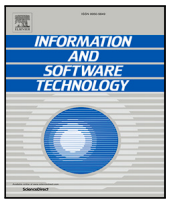
AI systems' negative social impact and factors

Information and Software Technology 192 (2026) 108038

Contents lists available at ScienceDirect

Information and Software Technology

journal homepage: www.elsevier.com/locate/infsof



AI systems' negative social impact and factors☆

Nafen Haj Ahmad [a], Linnea Stigholt [a], Leticia Duboc [b], Birgit Penzenstadler [c,a,d,*]

[a] Department of Computer, Science and Engineering, University of Gothenburg, Gothenburg, Sweden
[b] School of Engineering, La Salle University Ramon Llull, Barcelona, Spain
[c] Department of Computer, Science and Engineering, University of Chalmers, Gothenburg, Sweden
[d] Lappeenranta University of Technology, Lappeenranta, Finland

ARTICLE INFO

Keywords:
Sustainability
AI
Social impact
Ethics
Guidelines
Awareness
Responsibility

ABSTRACT

Context: AI technologies are rapidly being integrated into society, offering numerous benefits but also raising significant ethical and social concerns. While some AI systems aim to improve efficiency and decision-making, they can also cause harmful impacts on individuals and society.
Objective: This study examines both the immediate and systemic negative effects of AI systems, as well as the underlying factors that might contribute to these issues.
Method: Using a multi-vocal literature review, we analyze 28 AI systems and their associated impacts, including discrimination, psychological and physical harm, and unfair treatment.
Results: We identify key factors that might have led AI systems to operate in that manner and explain why these impacts may occur. Additionally, we propose initial concrete actions to mitigate these negative effects and promote the development of AI systems that align with ethical and social sustainability principles.
Impact: By shedding light on these issues, we aim to raise awareness among researchers and developers, encouraging the adoption of more responsible and inclusive as well as concrete AI guidelines.

1. Introduction

The world is currently undergoing a massive wave of digital transformation, with advances in technology that bring undeniable benefits across various sectors. From increased efficiency to new opportunities, digital tools, particularly those driven by Artificial Intelligence (AI),[1] are revolutionizing the way we live and work. However, the widespread adoption of AI also raises complex questions about its impact on society. While AI systems offer numerous advantages, they also bring about unintended consequences, including ethical concerns and social challenges that may not be immediately apparent.

A well-known example of this is the recruitment tool developed by Amazon, which was designed to streamline the hiring process by reviewing resumes and selecting the best candidates. This tool, however, was found to be discriminatory against women candidates, as it had been trained on a men-dominated dataset and thus reinforced gender biases [1]. Cases such as this highlight the importance of anticipating and addressing the potential negative social impacts of AI systems during their development stages.

Despite growing awareness of these challenges, the ethical dilemmas arising from AI can be difficult to predict and address, leaving developers in need of clearer guidance. Codes of ethics, such as those put forth by professional organizations such as IEEE and ACM [2,3], provide foundational principles to inspire developers to take responsibility for their creations and consider their broader societal implications. In addition, the OECD released a set of guidelines in 2019 [4]. However, questions have been raised regarding the effectiveness of these ethical guidelines in adequately addressing the nuances of AI development [5,6].

Hagendorff [7] examined 22 existing AI ethics guidelines and found that none sufficiently addressed the social impacts caused by developers' decision-making. He suggests that we need to close the gap between ethical and technical discourses and encourage individual self-responsibility. In light of this, better guidelines are needed — ones that

☆ This article is part of a Special issue entitled: 'Software and Society' published in Information and Software Technology.

* Corresponding author at: Department of Computer, Science and Engineering, University of Chalmers, Gothenburg, Sweden.
E-mail addresses: gushajana@student.gu.se (N.H. Ahmad), gusstighli@student.gu.se (L. Stigholt), l.duboc@salle.url.edu (L. Duboc), birgitp@chalmers.se (B. Penzenstadler).

[1] We use the definition provided by the EU AI Act: 'AI system' means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments; https://artificialintelligenceact.eu/article/3/.

https://doi.org/10.1016/j.infsof.2026.108038
Received 5 March 2025; Received in revised form 11 January 2026; Accepted 12 January 2026
Available online 17 January 2026

provide a deeper understanding of the real-world impact of AI systems bottom-up.

In this paper, we explore these concerns by identifying the *enabling* and *systemic* effects of AI systems on society. We also examine the factors that may have contributed to these outcomes during the development of these technologies. Hence, this study seeks to answer two research questions:

- **RQ1:** What are the types of negative social impacts that existing AI systems cause?
- **RQ2:** What are the common factors that can cause these negative social impacts?

We discuss the findings in the context of well-known codes of ethics, with an emphasis on the principles *Avoid harm*, *Be fair and take action* and *Respect privacy*. We offer suggestions for making ethical guidelines more concrete and applicable to the evolving landscape of AI development.

We believe that these findings can assist professional associations and companies in making their ethical guidelines more practical and effective. Additionally, they can guide software engineering researchers in further exploring these effects, their underlying causes, and potential strategies for mitigating or avoiding them.

The remainder of this article is organized as follows: Section 2 presents that relevant background and related work, Section 3 describes the research design, Section 4 presents the results, Section 5 supplies the discussion, and Section 6 concludes the article.

2. Background

In this section, we provide the background for the work at hand. First, we present related work on the social impacts of AI in Section 2.1. Then, we present the state of the art on the factors causing such impacts in Section 2.2. Finally, we present current related work on AI Ethics in Section 2.3. This progression is intentional: it moves from observed impacts, to their underlying drivers, to the ethical frameworks often used to address AI impacts. Structuring the section in this way establishes the conceptual foundations required for our analysis and mirrors the logic through which we interpret the data and develop our results in the main part of the article.

2.1. Social impacts of AI

As AI continues to evolve and become more integrated into various domains of life, there are important questions about its broader societal implications. Makridakis [8] discussed the potential for AI to replace human decision-making processes in society, highlighting that although AI is already part of various systems, its involvement is predicted to increase substantially in the future. However, the increasing role of AI raises alarms about the ethical implications of its use. Several studies have explored these potential impacts. Below, we will examine three application domains where AI's social impacts are becoming increasingly evident:

Employment: Makridakis [8] argues that AI systems could increase overall productivity and create new opportunities in some sectors. In particular, AI has the potential to enhance certain professions by augmenting human capabilities and replacing specific tasks in professions, but still requires human involvement, such as doctors or firemen [9]. However, the impact of AI on employment is complex. The reliance on AI systems is expected to create shifts in both employment and decision-making behaviors. Comparing it to the industrial revolution, Makridakis argues that the AI revolution could lead to significant disruption in the workforce, with both job displacement and the creation of new forms of employment [8].

A report by [10] further explores these concerns, specifically focusing on the U.S. labor market. The introduction of automation is predicted to result in higher unemployment rates, particularly among low-wage workers and marginalized groups such as Black and Latin workers. Self-driving vehicles, for example, could displace a significant number of truck, taxi, and delivery drivers. Thus, while AI holds promise for increasing productivity and creating new jobs, it also presents risks of widening inequality and job losses in certain sectors.

Medicine: AI also holds great potential in the medical field, where it can be used to develop personalized treatments and enhance diagnostic accuracy. According to [11], AI systems have already been deployed to diagnose diseases with remarkable success. However, these systems rely heavily on data inputs, which may inadvertently introduce bias if the data sources themselves are flawed or incomplete. Discriminatory patterns in data, such as those based on race, gender, or insurance type, can negatively affect the accuracy and fairness of AI outcomes. For instance, a healthcare algorithm designed to predict ICU mortality was found to exhibit gender and insurance-type biases, demonstrating how flawed data can perpetuate inequalities in critical decision-making processes [11].

Individual: AI systems have the potential to bring significant benefits to individuals, particularly in improving access to services and enhancing personalized experiences. For example, AI can help to improve customer service, provide tailored educational resources, or help personal finance management.

However, AI systems can also have profound implications for individual rights and freedoms. Vesnic et al. [12] argued that the implementation of AI in contexts where human emotions and empathy are involved may inadvertently restrict individual autonomy. When AI systems are used in situations that require personal interaction — such as in healthcare or customer service — there is a risk that people may be manipulated or subjected to interactions that lack genuine human connection. These reductions in meaningful human interactions, according to [12], can lead to feelings of alienation and diminish the quality of human relationships in society. As AI systems become more capable of mimicking human behaviors, the potential for such social impacts grows, further underscoring the need for careful consideration of their design and implementation.

These are just a few examples of the many social impacts AI could have on society. However, the scope of AI's influence extends far beyond these cases, with countless other potential consequences yet to be fully explored.

2.2. Potential root cause factors

According to West et al. [13], the groups of people who are most commonly affected by discriminatory AI systems are women, people of color, and minority groups. This can be traced back to the power dynamics in the AI sector and the discriminatory behaviors that get cemented into the logic of these systems. This happens as a result of misrepresentation. West et al. [13] explained that the reason behind such misrepresentation in the workplace can be that opportunities to work on influential AI projects are more commonly given to white males from specific social, economic, and educational classes. As a result, the **people who are in power** in the AI sector **are also the ones who benefit the most** from the developed systems.

Take the case of gender, for example. Leavy [14] expressed that an over-representation of male software designers can be a contributing factor to continuous gender inequality in software. It is therefore suggested that increasing diversity in the workplace will help generate solutions to gender bias issues caused by AI systems [14]. Leavy [14] pointed out that software relying on machine learning is trained from observing data, and if these **data are governed by stereotypical bias**, the machine would operate in a biased way. It was proposed that "gender ideology is embedded in language" [14, p. 14] and that addressing gender representation in language could therefore be an approach to minimize such bias. This bias in text could be traced back

Table 1
Overview of exemplary AI regulation and guidance.

Legislation	EU AI Act [16]
Codes of ethics	IEEE [2], ACM [3]
Principles	OECD [4], Asilomar [17]

to namings, orderings, descriptions, metaphors, and the presence of the word "women" [14].

A similar case can be made with respect to people with disabilities. Related to the concern about biased AI, Whittaker et al. [15] mentioned that the discussion is lacking in the topic of disabilities. Developers with disabilities, who could be a great asset to develop systems suited to their own disability, are faced with barriers. APIs, which are crucial for developing modern software, rarely work with fundamental accessibility tools and requirements such as screen readers forces disabled people out of the role of developers. Even though companies could make such tools more inclusive, Whittaker et al. [15] elaborated that an important issue is that developer tools are not controlled by the developers themselves, but by the providing companies, who decide how these tools function, who can use them, and what features they include.

In addition to issues related to the developers and their decision during when developing software, there is always a more fundamental question to be asked: should a particular AI system be designed in the first place?

West et al. [13] recommended that in order to tackle the discrimination issues posed by AI, there should be an assessment on whether specific systems should be designed in the first place. As an example, these authors pointed out that AI systems that measure physical characteristics to make decisions — whether to predict sexuality or for law enforcement — should be reassessed carefully. They also highlight that the view on gender being exclusively binary in AI systems discriminates against groups that do not identify in such terms.

Similarly to the impacts, these examples highlight just a few of the potential factors contributing to negative social impacts in AI systems. It is essential to explore a greater range of factors and how they relate to broader social impacts of AI.

2.3. AI legislation, codes of ethics, principles and guidelines

When AI-based systems are deployed and integrated into society, various ethical concerns and potential impacts must be addressed, which our global society does in the form of legislation, codes of ethics, and sets of principles, see Table 1.

Within **legislation**, the Artificial Intelligence Act (AIA) is a European Union regulation for AI that establishes a common legal framework within the EU since August 2024 [16]. However, as the AIA is not an ethics guideline, it is reasonable to assume that measures beyond compliance are required for ethical AI systems concludes Westerstrand [18] after investigating how far compliance with the EU AI Act takes AI providers in developing ethical AI.[2]

Two widely recognized **codes of ethics** are the IEEE Code of Ethics and the ACM Code of Ethics and Professional Conduct. The IEEE Code of Ethics is a foundational guideline that all technology practitioners should follow. It acknowledges the profound influence of technology and commits professionals to be responsible for their impact on society [2]. Another effort for this same association is the "Ethically Aligned Design" [19], a series that offers recommendations that address critical issues in the field of intelligent systems.

Similarly, the ACM Code of Ethics, issued by the Association for Computing Machinery, encourages professionals to take responsibility for the societal consequences of their work [3]. In addition, ACM published a statement on Algorithmic Transparency and Accountability, outlining seven principles to guide the development and deployment of algorithms [20].

Concrete AI guidance are the sets of **principles** proposed by the OECD in 2019 with an update in 2024 [4]. Whittlestone et al. [21] highlight the Asilomar AI Principles [17], the UK House of Lords' five AI principles, and Google's own AI ethics guidelines. Cowls and Floridi [22] propose that the various ethical principles from different organizations should be merged into a smaller set of key principles, as there is considerable overlap among them. This consolidation would align with the recommendations put forth by the IEEE Global Initiative [19]. In addition, Floridi et al. [23] propose a set of principles for AI in society, amongst them beneficence, non-maleficence, autonomy, justice, explicability.

Khan et al. [24] carried out a systematic literature review of proposed sets of AI principles and the resulting and/or remaining challenges. The most common AI ethics principles were transparency, privacy, accountability and fairness. The most significant challenges were lack of ethical knowledge and that the principles remain vague without further guidance. In response to this, we argue that concrete guidelines may help to bridge these challenges.

The **insufficiency** of codes of ethics is argued by Mittelstadt [5], stating they provide only an overview of principles rather than concrete, actionable advice. Furthermore, AI's long-term impacts are difficult to predict during development, which makes it challenging for practitioners to foresee the consequences of their decisions [5]. This view is corroborated by the work of Hagendorff [7], who analyzed 22 AI ethics guidelines and pointed out their limited influence on developers' decision-making processes. He claims that the common issue is that most of these guidelines are created by a small slice of the population (predominantly white, affluent men) and fail to adequately address important social contexts such as care, welfare, and ecological networks [7, p. 103]. Furthermore, they rarely address the potential for political abuse of AI systems. Hagendorff [7] suggests that a more balanced approach, incorporating both technological recommendations and social considerations, would reduce the phenomenon of distributed responsibility among developers and help them better understand the long-term impacts of their decisions.

Munn [6] declares codes of ethics "useless" for bridging the gap from the AIA to actually just technological systems as they do not solve injustices of social systems they arise from, but we see them as an (albeit ultimately insufficient) tool that software engineers can directly use if sufficient guidelines are provided. Hence, guidelines need to provide more concrete and effective guidance.

Mittelstadt [5] advocates for a **bottom-up approach** to develop guidelines based on an analysis of 84 public–private initiatives in AI ethics. Since AI is used across various domains, examining ethical issues by studying real-world cases provides a more accurate representation of the diverse ways AI is applied. This approach involves analyzing the social, ethical, and legal implications of specific AI use cases. We contrast their ethical guidance analysis with a systems impact analysis. We adopt a similar bottom-up approach as Mittelstadt [5], but with an emphasis on the social consequences of AI-driven system, exploring the potential systemic effects and identifying the underlying factors that contribute to these outcomes.

3. Research method

In this section, we will explain the research method used in the study at hand.

[2] She finds that AIA is only partially aligned with basic liberties and equality of opportunity, and weakly aligned with the difference principle [18].

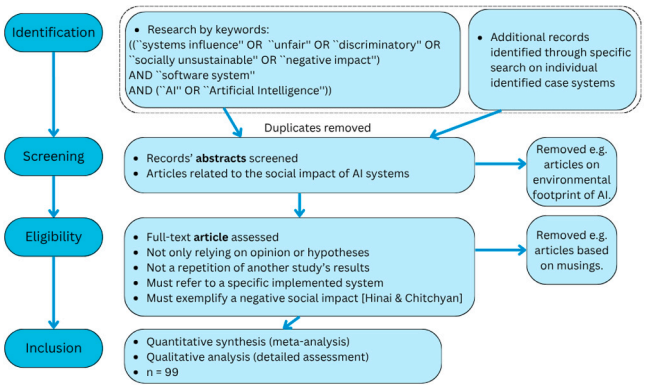


Fig. 1. PRISMA overview diagram of the applied process according to [28].

3.1. Data collection and selection

We conducted an exploratory **multi-vocal literature review** that used first and second tier grey literature [25]. We adhered to the guideline by searching a set of predefined keywords in Google Scholar. Google Scholar was chosen for its multidisciplinary nature, enabling us to capture research not only from IT but also from other fields that might address the social impacts of IT systems. In addition, we employed selective snowballing [26], following the most promising and relevant references from articles. This led us to additional papers and encouraged us to explore new terms. To discover further grey literature according to [25], we also searched for keywords in the regular Google search engine, where we identified examples of systems discussed that were not formally published in scholarly journals [27]. In our case, this primarily consisted of news articles and reports from various organizations.

We summarize the process including the specific overall search string in Fig. 1, a PRISMA diagram structured according to [28]. After preliminary searches with partial strings that led to promising leads, we identified an overall search string that included the most relevant results. We did not aim to perform an exhaustive literature review due to set time constraints for the overall project. The overall search string was defined as ((“systems influence” OR “unfair” OR “discriminatory” OR “socially unsustainable” OR “negative impact”) AND “software system” AND (“AI” OR “Artificial Intelligence”)). We excluded articles that were based on opinion or musings, articles that only repeated other studies’ results, and articles that did not discuss specific systems.

To assess whether an AI system had a negative social impact, we examined its effects on the social indicators outlined by Hinai and Chitchyan [29]. These indicators helped guide our decision about whether a system should be included in the study. The defined indicators are: employment, health, equity, education, security, social networks, services/facilities, resilience, human rights, social acceptance of technology, cultural, and political factors [29]. Systems that did not relate to negative social effects were excluded. Table 2 details each of these indicators. Additional records identified through specific search on individual identified case systems were added as relevant by following references.

We concluded sufficient coverage based on system types and application domains where AI was working supportively or centrally, and where that had led to a negative social impact to provide a meaningful set of analysis insights.

3.2. Data analysis

Once the AI-driven systems were selected, we proceeded to analyze their reported and potential impacts. For this analysis, we used the definitions of enabling and potential effects from Hilty and Aebischer [32].

Enabling effects refer to applying the IT system in its context of use. In our study, these were the social impacts directly described in the literature. When no enabling impacts were reported, we recontextualized the system to identify potential negative impacts. This involved looking for indirect or secondary effects that the system might have, even if they were not explicitly mentioned in the sources we reviewed.

Systemic effects, on the other hand, refer to the long-term, broader impacts of ICT, including changes in behavior and economic structures. In our analysis, these were identified by exploring how the enabling effects could lead to further, potentially negative outcomes over time. This was done by searching for additional literature that examined these long-term impacts. For example, if a system in a recruitment setting was reported to cause discrimination, we would look for literature on the long-term effects of discrimination, such as increased unemployment or reduced social mobility. In identifying systemic effects, we were also guided by the *Sustainability Awareness Framework* [33,34]. Specifically, we used questions related to the social and individual dimensions. For instance, we asked ourselves, “What effects can this system have on users with different backgrounds, age, groups, education levels, or other differences? What happens if systems like this are being used by many people, over extended periods of time (e.g. years)?” Finally, we classified them into common types of impacts.

Once the effects were identified, we attempted to find probably causes for them. First, we checked the sources to see if they mentioned any particular cause for the reported negative effect. In case they did not, we inferred the possible factors based on similar systems. Finally, we complemented the potential factors by means of root cause analysis [35]. This involved tracing underlying factors that may have contributed to the observed effects, such as organizational structures, technological limitations, or societal norms.

The factors are extracted from the sources that reported the system. In case the source did not mention a particular cause for the negative impact, we inferred the possible factors based on findings in similar technology. Some systems’ negative impact did not result from a design decision within the algorithm, but in the usage or context of it. Other systems used the same kind of technology but did not explain the specific underlying issue, like for the systems using facial recognition. Finally, we clustered causes in order to identify common types of contributing factors. The results of this analysis are presented in Section 4. The first two authors conducted the analysis and initial proposal of lists of impacts factors. The fourth author reviewed and provided quality assurance and validation during each step in discussion. The third author performed a thorough validation of the complete analysis. The replication package is available in [36].

4. Results

This section presents the results of our analysis; first the overview of the analyzed systems (Section 4.1), then the common negative impacts. Here we differentiate in between *enabling* impacts, brought about by usage of the system [33], and *systemic* impacts, the accumulation of immediate and enabling impacts over time [33]. Enabling impacts are presented in Section 4.2 and systemic impacts in Section 4.3.

It is worth noting that some impact categories within the enabling and systemic impacts do overlap. For example, we have chosen to highlight gender and racial inequality rather than grouping them under a broader category of systemic inequality. This approach allows us to draw attention to specific types of inequality that warrant special focus. A similar rationale is reflected in the Sustainable Development Goals (SDGs), where certain goals could arguably be subsumed under others — for instance, SDG 5 on Gender Equality could have been included under SDG 10 on Reducing Inequality — but the distinct importance of gender inequality justifies a separate category.

Table 2
List of indicators and descriptions.

Indicator	Description
Employment	This indicator includes various sub-indicators related to employment statistics and job conditions, such as "number of employed women," "number of full-time/part-time workers," "utilization of different working time arrangements," "compensation," and "job opportunities creation" [29], p.4.
Health	The health indicator encompasses topics like "the quality of health services provided to people," "health problems reported to authorities," "health risks," and "health practices" [29], p.4.
Equity	This indicator focuses on equal treatment and opportunities for all individuals, irrespective of gender, ethnicity, race, or social status, including people with disabilities. It covers aspects such as "income/wealth distribution," "social inclusion," "diversity of housing infrastructure," "provisions for the basic needs of disabled, elderly, or children with proper access," and "fair competition" [29], p.4.
Education	Education indicators address the availability and quality of educational facilities. These include "number of persons with higher education," "employees' educational level," "offered areas of employee training," "number of students per teacher," and "supporting educational institutions" [29], p.4.
Security	Security indicators focus primarily on various categories of crime and related concerns.
Social networks	Also referred to as "social cohesion," this indicator examines the connections between community members and their sense of belonging. Examples include "citizens' walkability to local places such as shops and community centers," "citizens' empowerment through participation in community activities and voluntary work or decision-making," "network and knowledge sharing," "tolerance of visible minorities," "identity," and "accountability in decision-making processes" [29], p.4.
Services/Facilities	This indicator addresses "the availability and access to services and facilities" [29], p.4.
Resilience	This refers to "the community's adaptability to changes" [29], p.4.
Social acceptance of technology	This indicator assesses a community's readiness to adopt new technologies [30].
Cultural	The cultural indicator concerns the preservation of a community's cultural identity.
Political	This indicator is focused on "governmental laws and people's trust in them" [29], p.4.
Human rights	This indicator examines issues such as child labor, forced labor, and discrimination. In addition to the points raised by Hinai and Chitchyan [29] regarding human rights, we also considered the conditions and statements outlined by the United Nations in the Universal Declaration of Human Rights. These statements apply universally to all individuals, regardless of race, gender, or language. For example, Article 23 ensures the right to work with the freedom of choice, and Article 9 prohibits arbitrary arrest [31].

4.1. AI-driven systems with reported negative social impacts

The 28 systems analyzed in this study are listed in Table 3. The table provides a short description of the systems selected in this study and the social indicator that they relate to. It also includes a "nickname", given by us, that conveys the purpose of the system. We have compiled this more concise version of the data to ease the understanding of the results by the readers. This table is extended in the appendix, in Table 7, and the replication package is available in [36]. For the 28 systems and their impacts and common factors, we analyzed data and information from 68 additional sources in both peer-reviewed and gray literature from 1985 to 2025, making it 99 sources in total with details provided in [36].

4.2. Negative enabling social impacts

This section discusses the results of the common factors for the negative enabling effects reported in the selected literature. Table 4 illustrates in which systems each of the common enabling effects was represented. In the following description, systems are referred to by their nickname, with their ID in parentheses.

Inequality in opportunities occurs when a system hinders some individuals from a possibility that others are given. For example, the *Amazon résumé scanner* (S3) was meant to review resumes and output the best candidates, but turned out to hinder women in the opportunity of receiving a job at Amazon [1]. In the *Health forecaster* (S7), an algorithm decided who should be enrolled in care management programs with extra resources and attention, based on a risk score. It was found that black patients, although sicker, received the same score as healthier white patients; so healthier white patients were being given a higher opportunity of recovering from their illness [42].

Potential for malicious use describes systems that have been, or have the potential to be, used to intentionally harm, mistreat or manipulate individuals. The *Tay social bot* (S6) was a chatbot created by Microsoft to engage in fun and normal conversations with users, but it was found to mistreat people by expressing racist comments [41], and could potentially manipulate people's opinions if it was kept up for a longer period of time. The *Cambridge Analytica data harvester* (S19) promised a financial gift in exchange for Facebook users filling out a survey. The app extracted a person's likes and friends lists from Facebook, identity, contact details and location; all of which was used to profile people [54] and for manipulating their opinions during America's presidential election in 2016 [69]. The *Sexual orientation predictor* (S22) claimed to predict someone's sexual orientation based on a picture of them. If used by people with homophobic opinions, people identified by the system risk both mistreatment and harm, specially in jurisdictions that criminalize homosexual activity [70].

Both the *Chinese trust score* (S26) and the *Uighur surveillance officer* (S27) also fall under this category, as they allow for authorities to

Table 3
Summary of AI-driven systems with reported negative social effects.

ID	System's nicknames	Description	Social indicator
S1	Beauty scorer	Beauty.ai was an AI system that evaluated and scored people's attractiveness based on facial features [37].	Equity
S2	COMPAS recidivism predictor	COMPAS is a risk assessment tool used in the criminal justice system to evaluate a defendant's likelihood of reoffending. This score is used in courtrooms and helps determine the time of release for prisoners [38].	Equity, Human rights
S3	Amazon résumé scanner	Amazon's recruitment tool was an AI system designed to assist in screening and ranking job applicants' resumes [1].	Employment, Equity, Human rights
S4	HireVue interview analyzer	HireVue is an AI-powered recruitment tool that analyzes video interview responses to assess candidates' suitability for a job [39].	Human rights, Employment, Equity
S5	Apple credit evaluator	Apple's credit card system uses an algorithm to assess applicants' creditworthiness and determine credit limits [40].	Equity
S6	Tay social bot	Tay was an AI chatbot launched by Microsoft on Twitter, designed to learn and interact with users in real time [41].	Human rights
S7	Health forecaster	Accountable Care Organizations (ACOs) use algorithms to predict which patients are likely to need complex and intensive healthcare, enabling proactive care management and cost control [42].	Equity, Health, Services/facilities, Human rights
S8	Google text translator	Google Translate is a web-based tool that translates text, documents, and websites between multiple languages using machine learning.	Equity
S9	Speech converter	Speech-to-text services that convert spoken language into written text. Koenecke et al [43] analyzed potential biases in such services provided by Amazon, Apple, Google, IBM and Microsoft.	Equity, Employment, Human rights
S10	Upstart smart lender	Upstart is an AI-driven lending platform that uses machine learning to evaluate creditworthiness and provide personal loans, considering factors beyond traditional credit scores [44].	Education, Equity, Human rights
S11	Predictive patrol officer	PredPol is a predictive policing tool that uses historical crime data to forecast future crime hotspots and help law enforcement allocate resources more effectively [45].	Human rights, Security, Equity
S12	Uber driver identifier	Uber uses Microsoft's Real-Time ID Check to verify drivers' identities through facial recognition [46].	Employment, Equity, Human rights
S13	Amazon face analyzer	Amazon Rekognition is a facial recognition tool that analyzes and identifies faces in images and videos [47]. The tool is used by police in the United States, for detecting, verifying and analyzing faces [48].	Equity, Security, Human rights
S14	Clearview identity finder	Clearview AI is a facial recognition system that uses a vast database of publicly available images to identify individuals, primarily for law enforcement and security purposes. [49].	Equity, Security, Human rights
S15	Examplify exam monitor	Examplify is a secure exam-taking software used by schools and universities to administer online assessments. The application uses face recognition to allow students to sign in [50].	Equity, Education
S16	Proctorio secure examiner	Proctorio is a remote proctoring software that uses facial recognition and monitoring tools to prevent cheating during online exams [51].	Equity, Education
S17	Giggle girls social networks	Giggle is a networking app designated for girls-only. The app verifies that users are girls by prompting the user to take a selfie when signing up for the platform. By using "bio-metric gender verification software" the app then confirms the gender of the new user [52].	Equity, Social networks
S18	Google image analyzer	Google Cloud Vision is an image recognition service that uses machine learning to analyze, label, and extract information from images [53].	Equity
S19	Cambridge analytica data harvester	The "thisisyourdigitallife" app, developed by Cambridge Analytica, collected personal data from Facebook users and their friends to build psychological profiles for targeted political advertising [54].	Political, Social networks
S20	Theft scoreer	The "Sensing project" is implemented by the police in Roermond, Netherlands. By using cameras, the police collect data of vehicles in the area in order to find potential pickpockets or shoplifters. The collected data was analyzed by an algorithm that then outputted a prediction in the form of a risk score [55].	Human rights, Security, Equity
S21	Facebook ads	The Facebook ad delivery system is used by companies to promote their products and services. The system uses an ad auction and machine learning to point ads to the appropriate people at the right time [56].	Equity, Employment, Human rights
S22	Sexual orientation predictor	Kosinski and Wang developed a system, which they claim to predict someone's sexual orientation based on their pictures [57]. They reported the system to have accuracy of 81% at predicting people who identify as homosexual [58].	Human rights

(*continued on next page*)

Table 3 (*continued*).

S23	Facial criminal tendencies guesser	Faception is a system that claims to be able to identify potential terrorists or pedophiles based on images. According to Michael Kosinski, a Stanford social psychologists who is an advisor at Faception, facial features can be connected to criminal tendencies [59].	Equity, Security
S24	Autonomous vehicles	Autonomous vehicles are self-driving cars that use sensors, AI, and machine learning to navigate and make decisions without human input [60–62]	Employment, Social acceptance of technology
S25	Deepfake video falsifier	Deepfake is a technology that allows creation of videos that seems to include real people saying and doing things they never really did [63]. Face2Face uses this technology to map and transfer facial expressions from one person to another [64].	Politics, Human rights, Social acceptance of technology
S26	Chinese trust scorer	Chinese social credit system is a data driven system that assigns a "score" to citizens to reward their behavior or to punish them. The score controls the kinds of benefits and rights that someone is entitled to, such as access to private school, air travel and real estate purchases [65].	Human rights, Equity
S27	Uighur surveillance officer	According to [66] Hauwei and an AI firm called Megvii tested a software feature called "Uighur alert". The feature is able to detect Uighur people from images. This was discovered by IPVM, a US based company specialized in video surveillance analysis. According to the two collaborating companies, they didn't have the intention of releasing the feature.	Human rights, Equity, Health
S28	Social media filters	Social media platforms like Snapchat and Instagram have introduced filters and lenses. The algorithms that these companies have created identify the face or faces which are visible for the camera and applies different types of effects [67,68].	Health

Table 4
Common enabling impacts.

Enabling impacts	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	S21	S22	S23	S24	S25	S26	S27	S28
Inequality in opportunities																												
Potential for malicious use																												
Law breaking																												
Directed policing																												
Privacy violation																												
Gender discrimination																												
Racial discrimination																												
Ethnic discrimination																												
Genetic discrimination (excluding skin color)																												
Negative financial impact																												
Negative impact on education																												
Wrongfully flagged																												

Table 5
Common systemic impacts.

Systemic impacts	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	S21	S22	S23	S24	S25	S26	S27	S28
Damage self-esteem																												
Trigger stress																												
Perpetuate division of socio-economic classes																												
Lack of diversity in workplace																												
Triggering dysfunctionality in families																												
Threat to safety																												
Perpetuate gender inequality																												
Perpetuate racial discrimination																												
Perpetuate ethnic discrimination																												
Perpetuate stereotypes																												
Influence opinions																												
Negative health impact																												
Decreased trust in authorities																												

harm, mistreat and manipulate citizens. The former assigns a "score" to citizens to reward their behavior or to punish them, as it happened with Journalist Lui Hu, who became blacklisted due to her writing about censorship and governmental issues [71]. The latter detects people belonging to the ethnicity of Uighur people, a Muslim minority that has been mistreated and oppressed by the Chinese government [66,72–74]; putting them in danger of being reported to authorities.

Law breaking refers to systems that break the law. For example, suing the creators of a system indicates the existence of illegal aspects in the system or in how it operates. This category includes systems that are illegal or banned in some places, like the *DeepFake video falsifier* (S25), which allows the creation of videos of people acting in ways that never actually happened [63], as it happened with false sexual images portraying Helen Mort [75] and never-issued statements from Obama

Table 6
Potential factors causing the negative social impacts.

Factors	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	S21	S22	S23	S24	S25	S26	S27	S28
Misrepresentation/poor diversity in the dataset	■		■	■					■		■							■		■								
Proxy bias					■		■			■																		
Existing social patterns/prejudice								■												■								
Issues in facial recognition												■	■	■	■	■												
Inappropriate use of AI	■	■															■		■	■		■	■		■	■	■	■
Lack of robustness (for external manipulation)						■																		■				
Unknown																					■		■					

shared in Buzzfeed [64]. In *Cambridge Analytica data harvester* (S19) the federal trade commission sued Cambridge Analytica's former chief executive and an app developer of the system [76].

Directed policing is concerned with systems that have, or can, direct the police towards a certain group of people or individuals, for example the *Predictive patrol officer* (S11) or the *Facial criminal tendencies guesser* (S23). The former is intended to help to predict crimes and has been used by the police to choose which areas to patrol [45] and even signaling individuals who have never committed a crime [77]. The latter is a system that claims to be able to identify potential terrorists or pedophiles based on image [59], which could potentially falsely accuse people because of their facial traits [59,78].

Privacy violation encompasses systems that collect data about others unknowingly or in a deceptive way. This data could include a range of different information such as images, names, identifiable information, location and contact information. This form of impact often occurs in surveillance contexts, for example, when people are publicly linked to certain places, activities or other people, due to their faces being recognized by a system. For example, *Amazon Face analyzer* (S13) and *Clearview identity finder* (S14) are used by police in the United States for detecting, verifying and analyzing faces [47–49]. The latter is based on images from Twitter, Facebook and Google that might even have been posted without consent.

Gender discrimination is when someone is being discriminated against only because of their gender. Since such attitudes are often based on generalizations, false beliefs, and on considering gender when it is irrelevant, the term is also related to stereotypes. Systems that either discriminate based on gender or as a result of gender stereotypes are therefore listed under this category, e.g. the *Amazon résumé scanner* (S3). Another example is *Tay Social Bot* (S6), who responded to users provocations with misogynistic tweets [41].

Racial discrimination covers systems that negatively impact people of color. For example, the *Health forecaster* (S7), mentioned earlier, was found to rate black patients with lower risk scores than white patients. Similarly, the *Amazon face analyzer* (S13) refers to a tool called "Recognition", that is capable of identifying and analyzing faces. The tool was found to perform better when identifying light-skinned people, compared to dark-skinned people [79], which can be harming to dark-skinned people, depending on the use the tool is given.

Ethnic discrimination includes systems that negatively affected people of certain ethnicity. For example, *COMPAS recidivism predictor* (S2) gives crime defendants a score indicating their likelihood of reactivism, and has been used in courtrooms to help to determine the time of release for prisoners [38]. Its algorithm has been noted to give higher score to African Americans compared to white defendants [80]. Furthermore, the *Upstart smart lender* (S10), an online lending platform that offer loan deals to students, was found to offer higher rates to Hispanic colleges and universities compared to graduates from institutions that were attended by people not belonging to minorities [44].

Genetic discrimination describes systems that negatively affect people due to genetic characteristics other than their skin-color. This category includes cases where discrimination occurs on the basis of physical appearance and/or capabilities. The *HireVue interview analyzer* (S4) assess how well a candidate would perform in a specific job, based on how well they did on their interview. This can be very harmful for people with disabilities, who may not conform to standard expectations for "doing well on an interview", specially those with speech impairments [81].

Negative financial impact incorporates systems that operate in favor of specific groups during hiring processes, offers unequal payment or unequal loan opportunities. Unequal or limited access to specific working fields can count as determinants of financial status. Those who are disadvantaged by these systems experience a negative financial impact. Both *Amazon résumé scanner* (S3) and *HireVue interview analyzer* (S4) can negatively financially impact women and disabled people, respectively, that were not selected to continue the job application process because of these characteristics.

Negative impact on education refers to cases where an AI system hindered access to education or made the educational processes more difficult. For example, *Examplify exam monitor* (S15) uses face recognition to allow students to sign in to online exams [50]. The system did not allow a dark-skinned student to sign in to his exam, saying it was unable to identify his face due to poor lightning, which, however, could not be resolved by adjusting the lighting in the room [50]. *Proctorio secure examiner* (S16) is also a testing system that schools use to conduct online exams. A black woman expressed that every time she used the tool it requested that she should shine more light on her face to validate her identity, while her white peers never had that problem [51].

Wrongfully flagged concludes the enabling impacts describing cases in which individuals become subjects to false positives, and, as a result, are accused of a crime they did not commit. A false positive in this case refers to when a system falsely identifies someone as another person or flags a person due to an overestimated risk score. The later form of flagging is often seen in cases where systems aim to prevent crimes before they happen. For example, the *COMPAS recidivism predictor* (S2) may lead to people to spend more time in prison because of an automatic recidivism score [38]. The *Predictive patrol officer* (S11), also described earlier, has been known to incorrectly flag individuals [77].

4.3. Negative systemic social impacts

This section presents the findings of the common factor analysis of the negative systemic effects identified in the selected literature. Table 5 shows which systems are associated with each of the common systemic effects. The line between enabling and systemic impacts is sometimes blurry. This is reflected in the common impacts, and explains why some enabling impacts continue to exist systemically by exacerbating the enabling impact. Furthermore, some of the systemic impacts listed here are also seen on the enabling level as they pertain to the individual, but since their *aggregation amplifies the impact*, we list them in this section instead of duplicating them.

Damaged self-esteem represents negative impacts on self-esteem, self-image, and sense of fulfillment. Damaged self-esteem also represents feelings of exclusion and feeling powerless in comparison to other

individuals. For instance, damaged self-image is a systemic impact of the *Beauty scorer* (S1), a system that claims to behave as an objective judge for human beauty contests [37]. Self appearance satisfaction is at risk when unattainable beauty standards are often portrayed in media and amplified in conversation [82]. Damage to self-esteem is also an apparent systemic impact of the predictor of recidivism of *COMPAS* (S2), as inmates with a long sentence have a higher stress level and a worse self-esteem than those with a shorter sentence [83]. *HireVue interview analyzer* (S4) can systemically impact disabled individuals' sense of fulfillment and increases feelings of exclusion, as they are more often denied the purpose and the community connections that normally come with having a job [84] As these systems become more widespread, their impact exacerbates.

Triggers stress relates to systems that trigger stress for different reasons. For example, systems like *Facial criminal tendencies guesser* (S23) often wrongly flag certain types of individuals, other people that identify with these individuals might start feeling stressed and afraid that this might happen to them in the future. Another example is when systems invade people's privacy, such as the Sexual Orientation Predictor (S22). Imagine these systems being constantly used against individuals from countries that criminalize LGBT identities. Even after fleeing their home countries as asylum seekers, individuals who have faced persecution for their sexual orientation are at a high risk of developing mental health issues, such as severe stress and depression [85].

Perpetuates division of socio-economic classes is concerned with systems that perpetuate or enhance differences in individuals' socio-economic status. According to [86], someone's social and economic status is measured by looking at education, income and occupation. For example, the Amazon résumé scanner (S3) falls within this category, as it could exacerbate the existing economic gender gap. If the algorithm is biased against women or other marginalized groups, it may disproportionately filter out qualified candidates based on patterns in past hiring data, reinforcing gender inequality in the workplace and widening the socio-economic divide. Another example are systems like the COMPAS recidivism predictor (S3). When an individual is convicted, their entire family is often immediately impacted, particularly financially. If systems disproportionately affect African Americans, leading to longer sentences for this group, the cycle of socio-economic disadvantage can persist. Children of those incarcerated individuals face prolonged hardship, possibly growing up in financially strained households with limited opportunities.

Lack of diversity in workplace represents systems that contribute to gender, racial and ethnic imbalances in specific job fields. It also includes misrepresentation of marginalized groups, like people with disabilities. Systems that contribute to poor diversity in workplaces are listed under this category. For example, *Google text translator* (S8) activities when translating from gender-neutral languages such as Finnish, Filipino and Hungarian. This translation has shown to be sexist [87]. Another example is the *speech converter* (S9) systems; If used in a job application process, it could affect diversity in the workplace, as studies have shown that applicants with Hispanic accents had a lower chance of getting the job compared to a standard American English speaker [88].

Triggering dysfunctionality in families includes systems that negatively affect family life and/or impacting the well-being of children. For example, systems like the *Apple credit evaluator* (S5) can prevent equal access to money and limits financial freedom, which in turn can lead to power and control imbalances in family dynamics [89]. Again, with a larger user base, this effect becomes systemic over time. Systems like the *Uighur surveillance officer* (S27) can put people in danger of being sent to camps for "re-education" [72]. The camps have been described as internment camps [90]. Descendants of Japanese who had been in internment camps during WWII, reported stories of family and material loss [91].

Threat to safety is present in systems that allow organizations and governments to target specific groups of people. *Autonomous vehicles* (S24) has the potential of being hacked, making it possible for malicious people, such as terrorist and criminals, to manipulate the system [62]. With systems for both security and safety being accessible online, increasingly connected, and exposed to more misuse potential, there is a high potential for these effects to become systemic. Furthermore, systems like the *Sexual orientation predictor* (S22) enable the identification and harassment of LGBT individuals, reinforcing stigma against this community. This normalization of discrimination can lead people to perceive anti-LGBT sentiments as acceptable, further encouraging harassment and persecution.

Perpetuation of gender inequality incorporates systems that negatively impact gender imbalances. For example, *Amazon résumé scanner* (S3) places female applicants at a disadvantage, which is an example of gender inequality. As systems like this become common in HR, women start to feel discouraged from entering certain fields and the effect becomes systemic. *Facebook ads* (S21) use an ad auction and machine learning to point ads to the appropriate people at the right time [56]. A study [92] found that the algorithm shows different jobs to females compared to males, even though the displayed jobs require the same qualifications. If this kind of bias becomes common place, women are kept from certain job opportunities and the gender inequalities get perpetuated.

Perpetuation of racial discrimination occurs when systems disadvantage racially marginalized groups and limit their opportunities based on race. The *Google Image Analyzer* (S18) is a computer vision service that automatically labels images using AI [53]. In an experiment, a thermometer held by a dark-skinned individual was incorrectly labeled as a "gun," while the same object was identified as an "electronic device" or "monocular" when held by a light-skinned person [53]. If systems like this become widely adopted for weapon detection in places such as schools, concerts, and malls, dark-skinned individuals are more likely to be wrongfully identified as threats, perpetuating racial discrimination against them. Additionally, when systems like the *Beauty Scorer* (S1) consistently portray white individuals as attractive while labeling dark-skinned individuals as unattractive, harmful media biases are reinforced [37]. This can lead to discriminatory behavior, such as social exclusion or the questioning of marginalized groups' rights.

Perpetuation of ethnic discrimination are systems that target specific groups based on their ethnic background. The *Theft scorer* (S20), for example, a system in the "Sensing Project" that uses cameras to collect data of vehicles in the area in order to find potential pickpockets or shoplifters [55]. Also, systems like the *Uighur surveillance officer* (S27) contribute to institutionalize ethnic discrimination by enabling automated profiling and persecution of ethnic minorities, reinforcing their marginalization [66]. It legitimizes mass surveillance, deepens societal biases, and sets dangerous precedents for AI-driven racial and ethnic profiling worldwide.

Systems in the previous three categories, were also listed in Table 3 under the enabling impacts "Gender discrimination", "Racial discrimination" and "Ethnic discrimination", respectively, as they in the direct impact, also contributes to perpetuating these kinds of discrimination.

Perpetuation of stereotypes either generates new stereotypes or manifests already existing prejudices in society. For example, the *Amazon Resumé scanner* (S3), perpetuates stereotypes that women are not fit for the tech field. Stereotypes can also be associated with societal norms and expectations, like beauty standards, which is seen in both the *Beauty scorer* (S1) and the *Social media filters* (S28).

Negative health impact includes systems that negatively impact individuals' health in any way. For example, the *Health forecaster* (S7) hindered black people from getting enrolled in care management programs, which in turn would have allowed them to receive extra care. Additionally, systems like *Examplify exam monitor* (S15) and *Proctorio secure examiner* (S16) can help to spread the adoption of online testing, which was in itself linked to students' eating and sleeping habits over time. Finally, systems like *Giggle girls social network* (S17), which uses biometric gender identification algorithms to decide whether someone

is allowed to log in, can contribute to anxiety in trans-girls [52], which again can lead to negative health effects.

Influence on opinions encompasses systems that had the ability to influence the public's opinion in an untrue or nontransparent way. For example, if the *Facial criminal tendencies guesser* (S23) falsely labels individuals, other people's opinions about them may be affected by this label. Hence, over time, the system may help to reinforce stigmatizations of certain collectives [59]. Another example is the *Cambridge Analytica data harvester* (S19), which clearly had the ability to influence the public's opinion. Over the last ten years, we have seen the accumulated impacts of fake news on a global scale [93,94].

Decreased trust in authorities includes systems that cause wrongful accusations, allow authorities to wrongfully identify individuals or enable authorities to invade citizens' privacy, are listed under this category. *Deepfake video falsifier* (S25) is included here, as such fake videos, when including fake images of authorities, may over time contribute to decreased trust in authority and journalism [95]. The *Predictive Patrol Officer* (S11) poses similar threats, if police surveillance systems disproportionately direct officers to Black communities. This can result in a higher incidence of police mistreatment in these areas and further deepen mistrust between residents and law enforcement.

4.4. Common potential factors for impacts

This section outlines the outcomes of the common factor analysis for the factors that may have let to the enabling and systemic effects above. Table 6 highlights the systems that correspond to each of the potential factors. We next describe each of these potential factors:

Misrepresentation/poor diversity in datasets refers to issues like under-representing or misrepresenting certain groups of people. In some cases, the datasets use data records that perpetuate biases or historical differences, even though those sets of data are a "true" representation of the past. This is the case for the *Amazon résumé scanner* (S3), which was based on resumes from applicants, submitted to the company during a 10-year period. Those resumes mainly came from men, due to that more men had applied to the tech industry [1]. In the *Predictive patrol officer* (S11), the algorithm relied on local report data from the police's records which supposedly should track accurate crime rates; however, if police heavily patrols a specific area or neighborhood, the data records would naturally over-represent people who live in these areas [77].

Proxy bias occurs when one attribute is used to determine another. For example, Cathy O'Neil [96] discusses that certain attributes, such as the geographic location of our homes, is a proxy for race, since many cities are so segregated. In the *Apple credit evaluator* (S5), the company and developers discussed that the algorithm did not use gender as an attribute and that it therefore could not be discriminating based on it. However, there could be proxies that caused the algorithm to include such biases [97]. The *Health forecaster* (S7) uses health costs to assess the need for care. However, due to unequal access to health care, less money is generally spent on black patients; leading the system to conclude that they are healthier than white patients [42].

Existing social patterns and prejudice may have made their way into algorithms and design decisions during the development of some of the AI-driven systems selected. For example, *Google text translator* learns how to translate by analyzing a huge amount of examples [87], which connect certain words with a certain gender based on the embedded sexism of our society. Another example is the *Theft scorer* (S20), a system in the "Sensing Project" that uses cameras to collect data of vehicles in the area in order to find potential pickpockets or shoplifters. According to [55], the police defined "mobile banditry" in the project as "pickpocketing and shoplifting committed specifically by individuals of Eastern European nationality" [55, p. 6], discriminating against those nationalities already by definition and so they would receive high scores by the system [98].

Inappropriate use of AI is the most frequent category for the selected system and is concerned with the choice of designing and implementing the system in the first place. The choice of designing these systems poses ethical issues and is inappropriate in their specific context. For example, the widespread malicious use of the *Deepfake video falsifier* (S25), suggests that there should be some restrictions and laws that hold the perpetrators accountable [75]. Similarly, the purpose of the *Chinese trust scorer* (S26) is also dubious; a data driven system that assigns a "score" to citizens to reward their behavior or punish them [65]. The score controls the kinds of benefits and rights that someone is entitled to, such as access to private schools, air travel and real estate purchases. The system lacks transparency, making it hard to recover from a low score [71], has the potential to fundamentally violate human rights, and its sheer existence enables authoritarian control through mass surveillance [99].

Vulnerability for manipulation refers to the ease of external manipulation of AI-driven systems. For example, the *Tay social bot* (S6) could be easily provoked, learned from "trolls" who suggested phrases and caused the chatbot to search the internet for replies that fitted their tweets [100]. The *Autonomous vehicles* (S24) can be hacked if people understand how their algorithms work. Hence, consequences of autonomous vehicles are dependent on how well we prepare for their existence, and to what extent we prepared the vehicles for ethical scenarios. Along this line, Lin [61] reminds us that "when technology goes wrong–and it will–thinking in advance about ethical design and policies can help guide us responsibility into the unknown". The latter is true not only for AI systems.

Unknown is not a category in itself, but is used here to list the systems for which no indication of a potential factor was found or could be conceived of. Since not all companies disclose the inner workings of their algorithms, the pattern *unknown* arose. The systems included here did not reveal any indication of a potential factor.

5. Discussion

There exist some ethical guidelines and principles that aim to guide development of software systems, such as the ones provide by IEEE and ACM [2,3]. Yet, our findings showed that there exist systems that violate them, which encourages us to ask: *"Why are ethical principles being violated?"*. Surely, the designers of most of these systems did not intend to do so. As observed by Mittelstadt [5], one potential problem is that they only provide high-level guidance but lack concrete advice. Furthermore, it is tricky to predict the long-term impacts of AI-driven systems and the effects that decisions made during development will have in the future [5]. Following Mittelstadt's advice, we followed a bottom-up approach, by examining different cases and seeking to arrive at a more realistic representation of the common impacts and their potential causes [5].

This section discusses our findings, in four subsections. The first two discuss, respectively, how the impacts and factors affect well-known ethical principles and relate them to the literature. The third subsection provides some initial guidelines to address the issues observed. The fourth subsection acknowledges the validity threats of our study. Lastly, we detail implications for research and practice.

5.1. Impacts and types of impacts

The ACM code of ethics and professional conduct defines seven general ethical principles [3]. These are: 1.1 Contribute to society and to human well-being, acknowledging that all people are stakeholders in computing; 1.2 Avoid harm; 1.3 Be honest and trustworthy; 1.4 Be fair and take action not to discriminate; 1.5 Respect the work required to produce new ideas, inventions, creative works, and computing artifacts; 1.6 Respect privacy; and 1.7 Honor confidentiality. The systems reviewed in this paper clearly violate three of these principles — (1.2), (1.4), and (1.6). The following section examines how each of these principles has been breached:

5.1.1. Principle 1.2 – "Avoid harm"

describes that technology should not cause any harm [3, p. 4]. This includes, for example, physical or mental harm, and harm to someone's reputation. It can also include misconducting information in a harmful way [3]. Generally, this principle is being broken by several of the systems identified in this work. The most obvious examples are systems that cause systemic impacts such as *negative health impacts*, systems that lead to *damaged self-esteem* and those that contribute to *influenced opinions*. Yet, systems that *trigger stress* and those that could be a *threat to safety* violate this principle. We discuss five examples.

Physical characteristics. As presented by Table 3, there are many impacts associated with systems that measure physical characteristics. Measuring physical characteristics could be used to classify gender or sexual orientation, such as in *Uber driver identifier* (S12), *Giggle girls social networks* (S17) and *Sexual orientation predictor* (S22). The impacts are: *threat to safety*, *potential for malicious use*, *sexist*, *negative health impact*, *decrease trust in authorities*, *genetic discrimination (excluding skin color)*, *negative financial impact*, *damage self-esteem*, *trigger stress* and *perpetuate gender inequality*. This long list of enabling and systemic impacts explains why West et al. [13] emphasized careful reassessing of systems that measure physical characteristics, especially of systems that view gender in an exclusively binary way.

Stress and self-esteem. We found that many of the identified systems have a negative systemic impact on users' stress levels and self-esteem. For example, systems like *Examplify exam monitor* (S16) and *Proctorio secure examiner* (S17) can cause stress by increasing feelings of surveillance, leading to anxiety about performance, privacy concerns, and fear of false accusations. *Beauty Scorer* (S1) can make individuals feel inadequate or insecure if they do not meet the system's criteria. Sometimes, these implications are difficult to pin on an AI system as there exists other, more traditionally known, causes. For example, parenting style is known to influence self-esteem in adolescents, as shown in [101], also school environments and experiences can have such influence [102]. It has also been found that there is a connection between high usage of social media and low self-esteem [103]. However, except for the *Giggle girls social network* (S17) and the *Social media filters* (S28), our findings present *damage self-esteem* as an impact of systems that are not in a social media context. Therefore, it is important to keep in mind that there are other contributors to damaged self-esteem, even though they may not be as present in the public discourse.

Family. Another systemic impact that we found is *triggering dysfunctionalities in families*. These systems can also impact *perpetuation of division of socio-economic classes*. The relation between these impacts was brought up by Conger et al. [104] who mentioned that one's social position has an influence on families over time. Rather than focusing on how technology impacts families in the sense of interaction and quality time, we focused on the economic opportunities that are lost due to unfair systems. For example, the COMPAS Recidivism Predictor (S2) may unfairly affect a person's chances of rehabilitation, while the *Apple Credit Evaluator* (S5) uses algorithms that can unfairly penalize people with limited financial history. Such systems can affect family members who are financially dependent on the people being judged. Even though the association between these impacts seem natural, it may not be as obvious to consider them together as impacts of AI systems.

Opinions and perspectives. Influence opinions can appear in different contexts. AI systems can affect the public's opinion in political contexts and also in employment processes. For example, *Cambridge Analytica data harvester* (S19) can significantly influence political opinions by using personal data to build psychological profiles and swaying voters based on their specific beliefs and biases. Similarly, *Beauty scorer* (S1) can lead to discrimination in hiring and promotion decisions that prioritize certain type of appearance over skills or qualifications. This means that influencing people's opinion can occur on an individual or small-circle level, and on a public level. The content of the circulated opinion implies different levels of risks, which is usually difficult to undo such impact.

Health. The categorization *negative health impact* implies that AI systems can influence people's health. In some cases, it can be done indirectly where AI systems are responsible for decision-making processes. For example, *Health forecaster* (S7) has been shown to discriminate against black people, when determining who would need extra care. This aligns with the concerns in [11] regarding usage of AI systems in the field of health. Often medical studies were only carried out on white males and it is unknown whether a similar treatment is even beneficial to a female patient or person of color, an effect commonly known for 30+years [105]. In [11], it is expressed that AI systems involved in health predictions can perpetuate discriminatory behavioral patterns, which is exactly the case in *Health forecaster* (S7). As shown in the results, the bias executed by such systems also contributes to *perpetuating inequality in opportunities* and *racial discrimination*.

5.1.2. Principle 1.4 – "Be fair and take action"

This principle urges practitioners to embed equity and inclusivity in their systems and avoid all forms of prejudicial discrimination [3, p. 5]. As seen in the results, some systems did not have the intention to discriminate, but the perpetuated bias lead to several negative impacts. These impacts were: *Racial discrimination, Gender discrimination, Directed policing, Ethnic Discrimination* and *Genetic Discrimination (excluding skin color)*. We provide three examples.

Historically marginalized people. In the results we see that several systems contribute to negative impacts such as *ethnic discrimination, sexism, gender discrimination* and *racial discrimination*. What these impacts have in common is that they affect historically marginalized people, as we have seen in the *Tay social bot* (S6) and in the *Uighur surveillance officer* (S27). This aligns with the statement by West et al. [13], that the most common groups of people who are discriminated against by AI systems are women, people of color and minority groups.

Disability. We found that several enabling and systemic impacts concern people with disabilities, such as the *HireVue interview analyzer* (S4). Impacts such as *perpetuating inequality in opportunities* negatively affect these people in their integration into society and isolate them even further. This aligns with what was expressed in [15], that the limited creation of technology that is suited for people with disabilities contributes to inequality. As shown in the results, not only does excluding people with disabilities from job opportunities increase inequality, but it also affects existing work environments, hence the systemic impact *lacking diversity*. As our findings show that people with disabilities are negatively impacted by AI systems, they strengthen the claim in [15], that the discussion of biased AI is lacking the topic of people with disabilities.

Diversity. Even though we found studies that implied that women, minority groups, and people with disabilities are more frequently discriminated against by AI systems [13], our findings also presented examples where other aspects of diversity were impacted. The contexts of AI that showed to have an impact on diversity were related to workplace and developers. This was discussed by Leavy [14] who suggested that an increased diversity in the workplace would help lessen biases such as gender bias in AI systems. Our findings indicate that this lack of diversity in workplaces is, but is not limited to, a result of poorly developed AI systems like in *Amazon resumé scanner* (S3). Through our study, we identify a loop of *Lack of diversity in workplace* in tech-related professions that contribute to perpetuate uniformity in the workplace.

5.1.3. Principle 1.6 – "Respect privacy"

This principle recognizes that technology collects and uses sensitive and private data about its users, however, the principle urges that this information is used for legitimate ends [3, p. 6]. In addition to that, individuals should know when their data are being collected and the purpose that they are used for [3]. As shown in our findings, the systems listed under *privacy violation* clearly do not meet the expectations of this principle. We present one example relevant for this principle.

Facial recognition. The most well-known invasions of privacy are related to facial recognitions systems. As seen in Table 4, there are several systems that use facial recognition for the sake of law enforcement. However, the biases of such systems have a variety of negative impacts, such as being *wrongfully flagged* or suspected of a crime. These unfair treatments have been observed, for example, in the *Predictive patrol officer* (S11) and the *Amazon Face analyze* (S13). The potential for magnified impacts resulting from facial recognition systems may indicate that their usage should be limited. For example, in July 2020 a number of municipalities in the U.S. banned facial recognition. This ban indicates awareness of facial recognition societal harms in the context of surveillance [106].

5.2. Factors for negative impacts

Many factors that we identified in our study are well-known within the field. For example, Chen et al. [11] showed that when data sources are unavailable or include discriminatory behavioral patterns, they can contribute to **bias in algorithms**. Leavy [14] also mentioned that when machine learning systems are trained by observing data that include stereotypical biases, the system will as a result operate in a biased manner. This is highly related to our findings and the identified factor *misrepresentation/poor diversity in the dataset* that we found for several systems. Leavy [14] also mentioned that gender is embedded in language, which causes systems to develop biases based on those stereotypes. For example, namings, orderings, descriptions, metaphors and the word "women" often hold biases. This is related to the factor *existing social patterns/prejudice*, and especially related to Google text translator (S8).

West et al. [13] offer that the approach to tackle discrimination caused by AI could be to reassess the decision to build specific systems in the first place. This is related to the factor *inappropriate use of AI* that we identified. This was also the most frequently appearing factor among the systems that we found.

Factors that cause bias in AI systems are of course also traced to algorithmic decisions. Corbett-Davies and Goel [107] argued that the three formal definitions of **algorithmic fairness** that have been notably discussed, in fact, have statistical limitations. One of these definitions was "anti-classification, meaning that protected attributes — like race, gender, and their proxies — are not explicitly used to make decisions" [107, p. 1]. The authors argue that the definition of anti-classification is difficult to achieve. This is because excluding certain attributes and proxies for the sake of fairness is not efficient enough, since it is hard to know which attributes act as proxies for which attributes. The authors meant that almost all attributes can reveal protected attributes. They also said that many of these attributes are considered legitimate to include in decision makings, such as education in a hiring process. On the other hand, it is also argued that there exists cases where including protected attributes is necessary to reach fair decisions. When a protected attribute adds predictive value, such attributes should be included, and when they do not add such value they can be excluded from the algorithm. However, Corbett-Davies and Goel [107] highlighted that if the latter is the case, then an accurate risk model could in theory be built by examples from solely one group of people. Hence, this paper recognizes that creating fair algorithms is difficult, and that even the definitions that aim at guiding development of fair algorithms have limitations. This further explains why companies struggle at developing fair and socially sustainable systems.

A potential factor we found for systems using facial recognition was that the **identity of developers** can have an effect on training sets or some other aspect of the development. This was demonstrated in [108] where Asian algorithms performed better for Asian people, and Western algorithms performed better for people from the west. This also relates to what was stated by West et al. [13], who meant that misrepresentation in the development team can have an influence on how discriminatory a system turns out. Both Leavy [7,14] also brought up the issue with over-representation of males in the tech-industry. This again suggests a potential relationship between the identity of the developers and the behavior of the system.

As mentioned by Leslie [109], another possible factor of bias in facial recognition is the designers lack of attention when testing the performance of systems. Especially when testing datasets that include historically marginalized groups. This suggests that human errors and behavior can have an influence on how discriminatory a system is. Even though we did not find this as a factor for the systems in Table 4, it can be a contributing issue that is difficult to identify due to, for example, lack of transparency.

One factor that was pointed out by Hagendorff [7] is that the guidelines he analyzed rarely discussed the **potential abuse** that AI systems can contribute to. This indicates that a lack of consideration of such political abuse can be a factor that causes unsustainable software systems. Although we did not identify this as a pattern, we did see some systems who had the potential of being used for such purposes. If potential use for political abuse were considered in development of AI systems, such systems could either be reconsidered, or a robust design that withstands such behaviors could be adapted.

Another factor that we did not find in the systems we identified is what Mittelstadt [5] discussed about companies' priorities when developing their systems. The article [5] mentioned that in comparison to the medical industry, the **well-being** of the users is **not** necessarily the main **priority** for tech companies. Usually, they have different objectives such as decreasing costs and meeting their stakeholders expectations. This could be an underlying factor for the systems we found, although we did not identify a direct relation.

As **diversity** in the workplace plays a role on how discriminatory a system is, the logical solution should be that companies in the AI sector hire more diverse employees. However, West et al. [13] mentioned that there are pipeline studies that investigate the reasons for this lack of diversity. It was identified that most of these studies are limited to the representation of women, and only consider binary genders. Also, the topic of women's representation is much more frequently discussed than other topics such as race. It is mentioned in [13] that these studies point out factors that most tech companies in the AI sector refer to, as reasons for their lack of diversity. Generally, the companies claim that there is a lack of diversity in the hiring pool itself, and ignore the pipeline studies' limitations. In reality, this is an excuse because there exist companies that did a good job in creating a diverse workplace that includes multiracial employees [13]. This means that there is more potential in the hiring pools than some companies claim.

The **lack of inclusivity** shown in some systems suggests that developers should create more inclusive systems. Some systems, like Facebook, started to participate in this by adapting to the culture of gender fluidity. As a result, Facebook now offers users the choice of picking their gender from 58 options instead of 2. Although this may appear as a progressive move, Facebook continues to apply binary schemes in their algorithm, in order to serve the goals of marketing, monetization and to increase ad revenue [110]. This makes their attempt of gender fluidity a mere front for the users. It also suggests that misrepresentation issues, in the case of Facebook, are rooted in the algorithm and difficult to solve due to economical constraints and the way marketing is currently implemented.

5.3. Initial guidelines to counter negative social impact

Based on our findings, we propose a set of initial guidelines to counteract negative impacts in AI systems that support the value-based AI principles of the OECD.[3] They contribute to overcoming the gap in between where the AI Act prohibits, e.g., deploying subliminal

[3] https://oecd.ai/en/ai-principles.

techniques[4] and proactively developing benevolent and socially just systems [111]. We use the impacts and factors discovered in our study, plus our background knowledge on ethical research and development, to formulate concrete actions that AI companies and engineers can take when developing AI systems:

5.3.1. Reassess the need for certain AI systems

Before developing AI systems, reassess whether the system is appropriate and necessary. Consider whether the use of AI could lead to unfair or discriminatory outcomes. Be particularly mindful of systems that used face recognition, that judge people by their physical characteristics and/or that use potentially discriminatory classifications, such as beauty or sexual orientation.

5.3.2. Address potential for misuse

Consider the potential for AI systems to be misused or abused, particularly in politically sensitive contexts and when systems can put historically marginalized people under threat by exposing them. The latter can be particularly dangerous with AI that generates conversations with users and uses sensitive and private data. Design systems with safeguards to prevent their use in harmful or unethical ways and that can withstand potential abuse. In other words, ensure that the design process includes considerations of potential misuse.

5.3.3. Ensure diverse and representative datasets

Ensure that training datasets are diverse, inclusive, and representative of all groups. This includes considering race, gender, and other historically marginalized groups. Be especially aware of people of color, as well as non-binary or gender-fluid people. Challenge datasets that may perpetuate stereotypes or discriminatory behavioral patterns. Actively seek out diverse sources of data, correct for less represented groups, and regularly update datasets to reflect changes in society.

5.3.4. Design for inclusivity

Create AI systems that are inclusive and sensitive to the needs of all users, including those with disabilities and non-binary or gender-fluid identities. Provide flexible and inclusive options for users, and avoid enforcing binary categories where they are not necessary. Ensure that systems are adaptable to different cultural, social, personal identities.

5.3.5. Involve diverse teams in development

Related to the previous two points, actively recruit diverse teams of developers, including people with disabilities and from different gender, racial, cultural, and socio-economic backgrounds. A diverse team helps to include a range of perspectives for reducing the potential for biased systems. Avoid over-representation of any single group, particularly in industries like tech, where certain demographics are often underrepresented.

5.3.6. Address existing social patterns and prejudices

Recognize that social patterns and prejudices can be embedded in data, such as through language and cultural biases. Be very mindful of "apparently innocent" functionalities, such as translation and text generation. Design AI systems that detect and mitigate these biases, and ensure that gender, race, and other characteristics are not inappropriately embedded in algorithms. Be particularly careful when such systems can be used by law enforcement and to give or deny people health and financial opportunities.

[4] https://artificialintelligenceact.eu/article/5/.

5.3.7. Focus on the well-being of users

Prioritize the well-being of users over corporate interests like cost-cutting, maximizing profits, or satisfying shareholders/direct users expectations. Pay particular attention to systems that can cause stress, negatively affect self-esteem, or who can influence opinions and perspectives. Ensure that AI systems are designed with the goal of improving society at large, such as ensuring equal opportunities.

5.3.8. Conduct thorough and inclusive testing

Test AI systems on diverse groups of people, particularly those from historically marginalized communities. Ensure that the performance of the system is evaluated across different demographic groups to identify and address potential biases. Test often to ensure that the system adapts to changing societal needs.

5.3.9. Implement transparency and accountability

Ensure transparency in the development and decision-making processes of AI systems. This is specially the case for systems that can generate financial, physical or psychological harm, such as denying people healthcare and economic opportunities. Establish clear accountability structures to address any negative outcomes, reduce as much as possible the burden of proof on the affected person, and continuously improve the system based on feedback. Make the underlying algorithms and data sources accessible for inspection and auditing.

5.3.10. Continually improve and update systems

Establish a system of continuous improvement and feedback to ensure that AI systems remain relevant and ethical. Regularly evaluate the impact of AI systems on different groups, and update them to address emerging issues or new biases. Foster a culture of ongoing learning and adaptation within the organization.

5.3.11. Consider open source development

Open-source methods can improve transparency, accountability and collaboration, as it allows independent experts, other stakeholders and the community to examine, audit and build upon these systems. However, many ethical concerns (e.g., biases in data or inadequate governance) can still persist whether systems are open versus proprietary, and in certain cases, opening up may not be feasible due to legal, commercial, or privacy constraints.

5.4. Limitations and threats to validity

Our study has some limitations that should be recognized.

External validity concerns the ability to generalize your findings beyond the specific context of your research. In terms of **generalizability**, this study is exploratory and based on qualitative analysis. We are not claiming that the study discovered all possible negative social impacts. Hence, the results are not necessarily generalizable. The main threat to external validity is related to the use of Google Scholar and Google to search for articles. Giustini and Boulos compared the results of search on a systematic literature review with the one generated by a combination of Google Scholar and Google, finding that together this search engine found about 95% of the papers in their control group [112]. The authors concluded that the process was inefficient and unsuitable for a literature review. Yet, the authors limited themselves to try to find all the papers included in their control group, not analyzing whether the search engines also returned other papers that were relevant, but have been missed by the original study. In another paper, Vanhala et al. [113] acknowledge Giustini and Boulos's finding and yet chooses to use Google Scholar, based on its multidisciplinary nature and arguing that 95% were good enough for their purpose [113]. We made a similar decision, by choosing to use these search engines due to their multidisciplinary nature, but also because we were also interested in collection the discussion of AI impacts from the grey literature. In order to increase the coverage of the search, we also

performed selective snowballing in the papers. Having said that, we do acknowledge this decision potentially adds a sampling bias, as the search on Google Scholar and Google may favor more accessible or widely cited articles, potentially excluding relevant but less popular research. Additionally, the search algorithms in both platforms prioritize certain results based on factors like relevance or citation count, which could skew the findings towards more well-known studies or specific perspectives. Moreover, limiting the search to Google Scholar and Google may not capture the full range of available literature that might have been returned by other databases (e.g., ACM digital library, IEEE Xplore, or Scopus). It is also worth noting that many impacts that were highlighted in our study aligned with what was found in other research. For example, that these systems generally affect women and people of color [13].

Construct validity is concerned with how well the study measures the concepts or constructs it intends to measure. In our study, these include the potential for unclear or inconsistent definitions of key terms like "discriminatory AI", which could lead to subjective interpretations across different studies. The search terms, or the combination of them, potentially also have led to overlooking studies that discuss related concepts using different terminology. For instance, studies that discuss "algorithmic bias", but do not use the terms we selected, may still address similar concerns. Furthermore, snowball sampling may reinforce specific perspectives, limiting the range of effects and causes considered. In addition, focusing on systemic effects based on previously reported ones could result in missing emerging or less-recognized impacts. Finally, searching for potential causes for the reported impacts may be biased by existing literature that predominantly emphasizes certain factors (e.g., algorithmic bias), potentially overlooking other underlying causes.

Internal validity focuses on whether the conclusions you draw about the negative social effects are valid and not influenced by confounding factors. Threats in our study include the potential influence of confounding variables, such as socio-economic, cultural, or political factors, which could affect the reported negative social effects and make it difficult to isolate the role of AI systems in causing those effects. Additionally, the interpretation of the findings could be biased by the reviewers, as inconsistencies or subjective judgments may arise if the effects are not analyzed and categorized in a systematic and objective manner, potentially leading to skewed conclusions about the social impacts of AI. To mitigate this threat, the first two authors performed the analysis steps, the fourth author reviewed each step in discussion, and the third author reviewed the entire analysis at the end.

5.5. Implications on research and practice

Our research questions are answered as follows:

RQ1 **"What are the types of negative social impacts that existing AI systems cause?"** We found inequality in opportunities, potential for malicious use, law breaking, directed policing, privacy violation, gender discrimination, racial discrimination, ethnic discrimination, genetic discrimination, negative financial impact, negative impact on education, wrongfully flagging, damage of self-esteem, triggering of stress, perpetuation of division of socio-economic classes, lack of diversity in the workplace, triggering dysfunctionality in families, threats to safety, perpetuation of gender inequality, ethnic discrimination and stereotypes, negative health impacts, influencing of opinions, and decrease of trust in authorities.

RQ2 **"What are the probable common factors that can cause these negative social impacts?"** We found misrepresentation and poor diversity in datasets, proxy bias, existing social patterns and prejudice, inappropriate use of AI, and vulnerability to manipulation.

We see the following major actions for **practice**:

(1) **Apply the Question Zero**: *Reassess the Specific Need for an AI System* when considering the development of one or the subscription to a service. What are we trying to accomplish and why?
(2) **Apply design guidelines** (Section 5.3): Address Potential for Misuse, Ensure Diverse and Representative Datasets, Involve Diverse Teams in Development, Address Existing Social Patterns and Prejudices, and Focus on the Well-Being of Users.
(3) **Apply implementation guidelines** (Section 5.3): Conduct Thorough and Inclusive Testing, Implement Transparency and Accountability, Continually Improve and Update Systems, and Consider open source development.

We see the following major implications on **research**:

(1) **Research on AI**, but really for any research: Ask Question Zero - what are we trying to accomplish and why? Followed by the question: should we do this?
(2) **Development of AI systems**: For research involved in the development of AI systems, we urge researchers to apply the same guidelines as for the practitioners above.
(3) **Ethical Reflection**: Given what we know about the negative social impacts of many (if not most) AI systems, we see an increase in importance of ethical research reflection, both in personal and institutional practice.

6. Conclusion

This study investigates the relationship between AI systems and social sustainability by examining the enabling and systemic impacts these systems have caused or could potentially cause. The research is guided by two primary questions: RQ1, "What are the types of negative social impacts that existing AI systems cause?" and RQ2, "What are the probable common factors that can cause these negative social impacts?" Through a multi-vocal literature review, we collected examples of AI systems with negative social impacts and identified commonalities among them. We categorized these impacts into enabling and systemic types and explored the underlying factors that could have potentially contributed to these harms. We also present some initial guidelines on how to mitigate these impacts and factors.

Our findings highlight that the diversity of negative social impacts calls for more comprehensive and thoughtful measures in the development and deployment of AI systems. While existing literature addresses some of these factors, there is still considerable room for improvement in the current ethical guidelines governing AI development. Social sustainability must be a central concern in the creation of AI technologies, ensuring that these systems do not exacerbate existing social inequalities or contribute to harm. The identified systems and their corresponding impacts underline the necessity for developing a methodology that will guide the creation of more AI technologies that contribute to social sustainability, instead of harming it.

The question of *how to create meaningful change* remains, in between the AI Act, codes of ethics, and initial frameworks and guidelines for ethical AI, so far they all fail to "mitigate the racial, social, and environmental damages of AI technologies in any meaningful sense" [6, p. 869]. Technological systems amplify the injustices and inequalities of the social systems they were built upon, hence the key is to evolve the underlying social and power structures into equitable ones, which subsequently can be mirrored in its technology.

Table 7
Impacts and potential causing factors of selected systems.

ID	Nickname: Description	Enabling impact (category)	Potential systemic impact (category)	Factor
S1	**Beauty scorer:** Beauty.ai was an AI system that evaluated and scored people's attractiveness based on facial features [37].	Out of a diverse set of contestants, the algorithm selected almost only white contestants as winners [37] (**Racial discrimination**).	When winners of such contests are presented as the "most beautiful" people it contributes to a misrepresentation of non-white groups on media (**Perpetuate stereotypes**). It is pointed out in [114] that such racial misrepresentation on media is related to acts of discrimination and inequality (**Racial discrimination**. In other words, presenting white people as attractive and dark-skinned people as unattractive on media has an effect on people's perception (**Influence opinions**, **Perpetuating stereotypes**). The author [114] highlight that negative perceptions of people could trigger discriminating acts. Such acts are, for example, unwelcoming these groups or questioning their rights (**Perpetuate Racial discrimination**). There is also a relation between ideals portrayed by media and self appearance satisfaction according to [82] (**Damage self-esteem**). It is described in [82] that there is a link between eating disorders and dissatisfaction.	The algorithm for "Beauty.AI" was taught to assess attractiveness and beauty based on a dataset of photos. It has come to light that, although there could be more issues, the main one was that the dataset did not include enough pictures of minority groups [37] (**Misrepresentation/poor diversity in the dataset**). Furthermore, one could challenge the purpose of the system, as it brings no real benefit for the society, as well as the potential for harm as it can encourage discrimination against certain groups and reinforce harmful beauty standards (**Inappropriate Use of AI**).
S2	**COMPAS recidivism predictor:** COMPAS is a risk assessment tool used in the criminal justice system to evaluate a defendant's likelihood of reoffending. This score is used in courtrooms and helps determine the time of release for prisoners [38].	The algorithm showed racial bias towards African Americans as black defendants more often received a high score yet did not re-offend, compared to white defendants who more often received a low score and did re-offend [80] (**Racial discrimination**, **Ethnic discrimination**).	The tool helps determine the length of sentences, and long incarceration have negative impacts on prisoners. It is pointed out in [83] that inmates with a long sentence have a higher stress level and worse self-esteem than those with a shorter one (**Trigger stress**, **Damage self-esteem**). Generally, long-term imprisonment creates a feeling that one's life has been wasted and that many people who are convicted with such a sentence experience feelings related to trauma. When a person is convicted, the family of that person is instantly affected. For example, a family would then have to rely on one source of income, instead of two. If more African Americans are imprisoned for a longer time, Children of those incarcerated individuals face prolonged hardship, possibly growing up in financially strained households with limited opportunities (**Perpetuate racial discrimination**, **Perpetuate division of socio-economic classes**). The author of [115] mean that children become victims of their parents' imprisonment. (**Triggering dysfunctionality in families**)	The inner workings of the algorithm are not disclosed, as mentioned in [38]. They investigated whether the issue was due to including protected attributes, but concluded that this was probably not the cause. Later they found that education levels and job-status were included, and could contribute to the bias. However, one could make several arguments against the existence of COMPAS on the first place, such as it influences life-altering decisions on individuals that are based on group statistics rather than evaluating the unique circumstances of each defended, not to mention that the system works as a black-box making if difficult for people to contest its reasoning (**Inappropriate use of AI**).
S3	**Amazon résumé scanner:** Amazon's recruitment tool was an AI system designed to assist in screening and ranking job applicants' resumes [1].	The tool turned out to be discriminatory towards female candidates [1] (**Inequality in opportunities**, **Gender discrimination**). By favoring male candidates, the tool denies equally-skilled women from a well paid job in the tech industry (**Negative financial impact**).	The tool could cause negative stereotypes to be generated and in that way discourage females from entering a particular field. In this case, the tech field. (**Perpetuating stereotypes**). It is expressed in [13] that lack of diversity in development teams in tech companies, such as at Amazon, have an influence on whether their systems turn out to be discriminatory (**Lack of diversity in workplace**). As women are discouraged to enter the field, there is a risk that this contributes to perpetuating discrimination and biases against females (**Perpetuate gender inequality**). In addition to that, exposure to unemployment during one's life could cause long-term mental health scarring, which is shown in a study [116] (**Negative health impact**). Finally, systems like this could exacerbate the existing economic gender gap. If the algorithm is biased against women or other marginalized groups, it may disproportionately filter out qualified candidates based on patterns in past hiring data, reinforcing gender inequality in the workplace and widening the socio-economic divide. (**Perpetuate division of socio-economic classes**)	The algorithm Amazon used was trained on a dataset that included resumes from applicants, submitted to the company during a 10-year period. Those resumes mainly came from men, due to that more men had applied to the tech industry. This caused the algorithm to assume that men were more desirable for the job. As a result, any resume containing the word "women" were disregarded by the algorithm [1] (**Misrepresentation/poor diversity in datasets**).
S4	**HireVue interview analyzer:** HireVue is an AI-powered recruitment tool that analyzes video interview responses to assess candidates' suitability for a job [39].	As people who are living with a disability express and conduct themselves in another way than the norm, the recruitment tool may not recognize their ways of conduct. This may cause the tools integrated facial analysis to eliminate these candidates even though they are qualified [39]. According to [117], it is illegal to discriminate against someone based on their genetic information or disability in the workforce (**Genetic discrimination**, **Inequality in opportunities**, **Law breaking**). When candidates are repeatedly eliminated from jobs that they are perfectly skilled to do, they may be forced to go to a less-skilled job (**Negative financial impact**).	The system contributes to preserving the low percentage of employed people with disabilities. As shown in [118], 17.9% of people with a disability were employed in 2020. This compares to the 66.3% of people without a disability that were employed at the same time. Routine discrimination against people with disabilities, can only perpetuate this situation (**Perpetuate division of socio-economic classes**, **Lack of diversity in the workplace**). According to [84], work does not only provide a source of income, it also provides a sense of purpose and self-worth to individuals. In order to feel included in their communities and to grow social connections, it is therefore crucial that individuals with disabilities are given the opportunity to be employed. (**Damage self-esteem**)	The dataset lacked representation of different individuals and the algorithm lacked diverse training that took into account the characteristics of people with disabilities who are later successful in their jobs [39]. If you train a system on data generated from good employees within a company, and you do not have any individuals with disabilities, the system is likely not going to prefer a person with a disability [119] (**Misrepresentation/poo diversity in the dataset**).
S5	**Apple credit evaluator:** Apple's credit card system uses an algorithm to assess applicants' creditworthiness and determine credit limits.	The algorithm was criticized by customers for favoring men. In one case, the algorithm offered a man a credit limit 20 times higher than what it offered his wife, denieing her form teh opportunity to have finatial aids for her projects (**Negative financial impact**). This happened even though his credit score was worse and they filed their tax returns jointly [40](**Gender discrimination**, **Inequality in opportunities**).	The system could contribute to increasing the financial gap between males and females, provide unequal opportunities and maintain power differences in gender related roles. This is discussed by Tharenou [120], who argues that pay gaps between genders contributes to a lower status for women in society and that it helps ensure "that the traditional gender-influenced hierarchical power structure is maintained" [120, p.203] (**Perpetuate division of socio-economic classes**, **Perpetuate gender inequality**, **Perpetuate stereotypes**). Similarly, such system, that limits accessibility to credit, prevents equal access to money and limits financial freedom. These effects, for example, have an impact on dynamics in families. Moss [89] mentioned that the household is an vulnerable setting where conflict of power and control can occur. Hence, a wide hierarchical difference that results from one's job and salary, could potentially increase such conflicts. (**Triggering dysfunctionality in families**)	The company and developers of the algorithm did not seem to know themselves how the algorithm worked or why it gave a certain output. It is discussed that the algorithm did not use gender as an attribute and that it therefore could not be discriminating based on it. However, it is mentioned that there could be proxies that caused the algorithm to include such biases [97] (**Proxy bias**).

(*continued on next page*)

Table 7 (*continued*).

S6	**Tay Social bot:** Tay was an AI chatbot launched by Microsoft on Twitter, designed to learn and interact with users in real time [41].	The chatbot tweeted racist and misogynistic tweets after other users shared provoking tweets that encouraged the chatbot to follow these themes [41]. (**Potential for malicious use, Gender discrimination, Racial discrimination**)	The existence of accounts such as Tay, which are easily influenced by negative reinforcement on social platforms, could end up influencing the opinions of users on social platforms. According to [121], people's opinions are in fact influenced by the general public's opinions on societal issues, that they see on social platforms (**Influence opinions**).	According to Microsoft, Tay learned from "trolls" who suggested phrases and caused the chatbot to search the internet for replies that fitted their tweets [100]. (**Lack of robustness - for external manipulation**)
S7	**Health forecaster:** In the US, the most prominent Accountable Care Organizations (ACOs) Accountable Care Organizations (ACOs) use algorithms to predict which patients are likely to need complex and intensive healthcare in the future. Those that are identified by these algorithms are then enrolled in care management programs where they receive additional resources and attention [42].	It was found that these algorithms included racial bias which in turn had an impact on who was enrolled in these programs. More specifically, it was discovered that black patients that received the same risk score as white patients, while they were in fact sicker (**Racial discrimination**). This means that white patients get enrolled into these programs, even though black patients with a lower score are equally sick [42]. (**Inequality in opportunities**)	According to [122], racial and ethnic minorities generally receives lower quality of care, and tend to experience greater morbidity and mortality. Even though these tools should work to improve the overall health in the country, the bias it contain causes it to improve the health of mainly white people. So, when this tool is used by care organizations, it preserves these race-based health disparities and contributes to an existing problem (**Perpetuate racial discrimination**, **Negative health impact**).	One of the issues is that the algorithm bases its decisions on health costs, meaning that it uses health costs to assess the need for care. Due to, for example, unequal access to health care, less money is generally spent on black patients. Since less money is spent on black patients, the algorithm believes that they are healthier than white patients, even though they may be equally ill [42]. (**Proxy bias**)
S8	**Google text translator:** Google Translate is a web-based tool that translates text, documents, and websites between multiple languages using machine learning [87].	The translation program assigns genders to professions and activities when translating from gender-neutral languages such as Finnish, Filipino and Hungarian. This translation has shown to be sexist [87] (**Gender discrimination**).	In a study [123] it is described that sexist language can negatively influence women's motivation and identification, and that it can trigger ostracism. It is mentioned that "ostracism threatens basic needs such as belonging, control over one's life, self-esteem and the need for meaningful existence" [123, p.63] (**Perpetuate stereotypes**, **Damage self-esteem**). Another paper [124], mentioned that sexist language reduce the importance of women as a social category and that such language maintain inequalities (**Perpetuate gender inequality**). In addition to that, [125] argued that sexist language perpetuate social roles which favor men, and that such language may contribute to an under-representation of women in male-dominated jobs. (**Lack of diversity in workplace**)	Since systems learn how to translate by analyzing a huge amount of examples, they will learn to connect certain words with a certain gender based on how it is used in those examples [87]. That certain occupations are given a certain gender is therefore a result of the poor diversity that those occupations have, and the embedded sexism our society perpetuate (**Existing social patterns and prejudice**).
S9	**Speech converter:** Speech-to-text services that convert spoken language into written text. Koenecke et al. [43] analyzed potential biases in such services provided by Amazon, Apple, Google, IBM and Microsoft.	Systems generally made fewer errors when processing an audio-snippet from white speakers than from black speakers. For white speakers the systems made 19 errors for every hundred word and for black speakers it made 35 errors. For black men, the systems performed even worse, with 40 errors for every hundred word [43]. As speech recognition is being integrated in services such as hiring processes, immigration decisions and transportation, those who struggle with being understood by such systems may be prevented from, for example, getting hired or moving to a new country [126] (**Inequality in opportunities**, **Negative financial impacts**, **Racial discrimination**). Louise Kennedy was denied permanent residency in Australia as she did not pass the test for oral fluency. The system assessing her used voice recognition technology to test speaking ability, but even though she is a native English speaker, the system did not understand her well enough [127].	The impact of Hispanic accents among applicants in job hiring processes was investigated in [88]. The participants of the study made decisions regarding job suitability and chances of a promotion. The results showed that applicants with Hispanic accents had a lower chance at getting the job in comparison to a standard American English speaker (**Perpetuate division of socio-economic classes**). The participants also viewed the applicants as less likely to get promoted in the job. (**Lack of diversity in workplace**) It is suggested in [88] that applicants with Hispanic accents experience access-related discrimination and treatment-related discrimination. The study also shows that accents influence important decisions that in turn influence economic classes. Using voice recognition systems that affect such decisions would continue and cement this discrimination, which in turn would affect the economic status of its users and influence their opportunities. If, for example, our phone's voice assistant only understand white speakers, then the data that gets collected comes mainly from white people. This results in that new voice assistants continue to only work for white people, and people with other backgrounds get left out (**Perpetuate race discrimination**). This means that not everyone can take advantage of assistive tools, and the gap in economic classes is therefore further nurtured.	One of the issues could be that databases contain less data of minority and women voices. For example, [126] bring up that TED talks are commonly used by speech scientist, and 70% of the people that hold TED talks are male. It is also mentioned in [43], that these systems are trained on data which lacks diversity and that a more diverse training set could reduce these differences in performance (**Misrepresentation/poor diversity in the dataset**).
S10	**Upstart smart lender:** Upstart is an AI-driven lending platform that uses machine learning to evaluate creditworthiness and provide personal loans, considering factors beyond traditional credit scores [44].	It is found that graduates from historically black and Hispanic colleges and universities, are assigned higher rates for their loans compared to graduates from institutions that are attended by people belonging to non minorities [44] (**Negative financial impact**, **Racial discrimination**). Higher interest rates may prevent them from attending their preferred educational institutions (**Inequality of opportunities**, **Negative impact on education**).	According to [128], lending systems generally make their decision based on the likelihood that the lender will be able to pay back the loan. This could explain the acting of this system. However, a study [129] conducted on students in the UK, showed that students coming from lower economic classes are more fearful of debt, which influences their attitudes towards higher education. Systems like Upstart might trigger the same kind of attitude for students applying to black/hispanic colleges, and as a result deter them from higher education (**Perpetuate racial discrimination**), and keep future generations from progressing financially in life (**Perpetuate division of socio-economic classes**).	Lender system often use data provided by their users. Though there are also non-traditional data that gets fed into these systems to assess the creditworthiness of applicants. This data includes search history, shopping patterns and social media activity. Such sources can lead to discriminatory decisions. Applicants are probably not aware that such data is being collected, so if they are rejected, it is difficult to know the reason [128] (**Proxy bias**).
S11	**Predictive Patrol officer:** PredPol is a predictive policing tool that uses historical crime data to forecast future crime hotspots and help law enforcement allocate resources more effectively. The police then takes these predictions in consideration when choosing which areas to patrol [45].	Robert McDaniel, a black resident in the south side of Chicago, was placed on the police's "heat list" which contains people who might potentially commit a crime. McDaniel was surprised, given that he has never committed any crimes before [77]. (**Directed policing, Wrongfully flagged**)	Being wrongfully accused of committing or attempting to commit a crime can negatively impact an individual's personality and sense of self, for example when it comes to credibility and dignity [99]. A person who has been wrongfully accused by authorities may also lose trust in authorities. According to [130] unfair treatment from the police in urban locations drive residents to view the police as less legitimate. The study also mentions that perceptions in police affect citizen's willingness for cooperation, and to report crimes. Another study [131] show that there is a relation between negative experiences with the police and one's satisfaction with the police. The study conducted a survey with a sample of white, African American and Hispanic people, and found that generally negative vicarious and personal experiences yield a dissatisfied view on the police. The study also reports that black and Hispanic respondents have reported a higher number of mistreatment from the police. A system like Predpol is likely to cement this attitude and view of the police, and even magnify the differences in reported treatment. (**Decreased trust in authorities, Perpetuate racial discrimination, Damage self-esteem**)	According to [77], police datasets do not include all criminal offences. The article explains that if police heavily patrols a specific area or neighborhood, the data records would naturally over-represent people who live in these areas. The algorithm that uses this data would as a result predict these areas as hot spots, which would motivate more police to patrol the area. This becomes a non-ending loop that feeds into itself (**Misrepresentation/poor diversity in datasets**).

(*continued on next page*)

Table 7 (*continued*).

S12	**Uber driver identifier:** Uber uses Microsoft's Real-Time ID Check to verify drivers' identities through facial recognition [46].	The tool locked out a transgender woman who was a registered driver, which resulted in her missing out on three days of work [46] (**Gender discrimination**, **Wrongly flagged**).	Besides the negative economic impact experienced in this case, in the long-term, social exclusion can lead to health and stress related issues as pointed out in [132] (**Negative health impact**, **Trigger stress**). Often the burden of proof of unfair decisions are placed on the marginalized users, rather than the companies developing the flawed technologies. It such misrepresentations become frequent, affected users may simply give up fighting for it, looking for alternative means of work and life (*Perpetuating gender inequality*). Another paper [133] explains that when trans or non-binary people experience misgendering, it influences their self-esteem, rejection and reaffirms the feeling of social stigmatization (**Damage self-esteem**).	Leslie [109] suggested that there should be ethical questions regarding the justifiability of development in the case of facial recognition systems and surveillance systems. (**Issues in facial recognition**)
S13	**Amazon Face analyzer:** Amazon Rekognition is a facial recognition tool that analyzes and identifies faces in images and videos [47]. The tool is used by police in the United States, for detecting, verifying and analyzing faces [48].	It was found that the tool performs better when identifying light-skinned people, compared to dark-skinned people. A study by MIT Media Lab was mentioned in [79]. The study found that the system falsely identified dark-skinned women as men in 31% of the cases. Meanwhile, the system was able to identify light-skinned individuals with a nearly 100% accuracy. Computer vision used for surveillance disproportionately affects women and dark-skinned individuals (**Race discrimination**). There exists several cases where this happened, for example, Ousmane Bah was wrongly accused for stealing at an Apple Store and Amara K. Majeed was wrongly accused of contributing to the bombings in Sri Lanka in 2019 [53]. Having to respond to the false allegations caused Bah "severe stress and hardship" [134]. Majeed received death threats as a result of the mistake [135] (**Directed policing, Wrongfully flagged**). Finally, when people are publically linked to certain places, activities or other people, due to their faces being recognized by a system, this can also violate their privacy (**Privacy violation**)	The systemic impact is similar to the systemic impact of S11.(**Decreased trust in authorities, Perpetuate racial discrimination, Damage self-esteem**)	According to Leslie [109], there should be ethical questions regarding the justifiability of development in the case of facial recognition systems and surveillance systems, similar to S12. (**Issues in facial recognition**)
S14	**Clearview identity finder:** Clearview AI is a facial recognition system that uses a vast database of publicly available images, such as Twitter, Facebook and Google, to identify individuals, primarily for law enforcement and security purposes. [49]. The data is then used by the algorithm to create a "faceprint" of individuals that clients, such as the Detroit police, can use for identifying people. The data that is collected also includes people who are captured in the background of complete strangers images.	The system wrongfully matched a black American named Robert Julian-Borchak to a crime he did not commit. The system matched his face to an unclear image obtained from a surveillance store tape. While it was discovered later as a mistake, the police had already arrested him, interrogated him and imprisoned him overnight [49] (**Directed policing**) Like S13, this system can also invade someone's privacy by publically linking them to places, activities or people (**Privacy violation**).	As suggested in [136], facial recognition used for surveillance could negatively impact the power balance between citizens. The same study emphasized that such systems threaten individuals privacy [136]. Another systemic impact is the systemic impact that is listed for S11. (**Decreased trust in authorities, Perpetuate racial discrimination, Damage self-esteem**)	Leslie [109] suggested that there should be ethical questions on the justifiability of development in the case of facial recognition systems and surveillance systems — similar to S12 and S13. (**Issues in facial recognition**)
S15	**Examplify exam monitor:** Examplify is a secure exam-taking software used by schools and universities to administer online assessments. The application uses face recognition to allow students to sign in [50].	Khan, a dark-skinned student could not sign in to his exam, as he was presented with a message saying that the system was unable to identify his face due to poor lightning. In order to solve the problem, he had to contact customer service and the matter took a couple of days to solve [50]. (**Racial discrimination, Negative impact on education**)	Surveillance systems, like Examplify, can help to widen the adoption of online testing, which was in itself has been linked to negative impact students' habits such as sleep and eating [137] (**Negative health impact**). In addition, if technological problems during online exams are perceived as common place, especially by a certain minority, this can lead to high level of anticipatory stress (**Trigger stress**, **Perpetuate racial discrimination**).	Leslie [109] suggested that there should be ethical questions regarding the justifiability of development in the case of facial recognition systems and surveillance systems — similar to S12, S13, and S14. (**Issues in facial recognition**)
S16	**Proctorio secure examiner:** Proctorio is a remote proctoring software that uses facial recognition and monitoring tools to prevent cheating during online exams [51].	A black woman expressed that every time she used the tool it requested that she should shine more light on her face in order to validate her identity [51]. (**Racial discrimination, Negative impact on education**)	The systemic impact is similar to the systemic impact of S15 (**Negative health impact**, **Perpetuate racial discrimination**, **Trigger stress**).	Leslie [109] suggested that there should be ethical questions regarding the justifiability of development in the case of facial recognition systems and surveillance systems — similar to S12, S13, S14, and S15. (**Issues in facial recognition**)
S17	**Giggle girls social networks:** Giggle is a networking app designated for girls-only. The app verifies that users are girls by prompting the user to take a selfie when signing up for the platform. By using "bio-metric gender verification software" the app then confirms the gender of the new user [52].	The verification software used by the app struggles to verify trans-girls, causing them to be locked out of the social networking platform [52]. (**Gender discrimination, Genetic discrimination**)	The systemic impact is similar to the systemic impact of S12. (**Negative health impact, Trigger stress, Damage self-esteem, Perpetuate gender inequality**)	The issue with this tool is not necessarily in the way the algorithm was designed, but rather in the choice of using the verification software. In today's society, it is a rather poor solution to use people's bone structure in order to verify their gender. Gender is concerned with the behavior, roles and expressions that we relate to, rather than the biological attributes that we were born with [138]. Someone who has the bone structure of a male, like many trans-girls do, may not identify as a male even though their biological attributes says so. (**Inappropriate use of AI**)
S18	**Google image analyzer:** Google Cloud Vision is an image recognition service that uses machine learning to analyze, label, and extract information from images [53].	A recent experiment looked into hand-held thermometers as they have become increasingly used as a result of the Covid-19 pandemic. When inputting an image where a dark-skinned individual held the thermometer, the system labeled the image "gun". However, when the thermometer was held by a light-skinned individual, the picture was labeled "electronic device" or "monocular" [53]. Flaws in Google's image labeling was also seen in 2015 when the tool labeled two black people as "gorillas" [139]. (**Racial discrimination**)	Tools that aim to recognize any kind of weapons are commonly used in places such as schools, concerts and malls. In some countries, law enforcement even use automated surveillance. It is likely that those systems perpetuate similar biases as what is seen Google Vision Cloud. Therefore, dark-skinned people risk being pointed out as dangerous even when they, for example, are holding a regular object [53] (**Perpetuate Racial discrimination**, **Decreased trust in authorities**).	One issue may be that dark-skinned people are more commonly seen in violent settings in the datasets that the algorithms are trained on. When the computer attempts to label the image,it is therefore more likely to choose a term related to violence [53]. (**Misrepresentation/Poor diversity in dataset**))

(*continued on next page*)

Table 7 (*continued*).

S19	**Cambridge Analytica data harvester:** The "thisisyourdigitallife" app, developed by Cambridge Analytica, collected personal data from Facebook users and their friends to build psychological profiles for targeted political advertising [54].	The collected data from Facebook was especially likes and friends list, the user's name, contact details and location. The data was fed into a model that became able to make personality predictions [54] that were used to customize political messages and agendas in order to sway people's opinions during America's presidential election in 2016 [69]. As a response, the federal trade commission sued Cambridge Analytica's former chief executive and an app developer of the system [76]. (**Potential for malicious use, Law breaking, Privacy violation**)	Wolleey et al. discuss digital misinformation and manipulation is [94]. It mentioned that automated software products can be used to create a "manufactured consensus" and to make people believe that the general public supports a certain idea (**Influence opinions**). In a study [121] it was pointed out that someone's pre-existing opinions influence that person's perception on societal matters. Other factors are mass media messages and interpersonal discussions, which are all possible in social media website like Facebook. Hence, curating political messages based on someone's pre-existing views could strengthen this bias. (**Influence opinions**)	Like the systems described by Leslie [109], this system that has dubious purposes. It exploited personal data without transparency or consent, violated privacy, and enabled the manipulation of political opinions. By collecting sensitive information like Facebook likes, friends, names, contact details, and location, without proper disclosure, the system violated individuals' rights to control their own data. Its use to tailor political messages and sway public opinion during the 2016 U.S. presidential election raised serious ethical concerns, as it allowed for the creation of targeted, manipulative content that exploited people's pre-existing biases. The system also facilitated illegal activities, such as data misuse and law-breaking, and contributed to the spread of misinformation and the potential for social division. This kind of technology undermines democratic processes, erodes trust in institutions, and sets a dangerous precedent for digital manipulation. (**Inappropriate use of AI**)
S20	**Theft scorer:** The "Sensing project" is implemented by the police in Roermond, Netherlands. By using cameras, the police collect data of vehicles in the area in order to find potential pickpockets or shoplifters. The collected data was analyzed by an algorithm that then outputted a prediction in the form of a risk score [55].	The design of the system becomes biased against Eastern European nationalities and/or Roma ethnicity as it focuses on "mobile banditry" being defined as carried out by those ethnic groups [55]. The individuals who are given a high risk score may be stopped by the police without knowing that they are being stopped for this reason. As explained in [98], proactive policing methods may decrease crime rates but they also often violate innocent peoples right to privacy. According to [55], the Sensing project also breach data protection rights. The report by Amnesty found that people of Dutch nationality in reality account for 60% of the people suspected of pickpocketing and shoplifting, while Eastern European people only account for around 22%. That the algorithm is designed to assign a rating to only Eastern European people therefore reinforces existing preconceptions. (**Directed policing**, **Privacy violation**, **Ethnic discrimination**, **Wrongfully flagged**)	The systemic impact is similar to the systemic impact of S11. (**Decreased trust in authorities**, **Perpetuate ethnic discrimination**, **Damage self esteem**)	The report by Amnesty [55] mentions that prejudices and stereotypes play a role in decisions made by the police in Europe. It also mentions that police records may be biased and not always reflect the truth when it comes to for example crime rates (**Misrepresentation in dataset**, **Inappropriate use of AI**, **Existing social patterns and prejudice**).
S21	**Facebook ads:** The Facebook ad delivery system is used by companies to promote their products and services. The system uses an ad auction and machine learning to point ads to the appropriate people at the right time [56].	A study by Imana et al. [92], found that Facebook's ad delivery can result in a "skew of job ad delivery by gender", also when controlling the qualification variable. The study [92] found that the algorithm shows different jobs to females compared to males, even though the displayed jobs require the same qualifications. For example, the algorithm targeted males when promoting jobs as software engineers for Nvidia and females for the same job at Netflix, the algorithm also targeted males when promoting sales associates for cars and females when promoting sales associates for jewelry. (**Gender discrimination**, **Inequality in opportunities**)	Several researches have investigated the impact diversity has on a firm's performance. A paper by Hunt et al. [140] brought up this question and lists multiple areas in which diversity have a positive effect. These areas are "advantages in recruiting the best talent, stronger customer orientation, increased employee satisfaction, and improved decision making" [140, p.9]. (**Lack of diversity in workplace**) That Facebook is targeting ads to people based on their genders will have a negative effect on those companies' performance. It also perpetuates existing gender biases. Having the ability to control who is shown an add, like Facebook have, also allows them to impact who receives a crucial economic opportunity. (**Perpetuate gender inequality**, **Perpetuate stereotypes**, **Perpetuate division of socio-economic classes**)	The reason is unknown as Facebook chooses to not disclose how their algorithm works [141]. (**Unknown**)
S22	**Sexual orientation predictor:** Wang and Kosinski developed a system that predicts someone's sexual orientation based on their pictures [57]. They reported the system to have accuracy of 81% at predicting people who identify as homosexual [58].	While the algorithm might have a relatively high level of accuracy, there is a potential that the system can be used maliciously by homophobic organization or governments. For example, according to [70], 71 jurisdictions criminalize homosexual activity, and 11 jurisdictions impose the death penalty on homosexual activity. So, the usage of the system can encourage such governments to track, arrest or harass LGBT people, even if they have not publicly announced their sexuality due to being concerned for their safety. (**Potential for malicious use**, **Directed policing**, **Gender discrimination**)	According to [85], asylum seekers that have experienced persecution for their sexual orientation are at high risk for mental health issues, such as severe stress and depression. In countries that criminalize LGBT, individuals often experience harassment, alienation and restricted access to their rights. The struggle of LGBT asylum seekers continue even after immigration as they often feel alienated because of cultural differences and shame about their persecution history. (**Trigger stress**) A system like this, can enable the identification and harassment of LGBT individuals, reinforcing stigma against this community. This normalization of discrimination can lead people to perceive anti-LGBT sentiments as acceptable, further encouraging harassment and persecution. (**Threat to safety**) Flores [142] explains the relation between LGBT people and mental health: sexual and gender minorities experience stress and anxiety different from what most other people face in their daily life. It is also explained that minority stress experienced by LGBT people cause poor health outcomes.(**Negative health impact**)	Like the systems described by Leslie [109], this is another example of a system that has dubious purposes. Wang and Kosinski [57] did not set out to discriminate - they aspired to advance the understanding of the origins of sexual orientation and the limits of human perception and note that, given that companies and governments are increasingly using computer vision algorithms to detect people's intimate traits, their findings expose a threat to the privacy and safety of gay men and women. The sexual orientation predictor system should never have been built due to its potential for misuse and the serious risks it poses to the safety and well-being of LGBT individuals. The algorithm's ability to accurately predict sexual orientation based on facial traits and postures can be exploited by homophobic organizations or oppressive governments, increasing the likelihood of discrimination, harassment, and even criminalization in regions where homosexuality is illegal or punishable by death. Furthermore, the system poses a significant threat to the mental health of LGBT individuals, especially those seeking asylum, by exacerbating stress, anxiety, and minority stress, which already contribute to poor health outcomes. The risk of this technology being used to track, persecute, or harm vulnerable individuals is a clear violation of human dignity and safety, making its creation highly unethical. (**Inappropriate use of AI**)

(*continued on next page*)

Table 7 (*continued*).

S23	**Facial criminal tendencies guesser:** Faception is a system that claims to be able to identify potential terrorists or pedophiles based on images. According to Kosinski, facial features can be connected to criminal tendencies [59].	Todrovo showed disagreement towards this notion and said that the cost of errors are high. According to [59] it is possible that certain biases are encoded in the algorithm that target a specific group. Faception claims that it was able to flag 9 out of 11 of the Paris attackers. However, Sirer showed concern regarding this last claim and mean that any algorithm that singles out people of Arab descent could identify those Paris attackers. He emphasized the risk that if this algorithm relies on facial traits then it will potentially falsely accuse 370 million Arabs out of 450 million [78] (**Ethnic discrimination**, **Directed policing**).	According to [59] this system will reinforce stigmatization based on appearance and possibly ethnicity. If systems such as this wrongly flag certain types of individuals frequently, other people that identify with these individuals might start feeling stressed and afraid that this might happen to them in the future (**Trigger stress**). Furthermore, this kind of technology can be used for purposes other than surveillance. For example, if the system is used for job application or a dating service, and someone is falsely flagged as a terrorist, then it would be difficult to change that impression of the recruiter or date. Such information, when learned during first interaction, can be digested when creating the first impression of someone. According to [143], first impressions or "implicit social cognition" is more stable than explicit social cognition, and can affect long term decisions. This is concluded as a result of observing no effect on implicit social cognition when presented with new information that counter the first narratives. A false positive produced by Faception could in a job application process, trigger the recruiter to have a negative implicit opinion of the applicant even after it has been identified as a false flagging. This would in turn affect that person's likelihood of getting the job. (**Stereotypes**, **Influence opinions**, **Perpetuating division of socio-economic classes**)	Faception did not share the details of the inner workings of their system. Even though [59] claim that they have a high level of accuracy, it is difficult to check the correctness of this statement. (**Inappropriate use of AI**, **Unknown**)
S24	**Autonomous vehicles:** Autonomous vehicles are self-driving cars that use sensors, AI, and machine learning to navigate and make decisions without human input [60–62]	In [60] the social implications of autonomous vehicles were discussed. For example, the article mentioned that implementation of such vehicles will impact the transportation labour force negatively. In the trucking industry fewer people will be needed to oversee the trucks, and the industry will require a skill change from the workers. Also, taxi-drivers may be affected by self-driving vehicles. Another concern with such vehicles are the ethical scenarios that the algorithms must be trained to choose between. Either choice in a decision between one person's life over another's, is ethically incorrect according to relevant professional codes of ethics. Leaving the outcome to chance seems wrong too when there may exist some reasons to prefer one scenario over the other. The ethical issue is that, no matter which strategy the vehicle adapts, a vehicle that is programmed to weigh one collision over another in a way resembles a targeting algorithm. A thought provoking scenario that autonomous cars could face is if the car is programmed to prioritize the health of its driver above all. Achieving this priority might cause more harm and possibly deaths of others. Meaning, setting the priorities of autonomous cars might not lead to the best consequences. The suggested answer to solving these ethically sensitive scenarios could be to give back the control to the driver. However, there could not be enough time to do so. These ethical concerns were discussed in [61]. (**Threat to safety**, **Negative financial impact**)	The introduction of AI in the workforce may create new, currently unimaginable, occupations. Though, it will also cause unemployment for humans in the transportation sector [61]. Besides employment, a concern regarding autonomous vehicles is that, if people know the decision process of these vehicles, it makes it possible for malicious people, such as terrorist and criminals, to manipulate the system. This is brought up in [62] who pointed out that, in order to hinder this manipulation, we need to allow some degree of uncertainty in the decision process. That would in turn introduces other problems. Hacking is also brought up in [61] where it is mentioned that nearly all computing devices have been subject to hacking. If these vehicles can be remotely controlled by owners or authorities, which is under development, they offer an easy-path for hacking. (**Threat to safety**, **Perpetuating division of socio-economic classes**)	If people are not ready or willing to adapt to the change that is forced by these autonomous vehicles, the introduction of it can have negative consequences. If there exists a willingness to undergo a shift in for example the trucking industry, the impacts may not be so devastating. Also, if the society is robust enough to adapt such vehicles, including enforcement of regulations that assigns legal responsibility, the consequences can be lessened. Hence, consequences of autonomous vehicles are dependent on how well we prepare for their existence, and to what extent we prepare the vehicles for ethical scenarios. This aligns with what said by Lin [61], "when technology goes wrong–and it will–thinking in advance about ethical design and policies can help guide us responsibility into the unknown" (**Lack of robustness**) [61, p.81].
S25	**Deepfake video falsifier:** Deepfake is a technology that allows creation of videos that seems to include real people saying and doing things they never really did [63]. Face2Face uses this technology to map and transfer facial expressions from one person to another [64].	A particular case is Helen Mort, who found violent sexual images of herself, where her face had been cropped from a non-sexual image [75]. Helen, as a result, was shocked and sad, and described feeling powerless. Sensity AI is a company working to detect Deepfake videos [75]. They found that 90%–95% of these videos are non-consensual porn. Another case, which is not initially negative but demonstrates the power of deep fake videos, is the well known example of the Peele/Obama video, shared by Buzzfeed [64]. The video shows Obama saying statements that he never actually said. The spread and creation of fake non-consensual sexual videos is only banned in two states in the US [75]. (**Law breaking**, **Potential for malicious use**)	According to [63], the usage of Deepfakes in harmful ways may include misrepresentation in the form of presenting an individual in an undesirable way or ruining someone's reputation. The believability of Deepfakes magnifies the damages of misrepresentation and manipulation. Exploitation is also possible, as identity theft can be feasible. In addition to that it can be used in unsolicited pornography where someone's voice and face can be used to create sexual scenarios in videos. According to [75], Deepfake can also facilitate revenge porn. These possibilities open the door to all kinds of threats, and leave damaging impacts on its victims as a result of this abuse. A study testing the believability of the Peele/Obama video showed that people were uncertain about the video and did not entirely dismiss it as fake [95]. The study suggested that such uncertainty would damage the trust in news on social media and affect the public's collaboration. The authors mentioned that the rise of deep fakes in the political context risks democracy and journalism [95] (**Decreased trust in authorities**, **Influenced opinions**).	According to [144] the way Deepfake technology works is that there are two neural networks, the generator and discriminator. The generator produces the video by using a data set. The other neural network works on distinguishing the video as a fake video or not. If the discriminator labels a video as fake, the generator networks seeks to understand how the discriminator discovered the mistake, and improves it accordingly. Hence, the generator continues to produce better and more "believable" videos in each iteration. It is pointed out in [144] that at some points a fake video will be indistinguishable due to improvements. There is a need to update laws regarding deep fake videos. As mentioned in the enabling impact, unsolicited Deepfake pornography is only banned in two states in the US. On the other hand, revenge porn is banned in 46 states. In England, fake nonconsensual video is not banned, while revenge porn is [75]. Banning one and not the other, when the impacts of unsolicited Deepfakes are so similar to revenge porn, indicates that laws needs to be updated. It is concluded in [75] that it is hard to collect evidence to criminalize the perpetrators. (**Inappropriate use of AI**)

(*continued on next page*)

Table 7 (*continued*).

S26	**Chinese trust scorer:** Chinese social credit system is a data driven system that assigns a "score" to citizens to reward their behavior or to punish them. The score controls the kinds of benefits and rights that someone is entitled to, such as access to private school, air travel and real estate purchases [65].	A concrete example of someone being affected by this system is Lui Hu. According to [71] Lui Hu is a journalist in China who has written about censorship and government issues, which led to fines and arrests. Lui Hu as a result became blacklisted, meaning that he was not able to practice his rights, such as flying. In general it is difficult to recover from being blacklisted or having a low score. (**Inequality in opportunities**, **Privacy Violation**, **Potential for malicious use**)	Kobie [71] mention that Mareike Ohlberg, a research associate at the Mercator Institute for China Studies, expressed concerns towards the system and pointed out that this system will likely increase social class differences in society. The system may also possess faulty data which could lead to a line of negative consequences. For example, faulty data could trigger a lower score or flag authorities. This means that the system could cause cases of being wrongfully accused. As mentioned in S11, being wrongfully accused can negatively impact an individual's personality and sense of self, for example when it comes to credibility and dignity [99]. (**Damaged self esteem**, **Perpetuating division of socio-economic classes**)	Like the systems described by Leslie [109], this is another example of a system that has dubious purposes. In-transparent calculation of social scores is unethical in itself. Moreover, there is no disclosure on who is affected and there seem to be many cases of high correlation of people who criticize the government receiving a low score in the system [71]. (**Inappropriate use of AI**)
S27	**Uighur surveillance officer:** According to [66] Huawei and an AI firm called Megvii tested a software feature called "Uighur alert". The feature is able to detect Uighur people from images. This was discovered by IPVM, a US based company specialized in video surveillance analysis. According to the two collaborating companies, they did not have the intention of releasing the feature.	According to [66], IPVM expressed that such a feature can be used to flag Uighur people and report them to the authorities. The Uighur people in China are a Muslim minority that has been mistreated and oppressed by the Chinese government [145]. "Uighur alert" as a system could become be an addition to an already existing technology dedicated to target this group of people [66]. The Chinese government places Uighur people in camps for "re-education" and to "wash their brains" [72]. In reality the camps are places for cultural genocide, according Adrian Zenz, a leading security expert on the far western region of Xinjiang, the Uighur homeland. Sophie Richardson, a director at Human rights Watch, describes that Uighur people in these camps are exposed to psychological torture [73]. Also, several resources such as human rights organizations and Chinese Communist Party (CCP), revealed that detainees are victims of crimes against humanity [74]. (**Potential for malicious use** and **Directed policing**)	The camps have been described as internment camps as according to [90]. An example of internment camp in history is the Japanese internment camps. These camps had a negative long term impacts on its survivors and future generations. A study [91] showed the long term effects reported by participants of Japanese descendants who are one generation away from the Japanese American internment camps experienced during WWII. The participants of the study reported stories of family and material loss. Some participants reported that their families experienced lost childhood and use assimilated coping strategies' to fit in, which limited their prospects for attaining status and affected their career choice. (**Triggering dysfunctionality in families**, **Perpetuate ethnic discrimination**) Some participants reported that the internment experience influenced their family members' confidence and self esteem. Furthermore, systems like this contribute to institutionalize ethnic discrimination by enabling automated profiling and persecution of ethnic minorities, reinforcing their marginalization. It legitimizes mass surveillance, deepens societal biases, and sets dangerous precedents for AI-driven racial and ethnic profiling worldwide (**Perpetuate ethnic discrimination**, **Perpetuate division of socio-economic classes**).	This is system has dubious purposes, similar to the systems described by Leslie [109]. This system should never have been built because it directly enables and legitimizes the mass surveillance and persecution of an already oppressed ethnic minority. By automating ethnic profiling, it institutionalizes discrimination, reinforcing systemic bias and facilitating human rights abuses. Such technology accelerates and normalizes state-led oppression, making it more efficient and harder to dismantle. Historically, systems that facilitate ethnic targeting, such as internment camp surveillance, have had devastating long-term effects on individuals and communities, leading to intergenerational trauma, loss of identity, and diminished opportunities. Beyond China, allowing such technology to exist sets a dangerous precedent for governments and institutions worldwide, risking the expansion of AI-driven racial and ethnic profiling, further eroding human rights protections on a global scale. (**Inappropriate use of AI**)
S28	**Social media filters:** Social media platforms like Snapchat and Instagram have introduced filters and lenses. The algorithms that these companies have created identify the face or faces which are visible for the camera and applies different types of effects [67,68].	The presence of filters and lenses are contributing to a change in people's perception of beauty. In an article [146] this matter of altered beauty standards is brought up, as well as the fact that Snapchat has been criticized for promoting "thin, westernized beauty ideals; the narrow nose, the lightening effect" (**Perpetuating stereotypes**). The YMCA's Be Real Campaign [147] found that 52% of young people think social media creates an expectation on how people are supposed to look. In a preliminary research by Amy Niu [148], she found that while Americans become more willing to conduct cosmetic surgery as a result of social media filters, Chinese people tend to feel better about themselves when using filters than when not using them.	According to [149] people who suffer from body dysmorphic disorder fixate on nose features, skin and face symmetry. Since filters tend to morph these features, they can influence someone's perception of their nose, skin and face. The same paper [149] describe that therefore, this kind of AI filters can have a negative impact on people who are vulnerable to appearance or body issues. In addition that, stress is associated with body dissatisfaction. An article [150] brought up another view of the problems with filters. It mentioned that filters that are created by fashion and beauty brands tend to smooth, contour and apply makeup to the face which reproduces feminine "hetero-sexy" beauty norms. It was mentioned that "A filter programmed to modify a face by smoothing wrinkles and warming skin tone uses digital code to attempt to 'return' a face to the norm, no matter what performance the person undertakes" [150, p.18]. (**Perpetuate Stereotypes**, **Influence opinions**)	This system perpetuates questionable beauty standards and stereotyping. The social media filter system should not have been built because it perpetuates harmful beauty standards by favoring narrow ideals like thinness, light skin, and small features, which marginalize those who do not conform. It contributes to mental health issues by distorting individuals' perceptions of their appearance, particularly among those vulnerable to body dysmorphia, leading to anxiety and depression. The filters reinforce gendered and stereotypical beauty norms, promoting a "hetero-sexy" image that limits self-expression and strengthens societal expectations. Additionally, the filters set unrealistic beauty standards, encouraging unhealthy comparisons and pressuring individuals to undergo cosmetic procedures. (**Inappropriate use of AI**)

CRediT authorship contribution statement

Nafen Haj Ahmad: Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation. **Linnea Stigholt:** Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation. **Leticia Duboc:** Writing – review & editing, Validation. **Birgit Penzenstadler:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

We would like to thank Gordana Dodig Crnkovic for providing valuable feedback on an earlier version of this work. We would like to acknowledge the Erasmus Mundus programme SE4GD under grant number EMJMD-619839.

Appendix

The table provided on the subsequent pages is the full version that was used to extract the data for the summary provided in Table 3 in Section 4.

The Table 7 provides: (1) an overview of each system by providing a short description; (2) lists the enabling impacts as it was reported in the data source; (3) gives an overview of the potential systemic impacts that the system may have; and (4) the identified factors that may have potentially caused the system to have a negative social impact.

Data availability

The data analysis is provided in the appendix and in a Figshare submission.

References

[1] J. Dastin, Amazon scraps secret AI recruiting tool that showed bias against women, 2018, [Online]. Available: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G. (Accessed 18 Oct 2025).
[2] IEEE code of ethics, 2021, [Online]. Available: https://www.ieee.org/about/corporate/governance/p7-8.html. (Accessed 18 Oct 2025).
[3] ACM code of ethics and professional conduct, 2021, [Online]. Available: https://www.acm.org/binaries/content/assets/about/acm-code-of-ethics-booklet.pdf. (Accessed 18 Oct 2025).
[4] https://oecd.ai/en/ai-principles.
[5] B. Mittelstadt, Ai ethics–too principled to fail?, 2019, arXiv preprint arXiv:1906.06668 https://robotic.legal/wp-content/uploads/2019/05/SSRN-id3391293.pdf.
[6] L. Munn, The uselessness of AI ethics, AI Ethics 3 (3) (2023) 869–877.
[7] T. Hagendorff, The ethics of AI ethics: An evaluation of guidelines, Minds Mach. 30 (1) (2020) 99–120.

[8] S. Makridakis, The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms, Futures 90 (2017) 46–60.

[9] G. Petropoulos, The impact of artificial intelligence on employment, Praise Work. Digit. Age 119 (2018) 119–130.

[10] K. Crawford, R. Dobbe, T. Dryer, G. Fried, B. Green, E. Kaziunas, A. Kak, V. Mathur, E. McElroy, A.N. Sánchez, D. Raji, J.L. Rankin, R. Richardson, J. Schultz, S.M. West, M. Whittaker, AI now 2019 report, AI Now Inst. (2019).

[11] I.Y. Chen, P. Szolovits, M. Ghassemi, Can AI help reduce disparities in general medical and mental health care? AMA J. Ethics 21 (2) (2019) 167–179.

[12] L. Vesnic-Alujevic, S. Nascimento, A. Pólvora, Societal and ethical impacts of artificial intelligence: Critical notes on European policy frameworks, Telecommun. Policy 44 (6) (2020) 101961.

[13] S.M. West, M. Whittaker, K. Crawford, Discriminating systems: Gender, race and power in AIs, AI Now Inst. (2019).

[14] S. Leavy, Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning, in: Proceedings of the 1st International Workshop on Gender Equality in Software Engineering, 2018, pp. 14–16.

[15] M. Whittaker, M. Alper, C.L. Bennett, S. Hendren, L. Kaziunas, M. Mills, M.R. Morris, J. Rankin, E. Rogers, M. Salas, et al., Disability, bias, and AI, AI Now Inst. (2019).

[16] https://artificialintelligenceact.eu/ai-act-explorer/.

[17] https://futureoflife.org/open-letter/ai-principles/.

[18] S. Westerstrand, Fairness in AI systems development: EU AI Act compliance and beyond, Inf. Softw. Technol. (2025) 107864.

[19] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems, [Online]. Available: https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html.

[20] ACM US Public Policy Council, Statement on algorithmic transparency and accountability, Assoc. Comput. Mach. (2017).

[21] J. Whittlestone, R. Nyrup, A. Alexandrova, S. Cave, The role and limits of principles in AI ethics:towards a focus on tensions, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019, pp. 195–200, [Online]. Available: https://dl.acm.org/doi/10.1145/3306618.3314289.

[22] J. Cowls, L. Floridi, Prolegomena to a white paper on an ethical framework for a good AI society, 2018, Available At SSRN 3198732.

[23] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, et al., AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations, Minds Mach. 28 (4) (2018) 689–707.

[24] A.A. Khan, S. Badshah, P. Liang, M. Waseem, B. Khan, A. Ahmad, M. Fahmideh, M. Niazi, M.A. Akbar, Ethics of AI: A systematic literature review of principles and challenges, in: Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering, 2022, pp. 383–392.

[25] V. Garousi, M. Felderer, M.V. Mäntylä, Guidelines for including grey literature and conducting multivocal literature reviews in software engineering, Inf. Softw. Technol. 106 (2019) 101–121.

[26] C. Wohlin, Guidelines for snowballing in systematic literature studies and a replication in software engineering, in: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering, 2014, pp. 1–10.

[27] Grey Literature, Karolinska Institutet, 2021, [Online]. Available: https://kib.ki.se/en/search-evaluate/grey-literature. (Accessed 18 Oct 2025).

[28] D. Moher, A. Liberati, J. Tetzlaff, D.G. Altman, P. Group, et al., Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement, Int. J. Surg. 8 (5) (2010) 336–341.

[29] M. Al Hinai, R. Chitchyan, Social sustainability indicators for software: Initial review, Science 79 (68) (2014) 29.

[30] G. Assefa, B. Frostell, Social sustainability and social acceptance in technology assessment: A case study of energy technologies, Technol. Soc. 29 (1) (2007) 63–78.

[31] United Nations, Universal declaration of human rights, 1948, [Online]. Available: https://www.un.org/en/about-us/universal-declaration-of-human-rights. (Accessed 18 Oct 2025).

[32] L. Hilty, B. Aebischer, ICT for Sustainability: An Emerging Research Field, Vol. 310, 2015, pp. 3–36.

[33] S. Betz, B. Penzenstadler, L. Duboc, R. Chitchyan, S.A. Kocak, I. Brooks, S. Oyedeji, J. Porras, N. Seyff, C.C. Venters, Lessons learned from developing a sustainability awareness framework for software engineering using design science, ACM Trans. Softw. Eng. Methodol. 33 (5) (2024) 1–39.

[34] B. Penzenstadler, L. Duboc, S. Akinli Kocak, C. Becker, S. Betz, R. Chitchyan, S. Easterbrook, O. Leifler, J. Porras, N. Seyff, C.C. Venters, The SusA Workshop - improving sustainability awareness to inform future business process and systems design, 2020, [Online]. Available: https://zenodo.org/record/3676514.

[35] B. Andersen, T. Fagerhaug, Root Cause Analysis, Quality Press, 2006.

[36] Replication package for AI systems' negative social impact and factors, https://doi.org/10.6084/m9.figshare.30234529.

[37] S. Levin, A beauty contest was judged by AI and the robots didn't like dark skin, Guardian (2016) [Online]. Available: https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people. (Accessed 18 Oct 2025).

[38] J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine bias, ProPublica (2016) [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. (Accessed 18 Oct 2025).

[39] A. Engler, For some employment algorithms, disability discrimination by default, 2019, [Online]. Available: https://www.brookings.edu/blog/techtank/2019/10/31/for-some-employment-algorithms-disability-discrimination-by-default/. (Accessed 18 Oct 2025).

[40] N. Vigdor, Apple card investigated after gender discrimination complaints, N. Y. Times (2019) [Online]. Available: https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html. (Accessed 18 Oct 2025).

[41] J. Vincent, Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day, Verge 24 (2016) [Online]. Available: https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist. (Accessed 18 Oct 2025).

[42] Z. Obermeyer, S. Mullainathan, Dissecting racial bias in an algorithm that guides health decisions for 70 million people, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019.

[43] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J.R. Rickford, D. Jurafsky, S. Goel, Racial disparities in automated speech recognition, Proc. Natl. Acad. Sci. 117 (14) (2020) 7684–7689.

[44] G. Aviles, Black and Hispanic students pay more for college loans, study finds, NBC News (2020) [Online]. Available: https://www.nbcnews.com/news/nbcblk/black-hispanic-students-pay-more-college-loans-study-finds-n1132816. (Accessed 18 Oct 2025).

[45] A. Gilbertson, Data-Informed predictive policing was Heralded as less biased. Is it? Markup (2020) [Online]. Available: https://themarkup.org/ask-the-markup/2020/08/20/does-predictive-police-technology-contribute-to-bias. (Accessed 18 Oct 2025).

[46] S. Melendez, Uber Driver Troubles Raise Concerns About Transgender Face Recognition, Fast Company & Inc, 2018, [Online]. Available: https://www.fastcompany.com/90216258/uber-face-recognition-tool-has-locked-out-some-transgender-drivers. (Accessed 18 Oct 2025).

[47] L. Ingham, Racist facial recognition technology is being used by police at anti-racism police at anti-racism protests, Verdict (2020) [Online]. Available: https://www.verdict.co.uk/facial-recognition-technology-racist-police-protests/. (Accessed 18 Oct 2025).

[48] B. Gilbert, Amazon sells facial recognition software to police all over the US, but has no idea how many departments are using it, Bus. Insider (2020) [Online]. Available: https://www.businessinsider.com/amazon-rekognition-police-use-unknown-2020-2. (Accessed 18 Oct 2025).

[49] J. Baily, J. Burkell, V. Steeves, AI technologies — like police facial recognition — discriminate against people of colour, 2020, [Online]. Available: https://theconversation.com/ai-technologies-like-police-facial-recognition-discriminate-against-people-of-colour-143227. (Accessed 18 Oct 2025).

[50] A. Asher-Schapiro, 'Unfair surveillance'? Online exam software sparks global student revolt, Reuters (2020) [Online]. Available: https://www.reuters.com/article/us-global-tech-education-feature-trfn-idUSKBN27Q1Q1. (Accessed 18 Oct 2025).

[51] S. Swauger, Software that monitors students during tests perpetuates inequality and violates their privacy, MIT Technol. Rev. (2020) [Online]. Available: https://www.technologyreview.com/2020/08/07/1006132/software-algorithms-proctoring-online-tests-ai-ethics/. (Accessed 18 Oct 2025).

[52] Z. Schiffer, This girls-only app uses AI to screen a user's gender — what could go wrong? Verge (2020) [Online]. Available: https://www.theverge.com/2020/2/7/21128236/gender-app-giggle-women-ai-screen-trans-social. (Accessed 18 Oct 2025).

[53] N. Kayser-Bril, Google apologizes after its vision AI produced racist results, AlgorithmWatch (2020) [Online]. Available: https://algorithmwatch.org/en/google-vision-racism/. (Accessed 18 Oct 2025).

[54] A. Hern, Cambridge Analytica: how did it turn clicks into votes? Guardian (2018) [Online]. Available: https://www.theguardian.com/news/2018/may/06/cambridge-analytica-how-turn-clicks-into-votes-christopher-wylie. (Accessed 18 Oct 2025).

[55] A. International, Netherlands: we sense trouble: Automated discrimination and mass surveillance in predictive policing in the Netherlands, Amnesty Int. (2020) https://www.amnesty.org/en/documents/eur35/2971/2020/en/.

[56] About Ad Delivery, Facebook for Business, 2021, [Online]. Available: https://bit.ly/3oNq2HI. (Accessed 18 Oct 2025).

[57] Y. Wang, M. Kosinski, Deep neural networks are more accurate than humans at detecting sexual orientation from facial images, J. Pers. Soc. Psychol. 114 (2) (2018) 246.

[58] B. Resnick, This psychologist's "gaydar" research makes us uncomfortable. That's the point, Vox (2018) [Online]. Available: https://www.vox.com/science-and-health/2018/1/29/16571684/michal-kosinski-artificial-intelligence-faces. (Accessed 18 Oct 2025).

[59] G. Lubin, 'Facial-profiling' could be dangerously inaccurate and biased, experts warn, Insider (2016) [Online]. Available: https://www.businessinsider.com/does-faception-work-2016-10?r=US&IR=T. (Accessed 18 Oct 2025).

[60] D. Bissell, T. Birtchnell, A. Elliott, E.L. Hsu, Autonomous automobilities: The social impacts of driverless vehicles, Curr. Sociol. 68 (1) (2020) 116–134.

[61] P. Lin, Why ethics matters for autonomous carsr, in: Autonomous Driving, Springer, Berlin, Heidelberg, 2016, pp. 69–85.

[62] A. Osório, A. Pinto, Information, uncertainty and the manipulability of artificial intelligence autonomous vehicles systems, Int. J. Hum.-Comput. Stud. 130 (2019) 40–46.

[63] B. Chesney, D. Citron, Deep fakes: A looming challenge for privacy, democracy, and national security, Calif. L. Rev. 107 (2019) 1753.

[64] C. Silverman, How to spot a deepfake like The Barack Obama–Jordan peele video, BuzzFeed (2018) [Online]. Available: https://www.buzzfeed.com/craigsilverman/obama-jordan-peele-deepfake-video-debunk-buzzfeed. (Accessed 18 Oct 2025).

[65] A. Kaplan, M. Haenlein, Siri, siri, in my hand: Who's the fairest in the land? on the interpretations, illustrations, and implications of artificial intelligence, Bus. Horiz. 62 (1) (2019) 15–25.

[66] A. Kharpal, China's Huawei tested A.I. software that could identify Uighur Muslims and alert police, report says, CNBC (2020) [Online]. Available: https://www.cnbc.com/2020/12/09/chinas-huawei-tested-ai-software-that-could-identify-uighurs-report.html. (Accessed 18 Oct 2025).

[67] Lens Studio, Lens Studio - Lens Studio by Snap Inc., 2021, [Online]. Available: https://lensstudio.snapchat.com/. (Accessed 18 Oct 2025).

[68] Introducing face filters & more on instagram, 2017, [Online]. Available: https://about.instagram.com/blog/announcements/introducing-face-filters-and-more-on-instagram. (Accessed 18 Oct 2025).

[69] O. Harvey, Did Facebook warn you that a friend used the "this is your digital life" app? Here's what that means, HelloGiggles (2018) [Online]. Available: https://hellogiggles.com/news/facebook-this-is-your-digital-life-app/. (Accessed 18 Oct 2025).

[70] Map of countries that criminalise LGBT people, Hum. Dignity Trust. [Online]. Available: https://bit.ly/2SkMM5G. (Accessed 18 Oct 2025).

[71] N. Kobie, The complicated truth about China's social credit system, Wired (2019) [Online]. Available: https://www.wired.co.uk/article/china-social-credit-system-explained. (Accessed 18 Oct 2025).

[72] Secret documents reveal how China mass detention camps work, CNBC (2019) [Online]. Available: https://www.cnbc.com/2019/11/25/secret-documents-reveal-how-china-mass-detention-camps-work.html. (Accessed 18 Oct 2025).

[73] Data leak reveals how China 'brainwashes' uighurs in prison camps, BBC (2019) [Online]. Available: https://www.bbc.com/news/world-asia-china-50511063. (Accessed 18 Oct 2025).

[74] "Break their lineage, break their roots" China's crimes against humanity targeting uyghurs and other turkic muslims, Hum. Rights Watch. (2021) [Online]. Available: https://www.hrw.org/report/2021/04/19/break-their-lineage-break-their-roots/chinas-crimes-against-humanity-targeting#. (Accessed 18 Oct 2025).

[75] K. Hao, Deepfake porn is ruining women's lives. Now the law may finally ban it, MIT Technol. Rev. (2021) [Online]. Available: https://www.technologyreview.com/2021/02/12/1018222/deepfake-revenge-porn-coming-ban/. (Accessed 18 Oct 2025).

[76] Federal Trade Commission, FTC sues cambridge analytica, settles with former CEO and app developer, 2019, [Online]. Available: https://www.ftc.gov/news-events/press-releases/2019/07/ftc-sues-cambridge-analytica-settles-former-ceo-app-developer. (Accessed 18 Oct 2025).

[77] K. Lum, W. Isaac, To predict and serve? Significance 13 (5) (2016) 14–19.

[78] S. Adee, Controversial software claims to tell personality from your face, New-Scientist (2016) [Online]. Available: https://www.newscientist.com/article/2090656-controversial-software-claims-to-tell-personality-from-your-face/. (Accessed 18 Oct 2025).

[79] T. Chappellet-Lanier, Study finds biases in Amazon Rekognition's facial analysis tool, FedScoop (2019) [Online]. Available: https://www.fedscoop.com/study-finds-biases-amazon-rekognitions-facial-analysis-tool/. (Accessed 18 Oct 2025).

[80] A. Brackey, Analysis of Racial Bias in Northpointe's COMPAS Algorithm (Ph.D. thesis), Tulane University School of Science and Engineering, 2019.

[81] D. Isetti, Disclosure of a communication disorder during a job interview: A theoretical model, J. Commun. Disord. 87 (2020) 106038.

[82] R. Engeln-Maddox, Buying a beauty standard or dreaming of a new life? Expectations associated with media ideals, Psychol. Women Q. 30 (3) (2006) 258–266.

[83] D.L. MacKenzie, L. Goodstein, Long-term incarceration impacts and characteristics of long-term offenders: An empirical analysis, Crim. Justice Behav. 12 (4) (1985) 395–414.

[84] A. Society, The importance of work for individuals with intellectual/developmental disabilities, 2021, [Online]. Available: https://www.autism-society.org/wp-content/uploads/2018/04/IDD-BRIEFING-Employment-importance-Final-2.22.18.pdf. (Accessed 18 Oct 2025).

[85] R.A. Hopkinson, E. Keatley, E. Glaeser, L. Erickson-Schroth, O. Fattal, M. Nicholson Sullivan, Persecution experiences and mental health of LGBT asylum seekers, J. Homosex. 64 (12) (2017) 1650–1666.

[86] E.H. Baker, Socioeconomic status, definition, Wiley Blackwell Encycl. Health Illn. Behav. Soc. (2014) 2210–2214.

[87] J. Grinevičius, A. Akavickaitė, People tested how Google translates from gender neutral languages and shared the "Sexist" results, Bored Panda (2021) [Online]. Available: https://cutt.ly/yb57Hte. (Accessed 18 Oct 2025).

[88] M. Hosoda, L.T. Nguyen, E.F. Stone-Romero, The effect of Hispanic accents on employment decisions, J. Manag. Psychol. (2012).

[89] N.E. Moss, Gender equity and socioeconomic inequality: a framework for the patterning of women's health, Soc. Sci. Med. 54 (5) (2002) 649–661.

[90] S. samuel, China's Jaw-Dropping family separation policy, Atl. (2018) [Online]. Available: https://www.theatlantic.com/international/archive/2018/09/china-internment-camps-uighur-muslim-children/569062/. (Accessed 18 Oct 2025).

[91] S.G. Arai, Intergenerational Experience of Japanese American Internment: The Grandchildren of the Camps (Ph.D. thesis), John F. Kennedy University, 2012.

[92] B. Imana, A. Korolova, J. Heidemann, Auditing for discrimination in algorithms delivering job ads, 2021, arXiv preprint arXiv:2104.04502.

[93] D.M. Lazer, M.A. Baum, Y. Benkler, A.J. Berinsky, K.M. Greenhill, F. Menczer, M.J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, et al., The science of fake news, Science 359 (6380) (2018) 1094–1096.

[94] S.C. Woolley, P.N. Howard, Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media, Oxford University Press, 2018, p. 4.

[95] C. Vaccari, A. Chadwick, Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news, Soc. Media+ Soc. 6 (1) (2020).

[96] C. O'neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, Crown, 2016.

[97] W. Knight, The apple card didn't 'see' gender - and that's the problem, Wired (2019) [Online]. Available: https://www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/. (Accessed 18 Oct 2025).

[98] C.F. Manski, D.S. Nagin, Assessing benefits, costs, and disparate racial impacts of confrontational proactive policing, Proc. Natl. Acad. Sci. 114 (35) (2017) 9308–9313.

[99] S.K. Brooks, N. Greenberg, Psychological impact of being wrongfully accused of criminal offences: A systematic literature review, Med. Sci. Law (2020).

[100] P. Mason, The racist hijacking of microsoft's chatbot shows how the internet teems with hate, Guardian (2016) [Online]. Available: https://www.theguardian.com/world/2016/mar/29/microsoft-tay-tweets-antisemitic-racism. (Accessed 18 Oct 2025).

[101] A. Milevsky, M. Schlechter, S. Netter, D. Keehn, Maternal and paternal parenting styles in adolescents: Associations with self-esteem, depression and life-satisfaction, J. Child Fam. Stud. 16 (1) (2007) 39–47.

[102] X. Zhang, X. Xuan, F. Chen, C. Zhang, Y. Luo, Y. Wang, The relationship among school safety, school liking, and students' self-esteem: Based on a multilevel mediation model, J. Sch. Health 86 (3) (2016) 164–172.

[103] M. Jan, S. Soomro, N. Ahmad, Impact of social media on self-esteem, Eur. Sci. J. 13 (23) (2017) 329–341.

[104] R.D. Conger, K.J. Conger, M.J. Martin, Socioeconomic status, family processes, and individual development, J. Marriage Fam. 72 (3) (2010) 685–704.

[105] R. Dresser, Wanted single, white male for medical research, Hast. Cent. Rep. 22 (1) (1992) 24–29.

[106] A. Kak, REGULATING BIOMETRICS global approaches and urgent questions, AI Now Inst. (2020) [Online]. Available: https://ainowinstitute.org/regulatingbiometrics.pdf. (Accessed 18 Oct 2025).

[107] S. Corbett-Davies, S. Goel, The measure and mismeasure of fairness: A critical review of fair machine learning, 2018, arXiv preprint arXiv:1808.00023.

[108] P.J. Phillips, F. Jiang, A. Narvekar, J. Ayyad, A.J. O'Toole, An other-race effect for face recognition algorithms, ACM Trans. Appl. Percept. (TAP) 8 (2) (2011) 1–11.

[109] D. Leslie, Understanding bias in facial recognition technologies, 2020, arXiv preprint arXiv:2010.07023.

[110] R. Bivens, The gender binary will not be deprogrammed: Ten years of coding gender on Facebook, New Media Soc. 19 (6) (2017) 880–898.

[111] C. Becker, Insolvent: How to Reorient Computing for Just Sustainability, MIT Press, 2023.

[112] D. Giustini, M.N.K. Boulos, Google Scholar is not enough to be used alone for systematic reviews, Online J. Public Health Inform. 5 (2) (2013) 214, [Online]. Available: https://doi.org/10.5210/ojphi.v5i2.4623.

[113] E. Vanhala, J. Kasurinen, A. Knutas, A. Herala, The application domains of systematic mapping studies: A mapping study of the first decade of practice with the method, IEEE Access 10 (2022) 37924–37937.

[114] M. Castañeda, The power of (mis)representation: Why racial and ethnic stereotypes in the media matter, Challenging Inequal.: Read. Race Ethn. Immigr. (2018).

[115] K. Philbrick, Imprisonment: The impact on children, Issues Forensic Psychol. (2002).

[116] M. Strandh, A. Winefield, K. Nilsson, A. Hammarström, Unemployment and mental health scarring during the life course, Eur. J. Public Health 24 (3) (2014) 440–445.

[117] U.S. Equal Employment Opportunity Commission, Prohibited employment policies/practices, 2021, [Online]. Available: https://www.eeoc.gov/prohibited-employment-policiespractices. (Accessed 18 Oct 2025).
[118] Persons with a disability: Labor force characteristics summary, 2021, [Online]. Available: https://www.bls.gov/news.release/disabl.nr0.htm. (Accessed 18 Oct 2025).
[119] A. Lee, An AI to stop hiring bias could be bad news for disabled people, WIRED UK (2019) [Online]. Available: https://www.wired.co.uk/article/ai-hiring-bias-disabled-people. (Accessed 18 Oct 2025).
[120] P. Tharenou, The work of feminists is not yet done: The gender pay gap—a stubborn anachronism, Sex Roles 68 (3–4) (2013) 198–206.
[121] G. Neubaum, N.C. Krämer, Monitoring the opinion of the crowd: Psychological mechanisms underlying public opinion perceptions on social media, Media Psychol. 20 (3) (2017) 502–531.
[122] L.E. Egede, Race, ethnicity, culture, and disparities in health care, J. Gen. Intern. Med. 21 (6) (2006) 667.
[123] S. de Lemus, L. Estevan-Reina, Influence of sexist language on motivation and feelings of ostracism (La influencia del lenguaje sexista en la motivación y el sentimiento de ostracismo), Int. J. Soc. Psychol. (2021) 1–37.
[124] R.R. Borah, M.M.H.G. Bhuvaneswari, Keeping sexism alive through social acceptability: A contextual study of sexist and derogatory slurs, 2020.
[125] M. Menegatti, M. Rubini, Gender bias and sexism in language, in: Oxford Research Encyclopedia of Communication, 2017.
[126] J. Palmiter Bajorek, Voice recognition still has significant race and gender biases, Harv. Bus. Rev. (2019) [Online]. Available: https://hbr.org/2019/05/voice-recognition-still-has-significant-race-and-gender-biases. (Accessed 18 Oct 2025).
[127] Computer says no: Irish vet fails oral English test needed to stay in Australia, Guardian (2017) [Online]. Available: https://www.theguardian.com/australia-news/2017/aug/08/computer-says-no-irish-vet-fails-oral-english-test-needed-to-stay-in-australia. (Accessed 18 Oct 2025).
[128] Algorithms and bias: What lenders need to know, White Case LLP (2017) [Online]. Available: https://www.whitecase.com/publications/insight/algorithms-and-bias-what-lenders-need-know. (Accessed 18 Oct 2025).
[129] C. Callender, J. Jackson, Does the fear of debt deter students from higher education? J. Soc. Policy 34 (4) (2005) 509–540.
[130] R.B. Taylor, B.R. Wyant, B. Lockwood, Variable links within perceived police legitimacy?: Fairness and effectiveness across races and places, Soc. Sci. Res. 49 (2015) 234–248.
[131] R. Weitzer, S.A. Tuch, Determinants of public satisfaction with the police, Police Q. 8 (3) (2005) 279–297.
[132] D. Raphael, Social determinants of health: Canadian perspectives, 2009, pp. 252–253.
[133] O. Keyes, The misgendering machines: Trans/HCI implications of automatic gender recognition, Proc. ACM Hum.-Comput. Interact. 2 (CSCW) (2018) 1–22.
[134] D. Brown, Student sues apple for $1 billion, claims face-recognition caused false arrest, USA Today (2019) [Online]. Available: https://eu.usatoday.com/story/tech/2019/04/23/apple-lawsuit-teen-claims-facial-recognition-tech-caused-false-arrest/3547479002/. (Accessed 18 Oct 2025).
[135] J.C. Fox, Brown University student mistakenly identified as Sri Lanka bombing suspect, Boston Globe (2019) [Online]. Available: https://bit.ly/3fNG402. (Accessed 18 Oct 2025).
[136] Big brother watch briefing on algorithmic Decision-Making in the criminal justice system, Big Brother. Watch. (2020) [Online]. Available: https://bigbrotherwatch.org.uk/wp-content/uploads/2020/02/Big-Brother-Watch-Briefing-on-Algorithmic-Decision-Making-in-the-Criminal-Justice-System-February-2020.pdf. (Accessed 18 Oct 2025).
[137] L. Elsalem, N. Al-Azzam, A.A. Jum'ah, N. Obeidat, A.M. Sindiani, K.A. Kheirallah, Stress and behavioral changes with remote E-exams during the Covid-19 pandemic: A cross-sectional study among undergraduates of medical sciences, Ann. Med. Surg. 60 (2020) 271–279.
[138] What is Gender? What is Sex?, Canadian Institutes of Health Research, 2020, [Online]. Available: https://cihr-irsc.gc.ca/e/48642.html. (Accessed 18 Oct 2025).
[139] J. Vincent, Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech, Verge (2018) [Online]. Available: https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai. (Accessed 18 Oct 2025).
[140] V. Hunt, D. Layton, S. Prince, Diversity Matters, McKinsey & Company, 2015, [Online]. Available: https://www.insurance.ca.gov/diversity/41-ISDGBD/GBDExternal/upload/McKinseyDivmatters-201501.pdf. (Accessed 18 Oct 2025).
[141] S. Biddle, Research says Facebook's ad algorithm perpetuates gender bias, Intercept. (2021) [Online]. Available: https://theintercept.com/2021/04/09/facebook-algorithm-gender-discrimination/. (Accessed 18 Oct 2025).
[142] A.R. Flores, Social acceptance of LGBT people in 174 countries: 1981 to 2017, 2019.
[143] A.P. Gregg, B. Seibt, M.R. Banaji, Easier done than undone: asymmetry in the malleability of implicit preferences, J. Pers. Soc. Psychol. 90 (1) (2006) 1.
[144] K.A. Pantserev, The malicious use of AI-based deepfake technology as the new threat to psychological security and political stability, in: Cyber Defence in the Age of AI, Smart Societies and Augmented Humanity, Springer, 2020, pp. 37–55.
[145] B.N. Service, Who are the uyghurs and why is China being accused of genocide? BBC (2022) https://www.bbc.com/news/world-asia-china-22278037.
[146] N. Serle, Is Snapchat changing beauty standards? Croft (2018) [Online]. Available: https://epigram.org.uk/2018/04/03/snapchat-beauty-ideals/. (Accessed 18 Oct 2025).
[147] YMCA, Be real campaign, YMCA (2018) [Online]. Available: https://www.ymca.org.uk/about/what-we-do/be-real-campaign. (Accessed 18 Oct 2025).
[148] M. Miller, Research Looks at How Snapchat Filters Affect Self-Image, University of Wisconsin-Madison, 2019, [Online]. Available: https://news.wisc.edu/research-looks-at-how-snapchat-filters-affect-self-image/. (Accessed 18 Oct 2025).
[149] S.C. Tremblay, S.E. Tremblay, P. Poirier, From filters to fillers: an active inference approach to body image distortion in the selfie era, AI Soc. 36 (1) (2021) 33–48.
[150] K. Hawker, N. Carah, Snapchat's augmented reality brand culture: sponsored filters and lenses as digital piecework, Continuum (2020) 1–18.