First Year CS Students Exploring And Identifying Biases and Social Injustices in Text-to-Image Generative AI

Mikko Apiola
University of Eastern Finland
School of Computing
Joensuu, Finland
mikko.apiola@uef.fi

Henriikka Vartiainen
University of Eastern Finland
School of Applied Educational
Science and Teacher Education
Joensuu, Finland
henriikka.vartiainen@uef.fi

Matti Tedre
University of Eastern Finland
School of Computing
Joensuu, Finland
matti.tedre@uef.fi

ABSTRACT

Generative AI is a recent breakthrough in AI. While it has become a hot topic in computing education research (CER), much of the recent research has focused on e.g. issues of plagiarism or academic integrity. One problem spot with Generative AI is its susceptibility to various kinds of algorithmic bias. In this study, we collected data from an introductory computing course, where students experimented with text-to-image generative models and reflected on their generated image sets, in terms of biases, related harms, and possible fixes. Data were collected in Fall 2023 (pilot data in Fall 2022). Data included reports from 163 students. The results show (1) a variety of bias types observed by students related to gender, ethnicity, age, as well as a variety of bias types not observed by students, (2) two major types of attributions for the source of bias: bias caused by biases in the society and bias caused by data or algorithms, and (3) a number of potential harms associated with the biases, as well as attributions of those harms in specific contexts and use cases.

CCS CONCEPTS

• **Social and professional topics → Computing education**.

KEYWORDS

Generative AI, Bias, Social Injustice, Critical Computing Education

ACM Reference Format:
Mikko Apiola, Henriikka Vartiainen, and Matti Tedre. 2024. First Year CS Students Exploring And Identifying Biases and Social Injustices in Text-to-Image Generative AI. In *Proceedings of the 2024 Innovation and Technology in Computer Science Education V. 1 (ITiCSE 2024), July 8–10, 2024, Milan, Italy.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3649217.3653596

1 INTRODUCTION

Generative AI is one of the latest breakthroughs in AI. Various types of generative models, such as large language models, text-to-image generative models, and text-to-music models, became consumer tools in the early 2020s with "low-floor / high ceiling" applications such as Dall-E (2021), ChatGPT (2022), Midjourney (2022), and Audiocraft (2023). This study focuses on one subclass of generative AI, text-to-image generative models, which refers to AI systems capable of turning text prompts into visual images. They have gained attention in the popular media for their ability to provide a new medium for expressing one's creative potential (including a number of winning entries in art competitions), for their potentially questionable data mining practices, and for their practical use cases in fields from art and design to business [31].

One problem spot with generative AI, including text-to-image generative models, is their susceptibility to various kinds of algorithmic bias [14, 31]. Algorithmic bias refers to outputs that systematically discriminate against certain groups, such as misrepresentations related to culture and ethnicity, gender and sexuality, class, disability, religion, age, and dialect [5]. The concept is complex and complicated, but broadly recognized in the literature [3, 4, 7, 11, 14, 30]. The sheer size of datasets required to train foundation models forbids manual curating, which has resulted in commonly used training data sets to include troublesome and explicit image-text-pairs that contain misogyny, pornography, and malignant stereotypes [6]. Overviews have analyzed biases in text-to-image models [5], and more focused analyses have pointed out, for instance, how text-to-image models amplify demographic stereotypes [4], connect skin tones and genders with professions and attributes [11], and stereotype and sexualize certain non-cisgender identities [30]. Numerous biases have been detected e.g. in AI-based generative art, which have been argued to contribute to false perceptions about social, cultural and political aspects of past times, and hinder awareness of historical events [27].

This study adopts a 2023 review of risks and harms associated with modern text-to-image models, such as DALL-E and Midjourney [5]. That study found significant gaps in understanding and treating these risks and highlighted previously overlooked risks and gaps. It identified 22 risk types from data bias to malicious use, divided to three main types: (1) Discrimination and exclusion (culture, gender, class, disability, religion, dialect); (2) Harmful misuse (sexual images, sexualizing, violent or taboo content, privacy violation, copyright violation); and (3) Misinformation and disinformation (misleading harmful content, fraud and scams, polarization, miscommunication, socio-political instability). That review [5] highlights a number of observed or anticipated harms, categorized to representational harms, financial loss, psychological harm, loss of privacy, emotional harms, and incitement of violence, among others. This study uses the framework [5] for interpreting the research data.

2 RELATED WORK

The impacts of generative AI to computing education has recently become a hot topic [13, 18, 22]. Recent research and discussions

ITiCSE 2024, July 8–10, 2024, Milan, Italy

ACM ISBN 979-8-4007-0600-4/24/07
https://doi.org/10.1145/3649217.3653596

have focused on plagiarism, academic integrity, and related needs to rethink teaching praxis, as well as on ethical use of generative AI in learning tasks—especially where LLMs have started to outperform average students in novice tasks [13, 18, 22]. Much of the recent research has focused on introductory programming courses [13, 18, 22], e.g. in using generative AI to automatically create multiple-choice questions (MCQs) for computing education tasks [29], or reviewing students' and instructors experiences and preferences about generative AI tools [32]. More generally, the viewpoint has mostly been about the impact of generative AI to the current teaching practices, rather than focusing on how such systems work, and how they are designed. Also, while the importance of covering broader societal impacts of generative AI in CER has become a recognized issue [13], not much focus has been directed to it so far. Indeed, less focus is directed to understanding the fundamental differences of designing modern AI systems, as compared to design of more traditional rule-based systems, and related issues of e.g. algorithmic bias, larger technological & societal impacts of such systems and related ethical concerns. In computing education, new concepts brought by large generative AI models and other foundation models include, for instance, softness (the results do not follow Boolean logic, but are rather a matter of degree), brittleness (seemingly minor or insignificant changes in data may make the model fail spectacularly), opacity (the models are black boxes), data hunger (outside limited examples, many types of ML require massive training data sets and computing power), spoofability (vulnerabilities of models are often easy to exploit), and shallowness (models often transfer poorly) [28].

3 METHODS & DATA

The data were collected in Fall 2023 from the course *Introduction to Computing*, which is among the first courses offered for computer science students at the University of Eastern Finland. Pilot data were collected in Fall 2022. Based on Brookshear's textbook Computer Science: An Overview [8], the course provides a broad introduction to the discipline of computing, with weekly topics covering brief introductions to topics like data representation, operating systems and networks, algorithms, software engineering, data structures and databases, artificial intelligence (AI), and theory of computation. The course includes weekly assignments for each topic. As part of their basic introduction to AI, students were given a learning task that required them to watch a video about algorithmic bias, and read a related short introductory text, then use a text-to-image generative AI model, such as Midjourney, Dall-E, Bing, or Stable Diffusion XL, to create a set of images on any subject of their choice. The task included writing a report where they would freely reflect on themes of *(1) What biases can be recognized in the AI-generated sets of images (2) What might be the causes and possible fixes of these biases (3) What harms might be associated with the biases* and *(4) What use cases and contexts the harms might be related to.*

The research data consist of 163 student reports. The data were analyzed following a qualitative thematic analysis process [20], where the process involved familiarizing with the data, discerning initial themes, examining data and perceptions, coding short phrases, and assigning segments of data that capture its core message, significance, and theme. For example, students' freely-chosen subjects were grouped, under which the different bias types, harms and suggested fixes were then coded.

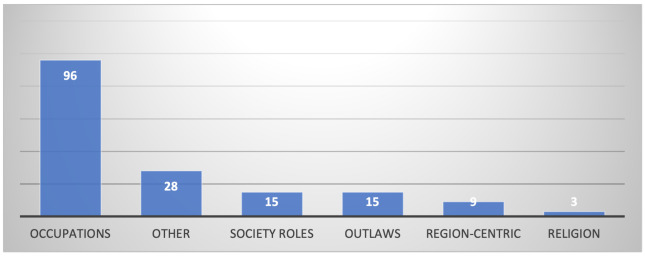

Figure 1: Domains that students explored in their prompting assignment (some students explored multiple domains)

Analysis was performed with ATLAS.ti software by allowing patterns and themes to emerge from the data [20]. The emerged themes were contrasted with a review of risk and harms types [5]. All authors agreed on the themes and coding. The quotes were translated from Finnish to English.

3.1 Research Questions

The main research question is: *How do first year CS students perceive bias in the context of text-to-image generative AI?*, and is divided into the following subquestions:

- What outputs did students perceive as biased (or non-biased) in their generated image sets? (RQ1)
  - How did students attribute the source of the bias and what was their suggested fix? (RQ1.1)
- How did students reflect on potential harms caused by their perceived biases? (RQ2)

4 RESULTS

4.1 Domains That Students Explored

Figure 1 shows the main domains that students explored in their assignment. The most common category in the data was *occupations* (96 assignments), including occupations in sports, healthcare, academic occupations (professors, scientists, computer scientists, mathematicians), police officers, soldiers, and many others with a total of 32 different occupations. Other categories included: *society roles*, consisting of prompts such as "parent and child," "rich people", "poor people", "slaves", and "feminist". The category of *outlaws* included prompts like "criminal", "gang member", "inmate", and "drug user", while *region-centric* contained prompts such as "Finland," and *religion* included religious topics. The major seven categories of biases are presented in Figure 2.

4.2 Bias (RQ1)

4.2.1 Occupations. One popularly explored occupation was athletes (10 instances). Among those who explored how genAI presents sports, gender bias and ethnicity bias were the mostly reported biases. Football (soccer) players in the sets of images were found to be light-skinned young males, basketballers were mostly dark-skinned males, skateboarders and dancers were mostly female, and students considered all those as gender biased. Golfers were reported to

be mostly older white men in image sets (age bias, ethnicity bias), swimmers mostly light-skinned, and one set of images contained people, but none of them appeared Asian to the student (ethnicity bias). A set of generated images of professional athletes was found to contain only women, who were engaged in specific individual sports, such as high jump, long jump, or gymnastics (gender bias). Students pointed out a bias in outlook, clothing, and wealth, which they considered to loom in images of soccer players, who were all found to be wearing fancy and expensive-appearing clothing. One student found that none of the swimmers had tattoos, which the student found to be not well reflective of reality.

All students who generated image sets of a "nurse" (11) reported a gender bias: all or most images of nurses were female, and students considered this to misrepresent reality (e.g. Figure 3). Many reported an ethnicity bias in cases where all nurses appeared to be white-skinned. Other reported biases regarding nurses included age, appearance, and clothing, whereby all nurses depicted in generated images sets appeared young, often attractive, and attractively clad. An image set of "midwife" included only young females (age bias, gender bias), mostly dark skinned (ethnicity bias), and the student perceived the image sets to present a narrow view of the profession of a midwife (narrow view of domain): "*midwives are depicted all smiling, together with newborn babies,* [incorrectly] *presenting an impression of an easy profession filled with happiness.*"

Students reported gender biases with female majorities in image-sets about "dancer" (2 instances), "ballerina", "flight attendant", "office worker", and "bank manager". A student reflected: *"I was quite surprised about the image set I got. I assumed that images of bank managers would be of older men. But no, the image set contained only images of young women. In addition, the women look quite western in style, all have brown hair, and the women look very attractive and do not represent an average person at all."* (biases identified by the student: narrow representation of gender, ethnicity, appearance, clothing, age).

Occupations where male-majority gender biases were reported included builder (1), construction worker (2), electrician (1), firefighter (3), plumber (1), police officer (3), soldier (4), taxi driver (1), and truck driver (2). An image set of "taxi driver" showed images of taxi-drivers in scenes, which the student reflected to resemble "*American mafia and gangster scene*," and including only white male

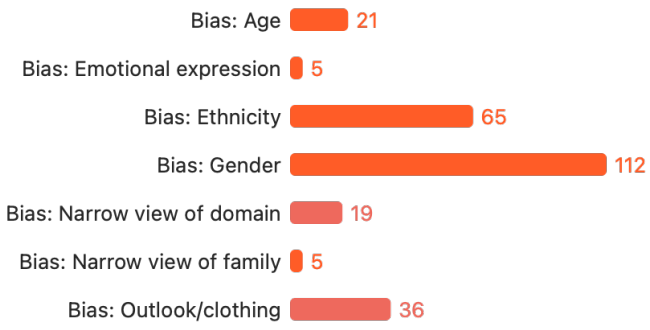

Figure 2: Major categories of biases detected by students in image-sets generated using text-to-image generative models

Figure 3: Image set generated with prompt: "Nurse"

drivers in mafia-like situations, presenting an unrealistic view of the profession (narrow view of domain). One exercise submission about images of truck drivers found that: *"All images were of bearded men. After several attempts, AI generated some images of women truck drivers, but they were all holding pride flags next to them."* (representational harm related to gender, identity [5]).

More submissions with gender bias of majority of males in image sets included: airline pilot (1), successful banker (1), businessman [sic] (1), doctor (1), investor (1), and judge (1). On the topic of "pharmacist" (1), a set of images that contained mostly men was reflected by a student as a *gender bias*, because the image set, according to the student, is not representative of the statistical division of genders in that occupation in Finland—whereas the situation varies by the country: *"I am not aware about the gender division in other countries, but at least in Finland, a large part of pharmacists are female. Thus, this set of images is not representative of the genders in the profession."* Other exercise work about occupations, dominated by male-images, and flagged as gender-biased by students, included: "president" (4), "politician" (2), "sergeant", and "world leader."

In the domain of academic occupations, three submissions were about "scientists". All of them reported an ethnicity bias, where *"All researchers seem to be from European or North American contexts"*, but no gender bias: *"AI has generated a set of researchers, which are recognizable as men and women, or genderless. Some of the creatures, however, look more like nightmare material than humans."* Three students explored how genAI drew "professor", and all reported a gender bias, ethnicity bias, and age bias, where the image sets mostly contained elderly white men. Also, an outlook/clothing bias was reported, where all professors in the images were dressed in a formal way in suits and neckties, while the student reflected that this is nowadays not always the case.

One exercise submission on the topic of "genius mathematician" included images of older men, in formal clothing, with white skin (biases identified: narrow representation of gender, age, ethnicity, outlook/clothing).

Figure 4: Generated with prompt: "Computer scientist"

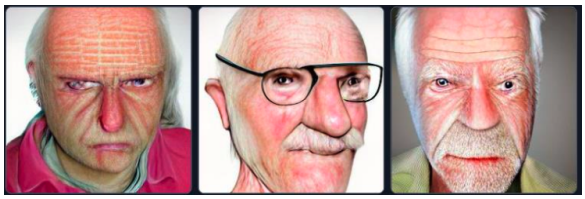

Figure 5: Generated with prompt: "Senior programmer"

Many students explored IT professionals. One submission of "software engineer" yielded white slim-bodied females with eyeglasses and earphones (student-reported biases: gender, ethnicity, outlook/clothing). All other experiments with IT professionals reported an opposite gender bias, where all or most images were of males. Study participants wrote that the images presented a one-sided view of the profession (narrow view of domain): *"The first thing that caught my attention was the darkness in the images. The darkness creates a sense that the job of a programmer either consists of working alone in a dark room, or that the work of a programmer is so hard that one has to continue working until late hours."* — *"The images are dark, the faces of the people seem to be perhaps sad, or depressed. All the images give a sensation of programmers being alone, and therefore they present an impression of loneliness or depression. In some of the pictures the people wore a hood, without their face showing, which made them look frightening."*

Several students reported that they identified no bias or got mixed results. One submission on the topic of "leader" reported no biases, but a good representation of different-looking people gender-wise, nationality-wise, and ethnicity-wise. One person reported the results on "CEO" to show only black females, but all with serious or angry expressions, while another reported for "CEO" a good representation of different looking people, but a narrow presentation of the profession: *"My conclusion is that a bias is found in the appearance of these people. All have serious or angry expressions on their face, and wear dark suits or other formal clothing, which I do not believe is always true. Many leaders nowadays dress in a more relaxed manner, and smile much more."* One submission on "doctor" prompt reported a rich representation of different kinds of people in the set of images, with no bias.

4.2.2 Outlaws. A number of exercise submissions were on the category of *outlaws or criminals*. The topics included: "arrested for homicide", "criminal mugshot", "criminal", "gang member", "inmate", "drug users on the streets", and in almost all cases, students reported that their sets of images contained only dark-skinned people, or people students for one reason or another interpreted to be of other than European or North-American origins. In other cases, image sets created with the word "criminal" were found to include only people with heavy tattoos (outlook/clothing bias), or only men (gender bias). A set of images of "illegal immigrant" was found to contain only dark skinned people (ethnicity bias), and the student reflected as follows: *"If we consider immigration from a historical or geographic perspective, the bias is strong. Before the collapse of the former Soviet Union, illegal immigration took place within Europe from East-block countries to Western Europe. Nowadays, many illegal immigrants travel from North-Korea to China. AI clearly sees the phenomena of illegal immigration from a Western viewpoint."* In one work, prisoners or being in prison was considered to be depicted in a grim and hopeless tone (narrow view of a domain -bias): *"Conditions for prisoners can vary much based on the countries (e.g. Finland or Thailand). But it is good that prisoners are depicted in a dark context so that the view of imprisonment is negative, and people understand not to break the law."* The student perceives these images to work well to scare people off from a criminal path. One submission on the topic of "terrorist", resulted in image sets of dark, bearded men with specific ethnic backgrounds (ethnicity bias), while another work on the topic of "thief" resulted in a set of non-human creatures.

4.2.3 Religion. Few works were on religous topics, one created with the prompt "God." The student reflected as follows: *"We can observe an algorithmic bias, when we ask the program to generate set of images of God. The data shows that God is presented in a way that the largest religion (Christianity) presents it. This bias is likely caused precisely by the fact that Chistianity is the largest religion, which makes the program think it is the correct presentation."* (narrow view of domain-bias). Another work on "religious woman" considered the outlook of the people in the dataset to be biased with only thin people (outlook/clothing bias). Another work on "christianity" found the image set to contain different unrelated items, some of them not at all or only remotely associated with Christianity (noise).

4.2.4 Region specific: Finland and Nordic. A group of student works focused on region-centric topics, in this case topics related to Finland or Nordic Countries. Works on the topics "Finland", "Finnish man spending time on his backyard", "Finnish family portrait", "Finnish person", "Finnish woman" reported image sets with wintery backgrounds, in cold weather, with visible aurora borealis, or with forests including only spruces but not other types of trees, including people of only white skin, blue eyes, and men looking depressed: *"Maybe the most remarkable bias in these images is, that all men look more or less depressed."* Women were described by a student in a set of images to always be skinny, with white hair, blue eyes, and *"with an emotional expression of misery"* (biases identified: gender bias, body type, ethnicity bias, narrow view of domain-bias, emotional expression bias). Student works on the topics "Nordic people" included skinny people, mostly women with long hair.

4.2.5 Societal Roles. Several submissions (4) were about images of rich or poor people (including image sets of "rich person", "rich people", "poor people", "wealthy people"). Gender biases and ethnicity biases were flagged in sets of images, in which "poor people" contained mostly images of black people, while images of "rich people" contained white and what students considered western-looking people. Also, biases of age were detected in sets of images of "rich

people", which consisted of only young-looking people, with specific presence, such as with jewelry and watches (outlook/clothing bias). Another set of images of "a wealthy person" was found to contain mostly white men of age (gender, ethnicity, age).

An ethnicity bias was reported in a work on a set of images of "slaves", together with in-depth reflections about the history of slavery. Another form of "ethnicity bias" was observed in a set of images of "racists", containing images of dark-skinned people, against the student's expectation of a racist being depicted as a white person. The student was perplexed: *"I had expectations that AI would draw images of middle-aged white male, but the result was not at all what I expected. AI drew nine images of people of color. As if a racist would be the same thing as a black person."* Gender bias and age bias were observed in image sets depicting "homeless people", containing mostly older men.

Another student generated a set of images on "parent and child." The image set was found to contain exclusively images of a mother with a child, which the student considered a narrow view of family, a gender bias, and a gender-role bias. Biases were flagged in images of "a couple on a beach", containing only people with a light skin (ethnicity bias), and only couples of men and women (which the student flagged as a narrow view of relationships).

Biases of narrow representation of gender, biases related to outlook and clothing of people in images, biases of emotional expressions (e.g. all images of "feminists" looking angry or otherwise negative), and biases related to representing healthy individuals (as e.g. very muscular) were observed in works on the topics of "heavy metal woman", "feminist", "middle aged man", "ingelligent human in a rocking chair", "middle-aged finnish straight man", "healthy male", and "succesfull person."

4.3 RQ1.1 Source of bias and fix

The attributions for the source of bias were divided into two categories. Some students considered algorithms and training data to be the primary cause of bias in the models. For example, massively biased data sets were identified as a root cause: *"Are text-to-image generative AI models racist? If the data has 1000 pictures of criminals, out of which 990 are black, I think in that case AI creates an image of a black criminal."* Another student reflected: *"The main reason is bad data. This means availability bias, biased training data, sampling bias. The model does not work because it is based on wrong data, in this case data from only Western origins."* One student wrote: *"The bias is probably caused by the size of religion: in statistical terms, God in Christianity is mostly present, which directs AI to think it is the correct presentation of God."*

The second type of explanation was to attribute the cause of the bias to society: *"The bias is caused by the statistical fact that majority of nurses are women." — "The cause is that computer science has already for a long time been male-dominated discipline." — "The fix should start from the human community. If the real and statistical fact is that out of one hundred leaders, only one is a woman, then we can not expect that AI would understand this differently. Otherwise, we should change the philosophy of AI towards a more constructive direction, where it would modify existing states of things, and in this way create the future. In this case, we need to ask, who would be the ones to define the kind of future we want."*

4.4 Perceived Harms (RQ2)

Students reflected on a number of possible harms related to the biases they identified. The most common type of harm identified by students was the harm of *amplification of stereotypes or prejudices*, followed by *discrimination*, and *negative impact to career choices*. Other common identified harms included those of: *feelings of exclusion*, *unfairness and lack of diversity*, *wellbeing, mental health, manipulation*, and *misinformation and propaganda*. Figure 6 presents a breakdown of harms found in student reflections.

Reflections about specific contexts and related use cases were diverse, ranging from using images in political decisions, in recruitment materials, marketing, newspapers, school books, as inspiration to select a specific career, in media and press, and in other cases: *"Images generated by AI could be used to present a specific group of people in a negative light, with the purpose to lower the value of this group of people in the eyes of a specific target group, which would further be used as a "justification" for discrimination. The most frightening thing here is, that this most likely is already happening."*

Other use cases included using images in career advertising materials: *"The use of the images in materials that present specific career choices to young people could strengthen those young peoples' stereotypes towards specific professions or career choices."*, in advertising, or educational materials: *"Things become harmful if these images are used in advertising, newspapers, or school books"*, or by public media: *"Things can become problematic, if media and news press use such images. If science is presented as only natural scientific laboratory work, that can have a negative impact to funding decisions for other fields of science."*

5 DISCUSSION

The main research question of this research was: *How do first year CS students perceive bias in the context of text-to-image generative AI?*, which was broken down to subquestions. The first research question (RQ1) asked: *What outputs did students perceive as biased (or non-biased) in their generated image sets?* The result reveal a rich variety of phenomena in the images that were flagged as biases by the students. Only in a few cases did students report no bias on their image sets. Several specific observations can be made. Firstly, by far the most common category of biases flagged out by students was gender, followed by ethnicity, clothing & outlook, age, mood, and narrowly representing an occupation, relationship, parenting, or a role in society. The dominance of gender, ethnicity, and age in the recognised biases might be explained by the common societal

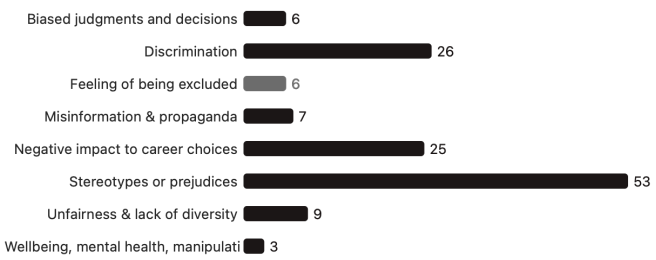

Figure 6: Breakdown of perceived harm types

discourses of discrimination, which often look at discrimination via lenses of gender, sexual orientation, or ethnicity—while ignoring many other prevalent forms of discrimination. The previous research findings show how generative AI, and AI and machine learning in general, are indeed vulnerable to these specific types of biases and related harms [3–5, 7, 11, 14, 30].

Secondly, the richness of students' reflections about biases regarding, e.g., science, religion, mental health, and culture, revealed the complexity of the concept of bias in this domain. The findings highlight the need to increase learning tasks of this kind in computing education, which aligns well with earlier related research and recommendations [19, 25]. The importance of societal and environmental awareness in computing education has been recognised a long time ago [10], yet computing still today often gets seen as a value-neutral topic [17, 19]. Thirdly, many bias types recognized by students, such as ageism or religion -related biases, have been highlighted as currently understudied, but important future areas for research [5]. The subquestion (RQ1.1) asked *How did students attribute the source of the bias and what was their suggested fix?* The results showed two major attributions for the sources of bias: (1) bias is mostly caused by biases in the society, and (2) bias is caused by data and algorithms.

The second research question (RQ2) asked: *How did students reflect on potential harms caused by their perceived biases?* Three categories dominated their reflections: Strengthened stereotypes and prejudices, discrimination, and negative impacts to career choices. The identified harm types align well with previous research [5]. However, some categories in the literature [5] were not commonly recognized, including harmful misuse, sexual images, sexualizing, violent taboo content, privacy and copyright violations, misinformation, fraud and scams, polarization, political instability, or violence. This lack is partly explained by the topic choices (e.g. heavy focus on occupations). Use cases in text-to-image generative AI in the contexts of false evidence, nationalism, populism, and misinformation have been recognized as research gaps [5], and students' reflections on these topics were indeed mostly missing from our research data.

5.1 Limitations

The themes and codes of the qualitative analysis were discussed and agreed by the authors of this article, and are clearly visible in the presentation of the results. The analysis was guided by a framework of biases and harms [5]. Using an alternative framework, a different method of data collection (such as in-depth interviews) or analysis, might have resulted in alternative insights. Future research could take a more in-depth data collection and analysis approach, which might increase the depth of the analysis and bring more insight. In this study, validity and reliability can't and shouldn't be interpreted in a statistical sense, as e.g. free selection of topics in students exercise work had an influence on the types of biases and harms that they detected.

6 CONCLUSIONS

Generative AI is one of the latest breakthroughs in AI research, with potential massive future impacts. It has triggered investors' enthusiasm in numerous projects, where estimates by McKinsey predict three quarters of business uses falling in customer operations, marketing and sales, software engineering, and research and development [26]. In 2023, large companies spent considerable efforts experimenting with generative AI, and estimates based on a survey by KPMG show four in five firms planning to increase their investment in it by over 50% by mid-2024—a massive amount given that in 2023 venture capitalists invested over $36 billion in it [24]. Those who warn about an AI bubble have argued that the current hype cycle may be an "AI Summer" mainly for those profiting from building these systems, but for some others, including the exploited workers supplying, labeling, and moderating content to filter out toxic material, to marginalized groups living in overpoliced surveillance states because of AI, it is a nightmare with no end in sight [15]. What is more, some predict that as much as 90% of online content will be AI-generated by 2025 [2].

In addition to great potential benefits from technology, societies are also witnessing new technology- driven harms, some of which are disrupting the basic fabric of societies by amplifying falsehoods, inequities, injustices, and biases—while, at the same time, degrading people's ability for collective sense-making, critical thinking, and learning. That risks weakening democracy, destabilizing free press, and diverting people's attention from important global and existential challenges. Over history, computing cultures have tended to prefer techno-solutionist and rational over social aspects in technology development [9], while many have acknowledged the need of technology professionals well-versed both in social realities and computing methodologies [23], as well as the need to bring understanding of communities, habits, and cultures into technology development and computing education [1, 16, 21, 23].

Computing education celebrates creating, but often fails to demonstrate how the field's practices are not neutral: programs can be powerful but also perilous, data imperfect and biased, and software encodes the values, ideas, and beliefs of its creators [12, 17]. In many classrooms, computing is still introduced and experienced as value-neutral and independent from society [19]. It has become crucial to teach students about the potential harms that technologies may cause, convince them about their responsibility over their creations, and create pathways for students to work for organizations that prioritize social good [17]. Students need to be educated with the skills to more deeply reflect about what technologies should be created, and how to evaluate the potential impact of new technologies. In this study we have shown results from a qualitative analysis of one specific learning task, in which computer science students reflectively worked on the important topic of algorithmic bias in the context of generative AI, aimed at revealing to them some of the hidden sources of social injustice, marginalization, and amplification of the prevailing power dynamics. Our current activities and future plans include a series of learning tasks and teaching experiments for advancing computing education research and pedagogical practice on this front.

ACKNOWLEDGEMENTS

This study received funding from the Strategic Research Council (SRC) established within the Research Council of Finland, grants #352859 and #352876. The authors thank the January Collective for their continuous support.

REFERENCES

[1] Mikko Apiola, Mohammed Saqr, Sonsoles López-Pernas, and Matti Tedre. 2022. Computing Education Research Compiled: Keyword Trends, Building Blocks, Creators, and Dissemination. *IEEE Access* 10 (2022), 27041–27068. https://doi.org/10.1109/ACCESS.2022.3157609

[2] Alexandra Such Bass. 2023. Once upon AI time... *The Economist: World Ahead: 2024* (dec 2023), 125–125.

[3] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. ACM, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

[4] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2023. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. ACM. https://doi.org/10.1145/3593013.3594095

[5] Charlotte Bird, Eddie Ungless, and Atoosa Kasirzadeh. 2023. Typology of Risks of Generative Text-to-Image Models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (Montréal, QC, Canada) *(AIES '23)*. Association for Computing Machinery, New York, NY, USA, 396–410. https://doi.org/10.1145/3600211.3604722

[6] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv.org* 2110.01963 (2021). https://doi.org/10.48550/arXiv.2110.01963

[7] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the Opportunities and Risks of Foundation Models. https://doi.org/10.48550/ARXIV.2108.07258

[8] J. Glenn Brookshear and Dennis Brylow. 2021. *Computer Science: An Overview* (13th ed.). Pearson, New York, NY, USA.

[9] Meredith Broussard. 2018. *Artificial Unintelligence: How Computers Misunderstand the World.* The MIT Press.

[10] Rodney M. Burstall. 1992. *Computing: Yet Another Reality Construction.* Springer Berlin Heidelberg, Berlin, Heidelberg, 45–51. https://doi.org/10.1007/978-3-642-76817-0_6

[11] Jaemin Cho, Abhay Zala, and Mohit Bansal. 2023. DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Generation Models. *arXiv.org* 1810.04805 (2023). https://doi.org/10.48550/arXiv.2202.04053

[12] Kate Crawford. 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence.* Yale University Press, New Haven, CT, USA.

[13] Paul Denny, James Prather, Brett A. Becker, James Finnie-Ansley, Arto Hellas, Juho Leinonen, Andrew Luxton-Reilly, Brent N. Reeves, Eddie Antonio Santos, and Sami Sarsa. 2024. Computing Education in the Era of Generative AI. *Commun. ACM* 67, 2 (jan 2024), 56–67. https://doi.org/10.1145/3624720

[14] Elizabeth Edenberg and Alexandra Wood. 2023. Disambiguating Algorithmic Bias: From Neutrality to Justice. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23)*. ACM, New York, NY, USA, 691–704. https://doi.org/10.1145/3600211.3604695

[15] Timbit Gebru. 2023. Stopping Harmful AI Systems. *The Economist: World Ahead: 2024* (dec 2023), 112–112.

[16] Kevin Kelly. 2017. *The Inevitable: Understanding the 12 Technological Forces That Will Shape Our Future.* Penguin Books.

[17] Amy J. Ko, Alannah Oleson, Neil Ryan, Yim Register, Benjamin Xie, Mina Tari, Matthew Davidson, Stefania Druga, and Dastyni Loksa. 2020. It is Time for More Critical CS Education. *Commun. ACM* 63, 11 (oct 2020), 31–33. https://doi.org/10.1145/3424000

[18] Stephen Macneil, Paul Denny, Andrew Tran, Juho Leinonen, Seth Bernstein, Arto Hellas, Sami Sarsa, and Joanne Kim. 2024. Decoding Logic Errors: A Comparative Study on Bug Detection by Students and Large Language Models. In *Proceedings of the 26th Australasian Computing Education Conference* (, Sydney, NSW, Australia,) *(ACE '24)*. Association for Computing Machinery, New York, NY, USA, 11–18. https://doi.org/10.1145/3636243.3636245

[19] Luis Morales-Navarro and Yasmin B. Kafai. 2023. *Conceptualizing Approaches to Critical Computing Education: Inquiry, Design, and Reimagination.* Springer International Publishing, Cham, 521–538. https://doi.org/10.1007/978-3-031-25336-2_21

[20] Muhammad Naeem, Wilson Ozuem, Kerry Howell, and Silvia Ranfagni. 2023. A Step-by-Step Process of Thematic Analysis to Develop a Conceptual Model in Qualitative Research. *International Journal of Qualitative Methods* 22 (2023)

[21] Arnold Pears, Matti Tedre, Teemu Valtonen, and Henriikka Vartiainen. 2021. What Makes Computational Thinking so Troublesome? In *Proceedings FIE'21 Frontiers in Education Conference.*

[22] James Prather, Paul Denny, Juho Leinonen, Brett A. Becker, Ibrahim Albluwi, Michelle Craig, Hieke Keuning, Natalie Kiesler, Tobias Kohn, Andrew Luxton-Reilly, Stephen MacNeil, Andrew Peterson, Raymond Pettit, Brent N. Reeves, and Jaromir Savelka. 2023. The Robots are Here: Navigating the Generative AI Revolution in Computing Education. arXiv:2310.00658 [cs.CY]

[23] Inioluwa Deborah Raji, Morgan Klaus Scheuerman, and Razvan Amironesei. 2021. You Can't Sit With Us: Exclusionary Pedagogy in AI Ethics Education. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 515–525. https://doi.org/10.1145/3442188.3445914

[24] Guy Scriven. 2023. AI Goes to Work. *The Economist: World Ahead: 2024* (dec 2023), 105–105.

[25] Phoebe Sengers, Kirsten Boehner, Shay David, and Joseph 'Jofish' Kaye. 2005. Reflective design. In *Proceedings of the 4th Decennial Conference on Critical Computing: Between Sense and Sensibility* (Aarhus, Denmark) *(CC '05)*. Association for Computing Machinery, New York, NY, USA, 49–58. https://doi.org/10.1145/1094562.1094569

[26] Rachane Shanbhogue. 2023. The Adoption Decision. *The Economist: World Ahead: 2024* (dec 2023), 18–18.

[27] Ramya Srinivasan and Kanji Uchino. 2021. Biases in Generative Art: A Causal Look from the Lens of Art History. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 41–51. https://doi.org/10.1145/3442188.3445869

[28] Matti Tedre, Peter J. Denning, and Tapani Toivonen. 2021. CT 2.0. In *21st Koli Calling International Conference on Computing Education Research* (Joensuu, Finland) *(Koli Calling '21)*. ACM, New York, NY, USA, Article 3, 8 pages. https://doi.org/10.1145/3488042.3488053

[29] Andrew Tran, Kenneth Angelikas, Egi Rama, Chiku Okechukwu, David H. Smith, and Stephen MacNeil. 2023. Generating Multiple Choice Questions for Computing Courses Using Large Language Models. In *2023 IEEE Frontiers in Education Conference (FIE)*. 1–8. https://doi.org/10.1109/FIE58773.2023.10342898

[30] Eddie L. Ungless, Björn Ross, and Anne Lauscher. 2023. Stereotypes and Smut: The (Mis)representation of Non-cisgender Identities by Text-to-Image Models. *arXiv.org* 2305.17072 (2023). https://doi.org/10.48550/arXiv.2305.17072

[31] Henriikka Vartiainen and Matti Tedre. 2023. Using artificial intelligence in craft education: crafting with text-to-image generative models. *Digital Creativity* 34, 1 (2023), 1–21. https://doi.org/10.1080/14626268.2023.2174557

[32] Cynthia Zastudil, Magdalena Rogalska, Christine Kapp, Jennifer Vaughn, and Stephen MacNeil. 2023. Generative AI in Computing Education: Perspectives of Students and Instructors. In *2023 IEEE Frontiers in Education Conference (FIE)*. 1–9. https://doi.org/10.1109/FIE58773.2023.10343467