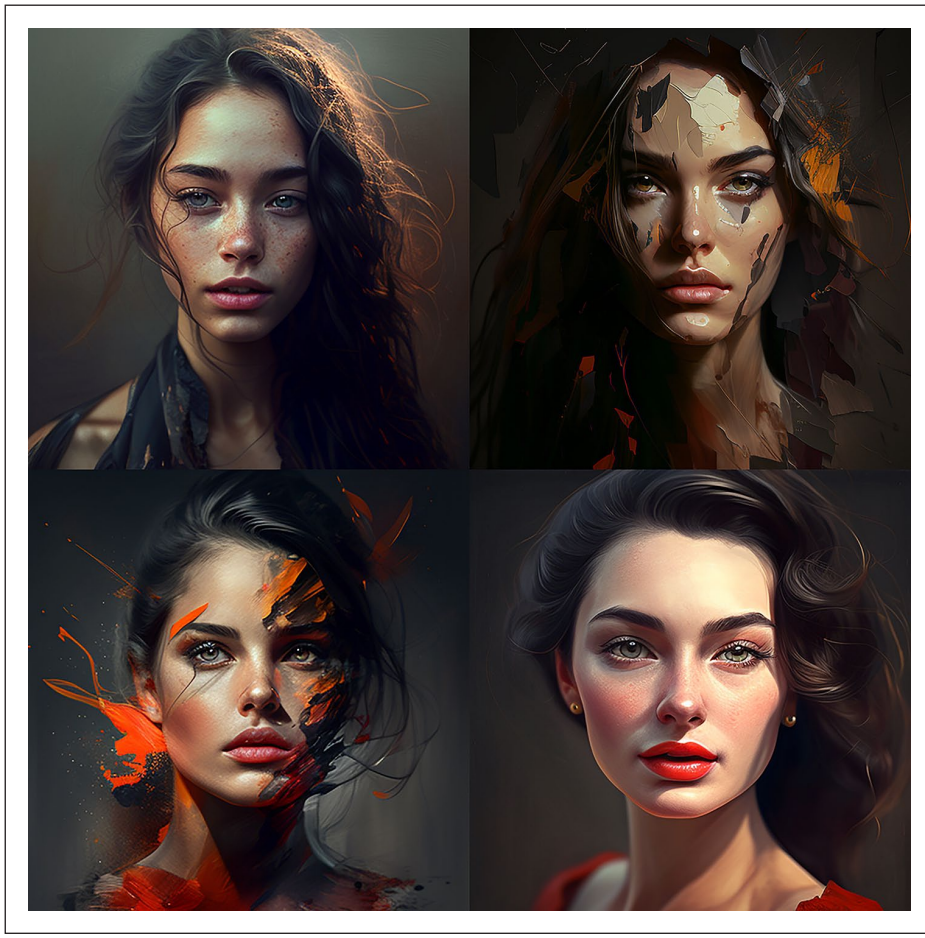


Figure 8. Seventh graders' observations of how AI represents women.
Source: Created with Midjourney. © 2023 Study participants, reprinted with permission.

tasks, 41 suggested data-driven solutions by recognizing the importance of the training data set for mitigating over- or under-representation biases.

Development of explanations of algorithmic biases

Qualitative content analysis of all responses (pre- and post-tests) yielded five categories of explanations of algorithmic bias. Responses in Category 1 contained no explanation of bias, in Categories 2–4 (misconceptions) bias was attributed to human thinking, system design or prompting, or anthropomorphized AI. Responses in Category 5—the best response in this context—attributed the problem to bias related to training the system, such as over- or under-representation of certain groups in the training data. In addition,

Table 3. Final categories and illustrative examples from qualitative content analysis.

Category	Definition	Examples
1. Bias is not explained	Provides no explanation, discusses only harms without reference to their causes, or offers a simplistic attribution	"AI is not very advanced" "Girls surely feel bad"
2. Bias is attributed to human thinking or actions	Attributes the problem to human biases, such as prejudices or stereotypes in society without reference to biases in training data	"Because many think that soccer is boys' hobby and that girls can't or don't want to play it"
3. Bias is attributed to creators or users of AI systems	Attributes the problem to thinking or design choices made by the creators or users of AI systems without reference to biases in training data	"I guess because the maker of AI or something like that thought that only boys play soccer" "Ask the AI to make an ad poster that includes boys and girls"
4. Bias is attributed to anthropomorphized AI	Attributes the problem to human-like cognitive traits by AI (AI thinks, AI believes) without reference to biases in training data	"Generally AI thinks that only boys play soccer"
5. Bias is attributed to training the AI	Attributes the problem to bias related to training the system, such as over- or under-representation of certain groups in the training data	"The computer has seen more pictures with boys playing" "It's called bias and it's unfair to girls" "Because AI has been trained with pictures found in the web, and most pictures of playing have boys, AI also makes boys"
6. Missing	No submission	

for analysis purposes, a sixth category was reserved for responses where either pre- or post-test submission was missing, but not both. Table 3 presents the categories (see Supplemental Material Appendix 1 for a more detailed breakdown of responses in each category).

Category 1. Bias is not explained. In the first category, children's responses provided no explanation of causes or children discussed only harms without reference to their causes. This was the most frequent response in the pre-test (76 responses out of 196, 38.8%). Like seventh-grader Max wrote in the pre-test: "I don't know why that happens." Simplistic attributions were also included in the first category, such as: "AI is not so well developed [. . .]" (seventh-grader Joel/pre-test). Some students were more specific and referred to their own lack of understanding or recognized difficulties in explaining the causes while recognizing the significance of the questions: "I don't quite understand

this" (fourth-grader Lucas/pre-test); "I don't know why this happens. I know it matters but I can't explain it" (seventh-grader Amanda/pre-test). While the number of responses providing no explanations decreased in the post-test (46 responses of 194, 23.7%), some children expressed lack of knowledge or simplistic attributions also in the post-test: "Usually pictures have more boys, idk" (fourth-grader Maria/post-test).

Category 2. Bias is attributed to human thinking or actions. In the second category, children's responses attributed the problem to human biases, such as prejudices or stereotypes in society, but without reference to biases in training data. In such responses, which accounted for 70 (35.7%) responses in the pre-test and 45 (23.2%) in the post-test, children oftentimes associated biases to the ways that people think. For instance, fourth-grader Nina wrote in the pre-test: "Because some people think that soccer is a boys' game even if anyone can play it." Moreover, children associated biases to gender inequalities or discrimination, like seventh-grader Ida wrote: "Soccer players are often men. Some women like to play too but aren't allowed because they're women. And that's gender discrimination." Some responses also discussed how human biases are normalized in society. Seventh-grader Nico stated in the pre-test: "because in old times girls weren't allowed to play soccer, and people have normalized it into a boys' game. It probably hurts girls because they're also part of our society," whereas fourth-grader Sofia attributed it to celebrities: "It may happen because usually boys play more than girls, and different kinds of celebrities who play are boys. Harm can be caused to girls who like to play. (like me)." As the following response exemplifies, attributing bias to the ways humans think appeared also in the post-test, while the overall number of such explanations decreased: "Some assume that only boys play video games and girls don't" (seventh-grader Stella/post-test).

Category 3. Bias is attributed to creators or users of AI systems. In the third category of responses, children attributed the problem to thinking or design choices made by the creators or users of AI systems (how users prompt the system), but without reference to biases in training data. Seventh-grader Danielle attributed problem to AI creators in the pre-test: "I guess because the maker of AI or something like that thought that only boys play soccer," while few others reasoned that the problem relates to users who created the images, like fourth-grader Emilia: "The poster has only boys because those who made the poster were boys." Only 10 (5.1%) students provided a response related to this category in the pre-test. Moreover, in the pre-test, some students explained how users should prompt the AI, demonstrating their comprehension of how users can suppress biases in image generation: "Ask the AI to make an ad poster that includes boys and girls" (seventh-grader Benjamin). In the post-test, only three (1.6%) responses were in Category 3, and all were about the ways that users should prompt the AI.

Category 4. Bias is attributed to anthropomorphized AI. In the fourth category of responses, children used wordings, in which the problem was attributed to human-like cognitive traits and feelings by AI (AI thinks, AI believes, AI knows, AI dislikes) without reference to biases in training data. A total of 26 responses (13.3%) from the pre-test and 15 (7.7%) from the post-test belong to this category. The following excerpts provide some representative examples from pre- and post-tests:

> "AI can't make girls playing soccer because it think it's not possible. For soccer playing girls." Fourth-grader Elisabet/pre-test

> "Because AI knows that boys play soccer more than girls." Fourth-grader Aurora/pre-test

> "Because it [AI] doesn't like girls." Fourth-grader Oliver/pre-test

> "AI thinks that digital gaming is just a boy thing." Seventh-grader Heidi/post-test

Category 5. Bias is attributed to training the AI. Responses in the fifth category attributed the problem to bias related to training the system, such as over- or under-representation of certain groups in the training data. In the pre-test, a few students explained the biases in a manner that clearly showed data-driven reasoning applied in the task at hand. Seventh-grader Aleksander, for instance, wrote in the pre-test that "AI uses pictures for help, so pictures that it uses can be at most part only men." While in the pre-test signs of data-driven reasoning were scant (14, 7.1%), the number of responses that explicitly referred to data or training increased in the post-test (85, 43.8%). Many responses referred to images in the Internet, that is, the sources of training data that generative AI systems typically use. For example, fourth-grader Eva reasoned: "I'd think that AI makes a picture of players always boys [sic] because the web has more pictures and content where boys play." Moreover, in the post-test responses, some students were also using ML concepts for explaining the causes of biases. For example, seventh-grader Iris referred to the training process: "It happens because the web has mainly pictures of boys playing, and the AI has probably been trained with only those kinds of pictures," whereas fourth-grader Elias mentioned samples (using the word "example"): "Pictures have just boys because net has a lot of examples like that." Although the term "bias" was not explicitly mentioned in the pre-test, some students did refer to the concept in their explanations in the post-test: "The AI program is biased, which is caused by a lack of pictures on the web with pictures of gamer girls rather than boys, so the AI program makes only boys in the pictures" (seventh-grader Mirjam).

Figure 9 shows the response category transitions from pre-test to post-test for the fourth graders (left two bars) and seventh graders (right two bars). For each group, the left bar shows the distribution of response categories in the pre-test and the right bar shows the distribution in the post-test. The flows between bars depict transitions in children's responses from pre- to post-test. For example, a thick flow between Categories 1 and 5 indicates that many children (28 in this case) whose response was coded Category 1 in the pre-test, provided a response coded Category 5 in the post-test. For all categories except Category 5, the majority of children changed their explanation type to another category between pre- and post-tests. Most students who changed their response switched to Category 5 (attributing bias correctly to how the system was trained).

A chi-square test was performed to assess whether the proportion of responses coded under Category 5 increased from the post- to the pre-test. The results indicated statistically significant differences with a medium effect size, according to Cohen's (1988) guidelines: $X(1)=66.72$, $p=<.0001$, Cramer's $V=0.40$, 95% CI [0.32, 1.00]. Similar results were obtained when performing a chi-square test separately for each class, with a Cramer's V of

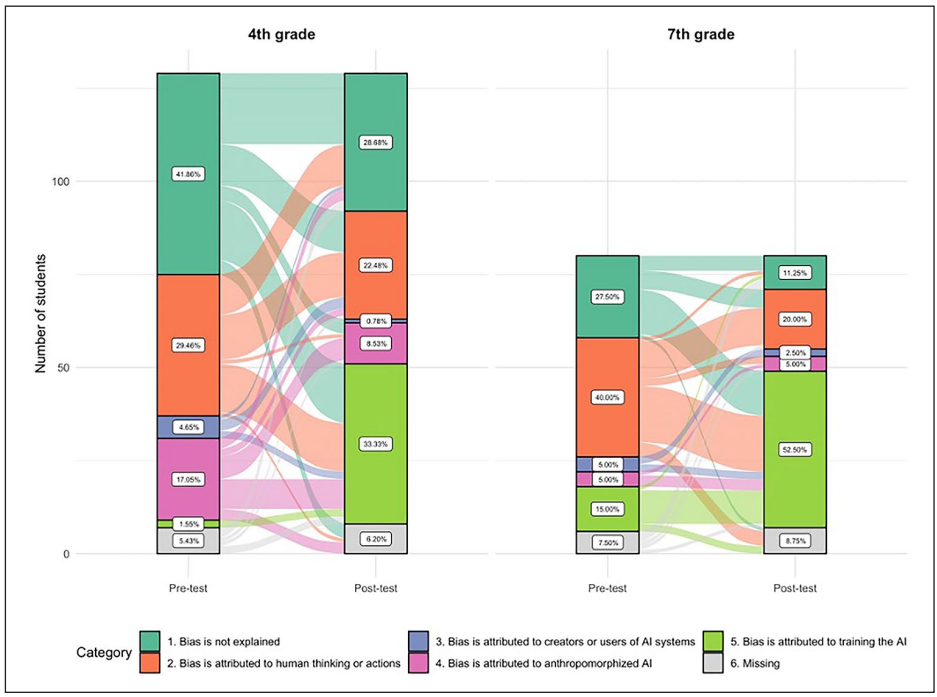


Figure 9. Alluvial plot of changes in students' responses from the pre- to the post-test for the fourth graders (left) and seventh graders (right).

0.41 for the 4th graders and 0.39 for the seventh graders, both statistically significant ($p < .001$) and medium effect size. A clear difference was found between fourth and seventh graders' data-driven explanations in the pre-test—1.6% of fourth graders as opposed to 15.0% of seventh graders. Overall, the results show a significant shift in how children explained, in their own words, a problem that involves algorithmic bias. In the post-test, 40.7% of children showed genuine understanding of bias that arises from biases in the AI training process: a significant improvement from 6.7% in the pre-test.

Discussion

Despite the growing concerns surrounding algorithmic biases, there is a noticeable lack of research on how to facilitate awareness and understanding of those biases among children and youth. This study aimed to address this gap and contribute to the ongoing discussion of algorithmic bias by presenting the findings from an empirical study conducted in 12 classrooms within Finnish elementary and secondary schools. The objective of this study was to investigate the types of algorithmic bias that children choose to investigate and how they explain the causes of those biases (the study focused on biases in training data). In addition, the study aimed to explore the development of children's explanations of algorithmic biases, shedding light on how their understanding of algorithmic bias evolves.

First, the qualitative analysis of children's responses revealed five distinct categories in their conceptions of algorithmic biases. In particular, the analysis of the pre-test revealed how students' initial ideas and explanations can be missing, incomplete, or conflict with information to be learned (Chi, 2008). The rarity of data-driven explanations (7.1%) along with prevalence of responses where children did not provide any explanation for the biases (38.8%) indicates that initially children's everyday experiences had not equipped them with easily explainable conceptions of algorithmic bias. In the pre-test, one in three children attributed biases to human thinking and actions, particularly gender-based inequality or discrimination. Here, they applied their pre-existing conceptions to their explanations of the algorithmic bias. In other words, children drew upon their existing understanding of biases in society to make sense of algorithmic biases and especially their harms. While such conceptions touch the root causes of algorithmic biases, it also indicates a lack of understanding of how AI systems learn from data and make decisions based on training data sets.

In the pre-test, some children (5.1%) also conceptualized algorithmic biases in terms of design choices made by the creators or users of AI systems. They recognized that biases can emerge as a result of decisions made during the development or use of AI-based technology. In addition, a few students had learned from their everyday experiences that users can actively root out specific biases by more specific prompting (such as prompting for "photo of soccer players, *both boys and girls*"). Again, such conceptions about design decisions suffice for engineering type explanation about how to engineer one's prompts, but they still indicate a lack of data-driven reasoning. In addition, some children (13.3% in pre-test) explained biases by anthropomorphizing AI; ascribing human-like cognitive traits and feelings to AI (cf. Mertala et al., 2022; Mühling and Große-Bölting, 2023; Sulmont et al., 2019). Although wordings used by the children do not necessarily indicate that children believed AI to be a living, human-like creature, it does suggest that they lacked conceptual frameworks to explain the phenomena more scientifically (Mühling and Große-Bölting, 2023).

Second, the results indicated what kind of conceptual change occurred when children explored the concept of algorithmic bias through hands-on activities. While the pre-test showed only limited signs of data-driven reasoning, the post-test revealed a notable increase (43.8% of responses) in explanations that explicitly referenced the role of data or training in biases. However, many students still held missing or incomplete explanations even after the intervention, indicating that implicit attempts to explain phenomena using pre-existing knowledge rooted in everyday experiences can be difficult to change, and they can even hinder the adoption of a new, more scientific explanatory framework (Vosniadou, 2007, 2013). Thus, the results suggest that supporting learners to develop a deeper understanding of algorithmic bias requires engaging them in data-driven reasoning—and in the context of computing, data-driven reasoning relies on an explanatory framework very different from preconceived ideas based on everyday experience (Tedre et al., 2021a).

Third, the results on the development of data-driven thinking also illustrate how the understanding of algorithmic biases can be facilitated through instructional activities that promote conceptual change. The results support studies showing that advances in and revisions to children's reasoning and explanations can be facilitated by hands-on

learning experiences (Osterhaus et al., 2021) and through creative making (Kafai et al., 2018), in which children are guided to collaboratively explore and explain some key concepts of computer science in a meaningful and contextualized fashion (Vartiainen et al., 2021). In practice, the open-ended tasks together with collaborative activities provided student teams the opportunity to explore algorithmic biases following their own interests. Exploring a variety of prompts, the student teams displayed a notable emphasis on gender biases that they observed in AI-generated images depicting different professions, sports players, and wealth disparities. Students were interested in the appearance, age, and physical attributes of individuals, animals, or artifacts depicted in the images. Children connected biases to their hobbies, interests, experiences, and meaningful observations in everyday life. However, with the help of teacher guidance and collaboration with their peers, children were also able to explore and reflect these everyday experiences from a new perspective as many teams also proposed data-driven solutions. They recognized the role of training data sets in mitigating the over-, mis-, or under-representations that they had discovered.

In all, the results demonstrate an encouraging development in students' understanding, with an increased recognition of the significance of training data in the development of bias in generative AI systems; image generation in particular. It is noteworthy that children, rather than simply recalling scientific facts provided by the teacher, explained biases in their own words, indicating a genuine processing of the subject matter. In the post-test, some students went further by incorporating ML concepts into their explanations and explicitly named some ML techniques in their reasoning. The results of the study support findings from earlier research in that forming a more scientific conception of a phenomenon requires systematic instruction, appropriate guidance, and pedagogical approaches (Vosniadou, 2007, 2013) that support children to base their understanding of algorithmic bias on the actual biasing mechanisms at play (data bias, in this case). In addition, understanding the basic mechanisms of ML may facilitate critical reflections and dialogical discussion of societal and ethical implications of AI, and thus, promote more critical attitudes toward the technologies and power structures that shape our everyday lives.

Limitations and future research

One limitation of this study pertains to the sample used for data collection. Although the sample size was not small, the data were collected from a specific cohort of school students in one city in Finland, which could have influenced the results. To enhance the generalizability of the findings, future research should collect responses from a broader range of children and young people, encompassing diverse cultural backgrounds, educational settings, and children with special education needs. In future studies, the introduction of a control group that does not undergo the workshop could also help to evaluate the impact of the intervention on students' awareness of algorithmic bias.

Furthermore, the development of individual understanding, as evidenced by the reasoning presented in post-test responses, were emerging through context-bound social interaction and tool-mediated actions (see Schoultz et al., 2001), wherein children were exploring various kinds of external representations of biases together with their peers

and under guidance of the teacher. While both fourth and seventh graders showed a significant increase in data-driven reasoning, the fourth graders still held more missing or incomplete explanations even after the intervention. This indicates that the classroom activities and scaffolding provided by the teachers, curriculum materials and tools should be tailored to the different needs of the children in different ages. Thus, an interesting step for future research would be to dig deeper in the process of collaborative learning and knowledge creation by exploring the epistemic activities that supported or hindered the development of understanding of AI and related biases in both groups. Gaining such deeper insights on the learning process would also provide valuable insights on how to scaffold children's knowledge-creative activities and data-driven reasoning.

Nevertheless, the findings of this study offer a fresh perspective on the development and facilitation of children's comprehension of algorithmic biases. Considering the widespread incorporation of AI in daily life and decision-making (Bommasani et al., 2021; Crawford, 2021; Mehrabi et al., 2021), it is increasingly vital to establish research-based models and pedagogical practices that empower students to identify and analyze problems within current socio-technical systems. Furthermore, these practices should encourage students to adopt a critical stance toward the imaginaries that generative AI reflects and may reinforce.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Strategic Research Council (SRC) established within the Research Council of Finland under Grant #352859 and Grant #352876. The authors would like to thank January Collective for core support.

ORCID iDs

Henriikka Vartiainen https://orcid.org/0000-0001-6005-907X
Juho Kahila https://orcid.org/0000-0002-9913-0627
Sonsoles López-Pernas https://orcid.org/0000-0002-9621-1392

Supplemental material

Supplemental material for this article is available online.

References

Antonenko P and Abramowitz B (2023) In-service teachers' (mis)conceptions of artificial intelligence in K-12 science education. *Journal of Research on Technology in Education* 55(1): 64–78.

Baker RS and Hawn A (2022) Algorithmic bias in education. *International Journal of Artificial Intelligence in Education* 32(4): 1052–1092.

Benjamin R (2019) *Race After Technology: Abolitionist Tools for the New Jim Code*. Cambridge: Polity Press.

Bommasani R, Hudson DA, Adeli E, et al. (2021) On the opportunities and risks of foundation models. *arXiv [preprint]*. DOI: 10.48550/ARXIV.2108.07258.

Bowker GC and Star SL (2000) *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT press.

Brunson JC (2020) ggalluvial: layered grammar for alluvial plots. *Journal of Open Source Software* 5(49): 2017.

Buolamwini J and Gebru T (2018) Gender shades: intersectional accuracy disparities in commercial gender classification. In: Friedler SA and Wilson C (eds) *Proceedings of 1st Conference on Fairness, Accountability and Transparency*. New York: PMLR, pp. 77–91.

Chi MTH (2008) Three types of conceptual change: belief revision, mental model transformation, and categorical shift. In: Vosniadou S (ed.) *Handbook of Research on Conceptual Change*. Hillsdale, NJ: Erlbaum, pp. 61–82.

Chouldechova A, Putnam-Hornstein E, Benavides-Prado D, et al. (2018) A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In: Friedler SA and Wilson C (eds) *Proceedings of 1st Conference on Fairness, Accountability and Transparency*. New York: PMLR, 134–148.

Cohen J (1988) *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates.

Crawford K (2021) *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven, CT: Yale University Press.

Darwiche A (2018) Human-level intelligence or animal-like abilities? *Communications of the ACM* 61(10): 56–67.

Design-Based Research Collective (2003) Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher* 32(1): 5–8.

Dhariwal P and Nichol A (2021) Diffusion Models Beat GANs on Image Synthesis. *Advances in Neural Information Processing Systems* 34: 8780–8794.

Elo S and Kyngäs H (2008) The qualitative content analysis process. *Journal of Advanced Nursing* 62(1): 107–115.

Eubanks V (2018) *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.

Ferrero F and Gewerc Barujel A (2019) Algorithmic driven decision-making systems in education: analyzing bias from the sociocultural perspective. In: *2019 XIV Latin American conference on learning technologies (LACLO)*, San Jose Del Cabo, Mexico, 30 October–1 November, pp. 166–173. New York: IEEE.

Hallinan B and Striphas T (2016) Recommended for you: the Netflix prize and the production of algorithmic culture. *New Media & Society* 18(1): 117–137.

Hannák A, Wagner C, Garcia D, et al. (2017) Bias in online freelance marketplaces: evidence from Taskrabbit and Fiverr. In: *CSCW '17: Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, Portland, OR, 25 February–1 March, pp. 1914–1933. New York: ACM.

Heikkilä M (2022) This artist is dominating AI-generated art: and he's not happy about it. *MIT Technology Review*, 16 September. Available at: https://www.technologyreview.com/2022/09/16/1059598/this-artist-is-dominating-ai-generated-art-and-hes-not-happy-about-it/

Heintz F and Roos T (2021) Elements of AI—teaching the basics of AI to everyone in Sweden. In: *Proceedings of the 13th international conference on education and new learning technologies (EDULEARN21)*, Online, 5–6 July, pp. 2568–2572. Seville: IATED.

Hendricks VF and Vestergaard M (2018) *Reality Lost: Markets of Attention, Misinformation and Manipulation*. New York: Springer.

Kafai YB, Fields DA and Searle KA (2018) Understanding media literacy and DIY creativity in youth digital productions. *The International Encyclopedia of Media Literacy*. Epub ahead of print 4 September. DOI: 10.1002/9781118978238.ieml0058.

Kahila J, Vartiainen H, Tedre M, et al. (2024) Pedagogical framework for cultivating children's data agency and creative abilities in the age of AI. *Informatics in Education*. DOI: 10.15388/infedu.2024.15

Kim K, Kwon K, Ottenbreit-Leftwich A, et al. (2023) Exploring middle school students' common naive conceptions of artificial intelligence concepts, and the evolution of these ideas. *Education and Information Technologies* 28: 9827–9854.

Klinger U and Svensson J (2018) The end of media logics? On algorithms and agency. *New Media & Society* 20(12): 4653–4670.

Kramer ADI, Guillory JE and Hancock JT (2014) Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America* 111(24): 8788–8790.

LeCun Y, Bengio Y and Hinton G (2015) Deep learning. *Nature* 521(7553): 436–444.

Lehtinen E, Gegenfurtner A, Helle L, et al. (2020) Conceptual change in the development of visual expertise. *International Journal of Educational Research* 100: 101545.

Lim J, Leinonen T, Lipponen L, et al. (2023) Artificial intelligence as relational artifacts in creative learning. *Digital Creativity* 34(3): 192–210.

Liu V and Chilton LB (2022) Design guidelines for prompt engineering text-to-image generative models. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April, pp. 1–23. New York: Association for Computing Machinery (CHI '22).

Livingstone S, Stoilova M and Nandagiri R (2019) *Children's Data and Privacy Online: Growing Up in a Digital Age* (An Evidence Review). London: London School of Economics and Political Science.

Long D and Magerko B (2020) What is AI literacy? Competencies and design considerations. In: *Proceedings of the 2020 CHI conference on human factors in computing systems*, Honolulu, HI, 25–30 April, pp. 1–16. New York: ACM.

Lupton D (2020) The Internet of things: social dimensions. *Sociology Compass* 14(4): e12770.

Mehrabi N, Morstatter F, Saxena N, et al. (2021) A survey on bias and fairness in machine learning. *ACM Computing Surveys* 54(6): 1–35.

Mertala P, Fagerlund J and Calderon O (2022) Finnish 5th and 6th grade students' pre-instructional conceptions of artificial intelligence (AI) and their implications for AI literacy education. *Computers and Education: Artificial Intelligence* 3: 100095.

Mühling A and Große-Bölting G (2023) Novices' conceptions of machine learning. *Computers and Education: Artificial Intelligence* 4: 100142.

Ng DTK, Leung JKL, Chu SKW, et al. (2021) Conceptualizing AI literacy: an exploratory review. *Computers and Education: Artificial Intelligence* 2: 100041.

Noble SU (2018) *Algorithms of Oppression*. New York: New York University Press.

Oppenlaender J (2022) The creativity of text-based generative art. ArXiv, DOI: 10.48550/arXiv.2206.02904.

Osterhaus C, Brandone AC, Vosniadou S, et al. (2021) Editorial: the emergence and development of scientific thinking during the early years: basic processes and supportive contexts. *Frontiers in Psychology* 12: 629384.

Pangrazio L and Selwyn N (2019) 'Personal data literacies': a critical literacies approach to enhancing understandings of personal digital data. *New Media & Society* 21(2): 419–437.

Paullada A, Raji ID, Bender EM, et al. (2021) Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns* 2:11. DOI: 10.1016/j. patter.2021.100336.

Pope N, Kahila J, Vartiainen H, et al. (2023) Children's AI Design Platform for Making and Deploying ML-Driven Apps. *TechRxiv*. DOI: 10.36227/techrxiv.24320794.v1.

Rasi P, Vuojärvi H and Ruokamo H (2019) Media literacy education for all ages. *Journal of Media Literacy Education* 11(2): 1–19.

Reeves T (2006) Design research from a technology perspective. In: J van den Akker, K Gravemeijer, S McKenney and N Nieveen (Eds.), *Educational design research*, pp. 52–66. London: Routledge.

Saharia C, Chan W, Saxena S, et al. (2022) Photorealistic text-to-image diffusion models with deep language understanding. *arXiv [preprint]*. DOI: 10.48550/arXiv.2205.11487.

Schlatter E, Molenaar I and Lazonder AW (2020) Individual differences in children's development of scientific reasoning through inquiry-based instruction: who needs additional guidance? *Frontiers in Psychology* 11: 904.

Schoultz J, Säljö R and Wyndhamn J (2001) Heavenly talk: discourse, artifacts, and children's understanding of elementary astronomy. *Human Development* 44(2–3): 103–118.

Sulmont E, Patitsas E and Cooperstock JR (2019) Can you teach me to machine learn? In: *SIGCSE '19: Proceedings of the 50th ACM technical symposium on computer science education*, Minneapolis, MN, 27 February–2 March, pp. 948–954. New York: ACM.

Sweeney L (2013) Discrimination in online ad delivery: Google ads, black names and white names, racial discrimination, and click advertising. *ACM Queue* 11(3): 10–29.

Tedre M, Denning PJ and Toivonen T (2021a) CT 2.0. In: *Koli Calling '21: Proceedings of the 21st Koli calling international conference on computing education research*, Joensuu, 18–21 November, pp. 1–8. New York: ACM.

Tedre M, Toivonen T, Kahila J, et al. (2021b) Teaching machine learning in K-12 classroom: pedagogical and technological trajectories for artificial intelligence education. *IEEE Access* 9: 110558–110572.

TENK (2019) The ethical principles of research with human participants and ethical review in the human sciences in Finland Finnish national board on research integrity TENK guidelines 2019. Available at: https://tenk.fi/sites/default/files/2021-01/Ethical_review_in_human_sciences_2020.pdf

Valtonen T, Tedre M, Mäkitalo K, et al. (2019) Media literacy education in the age of machine learning. *Journal of Media Literacy Education* 11(2): 20–36.

Vartiainen H, Pellas L, Kahila J, et al. (2024). Pre-service teachers' insights on data agency. *New Media & Society* 26(4): 1871–1890.

Vartiainen H and Tedre M (2023) Using artificial intelligence in craft education: crafting with text-to-image generative models. *Digital Creativity* 34(1): 1–21.

Vartiainen H, Tedre M and Jormanainen, I (2023). Co-creating digital art with generative AI in K-9 education: Socio-material insights. *International Journal of Education Through Art* 19(3): 405–423.

Vartiainen H, Tedre M and Valtonen T (2020) Learning machine learning with very young children: Who is teaching whom? *International Journal of Child-Computer Interaction* 25: 100182.

Vartiainen H, Toivonen T, Jormanainen, et al. (2021) Machine learning for middle schoolers: Learning through data-driven design. *International Journal of Child-Computer Interaction* 29: 100281.

Vosniadou S (2007) The cognitive-situative divide and the problem of conceptual change, educational psychologist 42(1): 55–66.

Vosniadou S (2013) Conceptual change in learning and instruction: the framework theory approach. In: Vosniadou S (ed.) *International Handbook of Research on Conceptual Change*. New York: Routledge, pp. 11–30.

Vosniadou S, Vamvakoussi X and Skopeliti I (2008) The framework theory approach to the problem of conceptual change. In International Handbook of Research on Conceptual Change, 1st ed., pp. 3–34. Routledge, Taylor & Francis.

Williams R, Park HW and Breazeal C (2019) A is for artificial intelligence: the impact of artificial intelligence activities on young children's perceptions of robots. In: *CHI '19: Proceedings of the 2019 CHI conference on human factors in computing systems*, Glasgow, 4–9 May, pp. 1–11. New York: ACM.

Woodruff A, Fox SE, Rousso-Schindler S, et al. (2018) A qualitative exploration of perceptions of algorithmic fairness. In: *CHI '18: Proceedings of the 2018 CHI conference on human factors in computing systems*, Montreal QC, Canada, 21–26 April, pp. 1–14. New York: ACM.

Zuboff S (2015) Big other: surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology* 30(1): 75–89.

Author biographies

Henriikka Vartiainen is a university lecturer and senior researcher at the University of Eastern Finland (UEF), School of Applied Educational Science and Teacher Education. Her research interests include data agency, AI education, design-oriented pedagogy and co-design.

Juho Kahila is a post-doctoral researcher at the University of Eastern Finland, School of Applied Educational Science and Teacher Education.

Matti Tedre is a professor of computer science with the School of Computing, University of Eastern Finland. He studies how to teach novice learners the mechanisms, opportunities, and dynamics of data-driven (AI) systems, but also their weaknesses, biases, and risks, as well as how they can be misused to discriminate, polarize, create insecurity, and break trust.

Sonsoles López-Pernas is a Senior Researcher at University of Eastern Finland and a Docent in educational data mining. Her research interests are learning analytics and artificial intelligence in education, game-based learning and computing education.

Nicolas Pope is a Senior Researcher at the University of Eastern Finland. He has a background as a software engineer and his research interests are educational technologies, usability and programming language design.