A sector-based approach to AI ethics: Understanding ethical issues of AI-related incidents within their sectoral context

Dafna Burema
Institute of Sociology, Technische Universität Berlin, Germany; Science of Intelligence, Research Cluster of Excellence, Germany

Nicole Debowski-Weimann
Volkswagen Group, Germany

Alexander Von JANOWSKI*
Responsible Technology Hub, Germany

Jil Grabowski
Volkswagen Consulting, Volkswagen Group, Germany

Mihai Maftei
German Research Center for Artificial Intelligence, Germany

Mattis Jacobs
Institute of Sociology, Technische Universität Berlin, Germany; Science of Intelligence, Research Cluster of Excellence, Germany

Patrick Van Der Smagt
Machine Learning Research Lab, Volkswagen Group, Germany; Department of Computer Science, ELTE University Budapest, Hungary

Djalel Benbouzid
Machine Learning Research Lab, Volkswagen Group, Germany

ABSTRACT

Acknowledging that society is made up of different sectors with their own rules and structures, this paper studies the relevance of a sector-specific perspective to AI ethics. Incidents with AI are studied in relation to five sectors (police, healthcare, education and academia, politics, automotive) using the AIAAIC repository. A total of 125 incidents are sampled and analyzed by conducting a qualitative content analysis on media reports. The results show that certain ethical principles are found breached across sectors: accuracy/reliability, bias/discrimination, transparency, surveillance/privacy, security. However, results also show that 1) some ethical issues (misinformation, safety, premise/intent) are sector specific, 2) the consequences and meaning of the same ethical issue is able to vary across sectors and 3) pre-existing sector-specific issues are reproduced with these ethical breaches. The paper concludes that general ethical principles are relevant to discuss across sectors, yet, a sector-based approach to AI ethics gives in-depth information on sector-specific structural issues.

CCS CONCEPTS

• **Social and professional topics**; • **Computing methodologies**; • **Artificial intelligence**; • **Philosophical/theoretical foundations of artificial intelligence.**;

KEYWORDS

Artificial Intelligence, AI deployment, AI Ethics, AI incidents, Sectors, Media Reports

ACM Reference Format:
Dafna Burema, Nicole Debowski-Weimann, Alexander Von JANOWSKI, Jil Grabowski, Mihai Maftei, Mattis Jacobs, Patrick Van Der Smagt, and Djalel Benbouzid. 2023. A sector-based approach to AI ethics: Understanding ethical issues of AI-related incidents within their sectoral context. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES '23), August 08–10, 2023, Montréal, QC, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3600211.3604680

1 INTRODUCTION

Artificial Intelligence (AI) has developed as a tool to improve efficiency, reduce costs, and enable new activities in various contexts with pilots and applications in, for example, fraud detection [6], hiring [17], and law enforcement [56]. At the same time, it is widely recognized that the deployment of AI is not without risks. In the past years, AI-related controversies arose across a variety of cases revealing issues ranging from surveillance, to biases and discrimination, and causing harm due to problems with the reliability and security of such systems [14, 30]. The need to account for these ethical issues has been widely acknowledged. With a "turn to ethics" [59:2], actors from industry (e.g., [35, 51]), the public sector (e.g., [5, 40, 54]), and non-governmental organizations (e.g., [2, 18]) have outlined principles to ensure the ethical, responsible and trustworthy development and deployment of artificial intelligence in AI ethics guidelines.

As important as such initiatives are for raising awareness for ethical issues of AI, they have been criticized as being too abstract [39, 57] and offering little to no practical applicability [66, 73, 75]. Furthermore, evaluations of AI ethics guidelines showed them to be too generic [63], vague [60], and hosting a multitude of possible interpretations [62], leading to a lack of clarity regarding how AI

*Work conducted during an internship at the Machine Learning Research Lab of Volkswagen Group.

AIES '23, August 08–10, 2023, Montréal, QC, Canada

ACM ISBN 979-8-4007-0231-0/23/08.
https://doi.org/10.1145/3600211.3604680

principles should be implemented, interpreted, or prioritized [12]. Based on such critique, some scholars question AI ethics guidelines in principle [52]. However, one possibility to close the "wide and thorny gap between the articulation of these high-level concepts and their actual achievement in the real world" [31:66] is to make AI ethics guidelines less abstract and ambiguous. To make abstract concepts such as ethical principles and values sufficiently concrete, they need to be viewed within a specific context.

This paper explores one approach to make ethics guidelines more tailored towards social context: focusing on sectors and their specific characteristics. A sector-based perspective to AI ethics enables understanding how AI systems are embedded in specific sectoral cultures with e.g. their norms, structures, activities, and routines, which is a perspective that is thus far overlooked in the AI ethics community. To elaborate, sectors are not explicitly used as a conceptual tool, but rather implicitly treated as relevant in relation to AI and robo-ethics in case studies such as elder care [15] or education [64]. To address this gap in the literature, this present paper aims to understand the feasibility of a sector-based approach to AI ethics. Are certain ethical issues found in specific sectors, or are ethical principles breached across sectors? In other words, it will be studied whether it has merit being sensitive to contextual, sector-specific information when understanding AI ethics, or whether overarching and rather general values are sufficient in doing so. Bearing this in mind, the guiding research question reads: How is sectoral context related to breaches of ethical principles?

In order to answer this research question, breaches of ethical principles are operationalized in terms of incidents with AI after deployment, as is listed in media reports in the AI, Algorithmic and Automation Incidents and Controversies (hereafter AIAAIC) repository [1]. Five sectors are selected for an empirical analysis on incidents with AI: healthcare, education and academia, police, politics, automotive. By comparing these sectors and their AI-related incidents, it could be seen whether such incidents occur in isolation (i.e. only within their respective sector), or across sectors. What follows next is an overview of related work in AI ethics with a focus on its principles and guidelines.

2 RELATED WORK

In order to situate this current study in literature, related work that addresses theoretical questions concerning principles and guidelines for ethical AI is discussed, followed by studies that also focus on the sectoral context of AI ethics.

2.1 Principles and guidelines for ethical AI

The widespread adoption of AI technologies is increasingly accompanied by calls for mitigating the risks that AI technologies pose. As a response, a variety of societal actors such as governments, policymakers and international organizations, businesses, professional associations, advocacy groups, and multi-stakeholder initiatives have produced ethical guidelines with the goal of defining and creating AI in accordance with ethical values and principles. Despite the multitude of guidelines coming from different institutional backgrounds, some overlap among the principles can be observed. According to Jobin et al. [43], eleven overarching ethical values and principles are found when comparing eighty-four AI ethical guidelines. These are, by frequency of the number of sources in which they were featured: transparency, justice and fairness, non-maleficence, responsibility, privacy, beneficence, freedom and autonomy, trust, dignity, sustainability, and solidarity. Another paper [31] states that eight main principles were found after analyzing thirty-six ethical guidelines: privacy, accountability, safety and security, transparency and explainability, fairness and non-discrimination, human control of technology, professional responsibility, and promotion of human values. When comparing these two papers and their results, partial overlap can be observed in the content of these ethical guidelines, e.g., transparency, privacy, fairness.

However, as Jobin et al. [43] note, relying on a numerical assessment of mentioned ethical values and principles, i.e., assessing which values and principles are mentioned how often, obfuscates divergences regarding "(1) how ethical principles are interpreted; (2) why they are deemed important; (3) what issue, domain or actors they pertain to; and (4) how they should be implemented". Thus, the landscape of AI ethics guidelines still is marked by extensive heterogeneity and is far away from "a unified framework that can guide the governance of AI" [60:11]. This raises questions about the applicability of AI ethics guidelines more generally, as it is difficult for AI practitioners to determine which ethical issues they may run into [61] and how they should interpret, account for, and operationalize proposed ethical values and principles [31, 43]. This challenge has also been investigated empirically. A behavioral ethics study on the effects of the ACM ethical guidelines [50] shows that the availability of the guidelines alone has no statistically relevant influence on ethical decision making and concludes that future research needs to find different ways that can influence ethical decision making. Vakkuri et al. [73] conclude in their study that the academic discussion around ethical values has been too conceptual and as a result, does not seem to have influenced the industry at large yet.

In short, there are still things left unclear within the AI ethics community. The rather broad character of AI ethics typically does not account for social complexities and the situated realities of ethical breaches. In the analysis, this is taken into account, as real world incidents are examined in relation to such ethical values and principles. In doing so, this study aims to understand how applicable general AI ethics principles are in different sectors.

2.2 Sectoral context and ethical AI

To account for the ways the social environment shapes both the development as well as the post-deployment phase of AI, researchers have called for broadening the analytical lens [3, 23, 24]. As discussed in the introduction, this could be achieved by introducing a sector-based approach, allowing to account for sector-specific characteristics. The field of AI ethics does have numerous case studies that ontologically assume the relevance of understanding sectoral context in relation to AI ethics. For instance, in her analysis on novel elder care technologies, Burema [15] argues that such technologies embed a neoliberal understanding of the welfare state. In other words, the (un)ethical nature of such technologies was assessed in the context of the sector: aging, and the welfare system.

Thereby, the author does not isolate the technology from its sectoral environment.

One of the few publications that does *explicitly* refer to sectors in understanding AI ethics is the European Commission's High-Level Expert Group on Artificial Intelligence [25]. The authors argue that the AI ethics recommendations the EU has made thus far are too general in their nature, and in-depth knowledge is needed for specific sectors. In their paper, they choose three sectors to make specific recommendations for the creation and deployment of AI in relation to three sectors: health care, the public sector (e-governance and justice/law enforcement), manufacturing, and (industrial) Internet of Things (IoT) sector. Though these authors thereby explicitly acknowledge the relevance of being sensitive to sector-specific contexts of AI ethics, the work reads as three different case studies on three different sectors in which the content of these recommendations was made based on workshops with experts from the respective fields, i.e. without data about deployment.

In contrast, this study does two things differently: 1) instead of solely describing ethical issues for each individual sector, this paper compares the ethical issues of sectors to see how sector-specific mechanisms are (not) relevant when discussing AI ethics, 2) instead of relying on expert interviews, this paper analyzes incidents in particular sectors after deployment, which provides the opportunity for analyzing AI systems and their use "in the wild". Furthermore, by looking at incidents post-deployment, this study takes a broad definition when defining a sector compared with the approach of the European Commission [25]: it does not only concern industry actors situated in a sector, but also users. A sector, therefore in this paper, functions as a broader frame of reference where the incidents took place.

3 METHOD

This study sampled incidents based on media reports shown in the AIAAIC database developed by Pownall [1]. At the time of writing this paper, this database covers more than 950 entries of incidents and includes several variables, such as sector, country, year, and URL links of media reports. For this study, two variables are of interest: sector and URL links that relate the incident to the media report. The content of these media reports is analyzed qualitatively with a thematic analysis according to sector as is explained later. First, the selection procedure for the sectors is explained, as there were many to choose from in the database.

This paper does not follow a predefined operationalization of sectors, where certain sectors are chosen over others before sampling. Rather, sectors are selected based on feasibility and sample size (i.e. the number of incidents per sector in the database): it needs to be feasible to code the data in a limited amount of time while keeping a sufficient number of cases. For that reason, the biggest sector, "technology", is omitted for this study because the database shows more than 220 incidents and is therefore not feasible to code qualitatively in a restricted amount of time, as well as the smaller sectors such as "religion" which shows one incident with a robot priest and is therefore too small in sample size to draw any conclusions. This sampling procedure resulted in the selection of five sectors: police, education and academia, politics, automotive, and healthcare. Initially, taking all the sectors together, a total of 180 cases were identified. After the data was cleaned by two analysts, the sample size was reduced to n=125 cases: police (n=39), education/academia (n=34), politics (n=16), automotive (n=21) and healthcare (n=15). The exclusion criteria applied are: duplicates, not accessible media reports (e.g., paywall), cases that do not relate to AI technologies per se, technologies that are not deployed yet, or media reports that do not discuss an incident (e.g., commentary texts that expressed an author's opinion about an incident or an entire field). Furthermore, cases that were labeled incorrectly according to sector, were moved to the respective sector. In total, 55 cases were excluded due to these reasons, resulting in a sample size of n=125. It should be noted that the database was retrieved in February 2021. Since then, more cases were added to the database and it has been reorganized.

The content of the media reports is analyzed qualitatively with a thematic analysis. In essence, this is a tool for data reduction by first exploring the data, then establishing initial codes, and finally establishing themes by comparing and contrasting codes. The unit of analysis is the incident itself, not the media report. In other words, the analysis is not conducted on a semantic level (e.g., framing analysis or discourse analysis) but rather on a descriptive level (i.e., understanding the critical elements of the AI incidents by directly assigning a descriptive code). To elaborate on this process, first the media reports are read in order to understand the nature of the incidents. Then, the media report is coded in terms of the ethical issue that is described in the report (i.e. breach of ethical principle). Since these reports typically deal with a case that has multiple issues, one media report is able to include more than one ethical code. Additionally, special attention is paid to sector-specific activities: where exactly did the incident in the sector take place?

Thus, two pieces of information are analyzed and coded from the media reports: the **ethical issues** (i.e. what ethical principle has been breached?) and sector-specific **activities** (i.e. where is this incident situated within the sector?). Concerning the former, it should be noted that the database already coded each incident according to the respective ethical issue (e.g., accuracy/reliability, transparency, etc.). However, all incidents are re-coded with the purpose of this study in mind, albeit sometimes with the same terminology. The reason for reassessing each incident in terms of their ethical issue is because certain incidents were initially coded in ways that were not aligning with this research's aim. For instance, codes such as "marketing" and "ethics" were found to describe certain incidents. Still, bad marketing is not inherently an AI-related ethical incident, and using "ethics" as a label to describe unethical AI deployment is too generic.

After getting to know the data and developing initial codes, the rest of the coding process involves steps in data reduction: how are these initial codes related to one another (i.e., is there overlap found?), and are there themes able to be established? This is an iterative process, especially for the data and codes that describe sectoral activity. This coding procedure resulted in a couple of themes that describe the ethical issue (i.e. what ethical principle has been breached?) as well as themes that describe their sectoral context in terms of activities (i.e. where is this incident situated within the sector?), as is discussed in the results. The final step is to compare the results across sectors: are certain themes only occurring in particular sectors, or is there overlap found? Can we

speak about general ethics or should AI ethics be tailored towards sectoral contexts?

4 RESULTS

The results are presented in two parts: first a description of the incidents per sector are described in detail. Here, the core ethical principles that are breached (e.g. transparency) are described as well as the sectoral activities (e.g. tracking, monitoring and identifying people in the police sector). Then, two tables are presented in which the sectors and ethical issues are compared.

Before discussing the results in-depth, it should be noted that the data used for this study are media reports. Therefore, the list of incidents are not exhaustive due to media bias, as some topics might be picked up more than others in favor of media logic. Thereby, not all incidents that occurred after deployment and their ethical issues in their respective sector are reflected in the results. Also, it means that the incidents were not observed first-hand, but are filtered through observations of the reporter and its editorial process. This issue of relying on media reports for the analysis is further discussed in the limitations section of this paper.

4.1 Description per sector

4.1.1 Police. When AI is deployed in the police sector, it concerns issues related to tracking, monitoring, or identifying people. Often but not always, this is done with the help of personal data. AI technologies can be used in both ongoing police investigations and predictive policing. What all cases (n=39) in the database have in common is the use of AI for either visual detection of objects or people, or administrative purposes. When AI gets used in this sector, ethical themes relate to accuracy/reliability, bias and discrimination, transparency, surveillance and privacy, as is explained next.

Accuracy/reliability relates to cases that misidentify people, sometimes leading to wrongful arrests [44]. This ties in with another theme: bias and discrimination, as certain racial minorities are often misidentified as also the case of [44] shows. The issue of transparency relates to not knowing when personal data is being used, and for what purpose. For instance, Biddle [11] discusses how the Los Angeles police department requested home security videos of Amazon Ring users to identify protesters in the Black Lives Matter protests. Though the author hints to the possibility of using video footage for facial recognition, and calls surveillance through Ring a "ubiquitous camera network" there is much unclarity about the use of data: "Policies guiding how long cops can retain privately obtained data like Ring videos—and what they can do once it lands on their hard drives—are rare and typically weak". This latter example also ties in with privacy/surveillance issues: as new technologies were primarily used by the police to track, monitor, or identify people, it by default taps into issues of privacy and surveillance. The use of personal data to observe citizens is for instance found in China, where illegal street crossings are being detected with facial recognition software at intersections. After being detected, pictures of supposed offenders are publicly displayed at those intersections on LED screens, and a fine is announced via text message to the offenders [70]. In other words, law enforcement is able to observe its citizens closely with AI, in this case leading to public shaming and fining.

4.1.2 Education and academia. **Education -** The incidents related to this sector (n=25) concern teaching and administrative activities that can be divided into three types: 1) evaluation and grading 2) monitoring and tracking behavior of students; 3) physical and digital access. These three types of activities show a mix of different ethical issues, as is explained next.

Issues with grading show problems with accuracy/reliability and bias/discrimination. Meaning, the systems were not doing the tasks that they were supposed to do but also affect certain sociodemographic groups differently than others. To elaborate, the algorithms used are not accurate or reliable, for instance, when grading tests [19] or predicting students' grades that otherwise could not be performed due to Covid 19 [26]. Bias and discrimination were found when the technologies disadvantage certain groups over others, typically (and at the intersection of) gender and race, such as AI that predicts student success [28], or assesses PhD applications [58].

Tracking and monitoring the behavior of students predominantly breaches principles of privacy and surveillance, but also bias/discrimination and security. Concerning the latter, cybersecurity breaches were found in, for instance, online learning environments and proctoring software [46] though this does not inherently have to do with AI per se but rather could be seen as a side-effect when AI gets implemented. Examples of privacy and surveillance breaches are proctoring software used to administer tests [29], or facial recognition used in Australian schools to check attendance [7]. Bias and discrimination occurs when for instance proctoring software does not identify students with dark skin tones [20].

Access refers to physical access to school and its environment or access to digital learning environments of schools on the basis of biometric data. The incidents related to restricting access due to misidentification. In doing so, the systems are biased/discriminatory or inaccurate. For instance, in the Lockport city school district in the US, media reports mention how the system disproportionally misidentifies black students [27]. Furthermore, there are privacy issues as biometric data are stored, processed, and shared to regulate access [49].

Academia - In academic publications (n=9), the ethical issues concern ethically disputable premises of hypotheses and underlying arguments used to test and create AI. In other words, when publishing, scholars have to specify what ideas they are testing or developing and why, in this case all publications develop an AI or a component thereof. The ethical issues of these incidents do not refer to the output (i.e. how well the AI is performing), rather, the very starting point of the academic publication: the initial ideas that lead to the development of a newly developed AI system. Examples are a publication that developed AI to detect people's sexual orientation with facial recognition [47], or similarly a publication that uses facial recognition to predict political orientation [79].

4.1.3 Politics. The sector "politics" in the dataset refers to the communication of political viewpoints in which deepfakes and twitterbots are created by citizens and political organizations alike to credit and discredit political figures and/or their agendas (n=16). The incidents that occurred in this sector concern ethical issues with misinformation and transparency by communicating messages without disclosing that AI was involved in the construction of the

messages. To clarify, 15 incidents (out of 16 cases) concerned the communication of a message by a deepfake of a politician. Without a disclaimer that such technologies were involved when creating the message, this can be misleading about the authenticity of the message. The content of these deep faked messages ranges from creating fake political statements from politicians [71], to videos used in political election campaigns [53] and advertisements by lobby groups [80]. Only one case was found that did not directly involve audiovisual deepfakes: Twitter bots that disseminated misinformation about climate change [9]. Nonetheless, what all cases have in common, regardless of the exact technology used, is that the incidents concern the communication of political ideas with AI to the general public.

4.1.4 Healthcare. In healthcare (n=15), the activities where AI-related incidents were found concern care provision and medical analyses (i.e. prevention/prognosis/diagnosis), data management (i.e. storing/sharing/tracking medical data), and allocation of care.

Care provision and medical analyses refer to the actual "doing" of care: Prevention, prognosis, and diagnosis. Flawed COVID-19 prediction models [72], and digital symptom checkers [33] show issues with AI's accuracy/reliability. There were issues found in AI with bias/discrimination towards certain populations (most typically gendered and ethnic/racial) e.g., estimating kidney function [65] and in chest x-ray classifiers [76]. Finally, scientists criticized Google's lack of transparency in their breast cancer predicting AI [77].

Data management refers to the administration and logistics of handling personal and medical information: storing, sharing, and tracking of medical data. This concerns issues with surveillance/privacy such as the case of Amazon's Halo Band [32], a fitness tracker that constantly tracks medical data of the person wearing it, and an incident of asking for private medical data on the platform Facebook by a chatbot linked to the account of Israeli politician Netanyahu [69]. Also, transparency is an issue with storing, sharing, and tracking medical data, as for instance the transfer of medical data from a healthcare provider to Google was criticized for not informing the patients [22].

Concerning the allocation of care, two incidents were found: one involving the allocation of care work (i.e. how many hours a caregiver ought to spend with their patient) [45], and a case that concerns the allocation of Covid-19 vaccines [78]. Both of these incidents showed issues with accuracy/reliability, as the people in need of care were not able to access it due to inaccurate algorithms. At the same time, there were issues with transparency, as it was unclear in the case of Covid vaccination allocation how the algorithm makes its decision [78].

4.1.5 Automotive. All identified incidents in the automotive sector (n=21) involve self-driving cars in traffic. Three main causes were identified: external, human, and other (i.e. difficult to determine who/what caused the incident).

External incidents refer to incidents with self-driving cars in traffic due to external manipulation by researchers for the sake of calling for more security in self-driving cars [13, 68]. Self-driving cars were manipulated with, for instance, shiny stickers, drones with projectors, or through taking remote control to move seats, trigger indicators, wing mirrors, and windscreen wipers.

Human incidents concern incidents with self-driving cars in traffic due to human error. In these cases, drivers watched movies [34] or slept while using the autopilot [16], leading to slow responses of the driver when approaching subjects such as pedestrians or other cars or due to exceeding speed limits. Sometimes, human error does not necessarily refer to the human driver of an autonomous vehicle. Rather, two incidents in the database show how human error occurs when also other participants in traffic make estimation errors allegedly [10, 55].

In other types of incidents, it is difficult to determine the cause of the incident due to either the nature of the incident or lack of details reported about the incident. For instance, cases where the car could not detect a white vehicle due to bright weather while human drivers allegedly were not attentive enough [36, 37], or car crashes where details of the incident are missing [21, 38]. However, even though it is difficult to pinpoint responsibility and cause, it does not mean that there is no indication of possible technical issues: e.g. when the autopilot emergency braking systems were not used when an object or traffic situation was not (timely) detected [37] or all the lack of defensive driving when approaching a pedestrian, e.g. allegedly stopping too close to the subject [74].

Interestingly, there is a case involving two autonomous vehicles, i.e., a traffic situation where the key players are technologies, not humans. Two self-driving cars nearly collided when one car tried to switch lanes while being cut off by the other car. The crash was prevented as the first car detected the other one on time and waited until the lane was clear again [42].

All these incidents relate to safety, accuracy/reliability, and security issues with autonomous driving vehicles. Safety refers to (the lack of) physical harm when, for instance, a self-driving car crashes or is involved in any type of physical accident [e.g., 38]. Accuracy/reliability shows the lack of accuracy and reliability in the use of sensors e.g. for recognizing objects in traffic [e.g., 37]. Security deals with safety from external manipulation [13, 68].

4.2 Comparing sectors

While the analysis above provides a rich description of each specific ethical issue and how it relates to its respective sector, boiling down the results to key insights, one can identify the following overlap between sectors (table 1).

This indicates that there is merit in the approach of general AI ethics guidelines and principles because several issues are not sector specific but cut across different sectoral contexts: accuracy/reliability, bias/discrimination, transparency, surveillance/privacy, security. Yet, even though the findings show that ethical values span across sectors, there are sector-specific characteristics found in the data when looking at specific sectoral activities as is explained next.

In addition to understanding ethical issues within their sectoral context, specific sectoral activities are also studied. Focusing on these activities reveals further contextual characteristics of the sector: the actions that are inherent to the sector that AI systems got involved in. This gives additional information on the nature of the incidents.

When reading this table, three things are observable, 1) certain ethical issues are found in only one sector, 2) the same ethical issue

Table 1: Ethical issues listed by sector

Sector	Ethical issue
Police	Accuracy/reliability, bias/discrimination, transparency, surveillance/privacy
Education	Accuracy/reliability, bias/discrimination, transparency, surveillance/privacy, securitys
Academia	Premise/intent
Politics	Misinformation, transparency
Healthcare	Accuracy/reliability, bias/discrimination, privacy/surveillance, transparency
Automotive	Safety, security, accuracy/reliability

Table 2: Ethical issues and sectoral activities listed by sector

Sector	Ethical issue	Sectoral activity
Police	Accuracy/reliability, bias/discrimination, transparency, Surveillance/Privacy	Predictive or investigative tracking/identification/monitoring
Education	Accuracy/reliability, bias/discrimination, transparency, surveillance/privacy, security	Administrative work and teaching (Regulating access, tracking student behavior, evaluating work)
Academia	Premise/intent	Academic publishing
Politics	Misinformation, transparency	Political communication and persuasion
Healthcare	Accuracy/reliability, bias/discrimination, privacy/surveillance, transparency	Care provision and medical analyses, data management, allocation of care
Automotive	Safety, security, accuracy/reliability	Self-driving cars

that is being breached across sectors leads to different consequences and refers to different meanings 3) pre-existing sector-specific structures are reproduced, as is explained next.

First, there are some ethical issues that are inherently sector-specific. In academia, the only ethical concern found is the issue of having questionable premises or intentions when developing the technologies (table 1). In other words, the AI systems were not at fault when the incident occurred, rather the worldviews/theories that humans hold when developing the technologies. While this in itself could happen in other sectors, i.e. having bad intentions or unethical ideas about the sociotechnical, when looking at the sectoral activity, it shows that this relates specifically to scientific publishing (table 2). To elaborate, the realm in which this incident occurred is where academia produces its knowledge, i.e. the creation of scientific output. Framed differently, the premise/intent principle that is being violated relates to scientific ideas that are being published: ethically questionable hypotheses and theoretical premises prior to developing the AI, that guides the narrative in the publication. To give an example, an AI "gaydar" was developed in academia and got published [47]. Regardless of how the technology itself is working, the main theoretical starting point was ethically questionable, i.e. the "need" that one can or should scan faces to detect one's sexual orientation. In other words, the main concern for academia in relation to AI deployment is the social and moral theories that scholars that develop AI hold. Again, this is not something only academia struggles with, as developers in all sectors could have questionable premises/intent when developing their technologies. However, the data shows that it is inherent to academia to focus on scientific publishing, which by default links the questionable premises/intent with unethical hypothesizing. In other words, this intersection is a quirk specific to the sector of academia (i.e. coming up with unethical hypotheses and theoretical premises which then get tested and published).

Similarly, the ethical issue of "misinformation" was only found in the sector of politics, albeit together with the breached ethical principle of "transparency". When looking at the sectoral activity, one can see that it relates to political persuasion (table 2). All incidents in the political sector thereby relate to creating a political message with AI, but this message is 1) not disclosing that AI was made in the making of the message, thereby not being transparent and 2) having elements of deceit/not being factual, thereby being able to misinform the audience. To elaborate, the incidents all relate to deepfakes and twitter bots that spread political messages but are not informing the audience about the nature of such messages. While such technologies could also be used in other sectors for other purposes than politics, for instance, in popular culture for satire purposes, the sector "politics" specifically struggles with this phenomenon, as the data for instance do not show other AI-related incidents in other realms of the political sector (e.g. using AI for administrative work in the political sector) or other ethical values being breached that were repeatedly found across other sectors (e.g. accuracy/reliability). This does not mean that there are no other activities in the sector of "politics" where AI could get deployed in, but rather that the most pressing issue (according to media logic and the public sphere) where AI gets deployed are activities related to public persuasion. Political communication and persuasion are activities specific to the sector, to, for instance, assert dominance

of certain political ideas over others. By doing so, the sector "politics" is prone to misinformation and transparency when AI gets embedded into this context. The results show that the combination of ethical principles being breached, i.e. misinformation and transparency play into those pre-existing sector-specific characteristics: the struggle of competing belief systems.

Second, one can see that if the same ethical principles are breached, it leads to radically different consequences in different sectors, as the principle intersects with the sectoral activity. The same principle can manifest differently in different sectors due to the sectoral activity being involved. To give an example, when the principle "security" is being breached, which in all cases means the event where an external person hacks into a system, the consequences for the sectors "education" and "automotive" are very different. In "education", as seen in table 2, a security breach happens in administrative and teaching activities, i.e. everything that relates to grading or administering data and information about students. The worst outcome that could happen in a security breach, is that an external person would be able to access, retrieve, and modify personal data. However, for the automotive industry, a security breach could potentially have physical consequences as all AI-related incidents refer to self-driving vehicles (table 2). If external actors are able to hack the autopilot of cars, the possible effects are bodily. This does not mean that one ethical problem is lesser than the other. Rather, it means that in order for people to truly understand the nature of the breach of an ethical principle and its potential consequences, it has merit trying to understand the sectoral activity it is embedded in.

Related to this argument, not only are the consequences of ethical breaches different for different sectors, regarding transparency, the meaning of the ethical value in itself can be different for different sectors. Transparency in the political domain focuses primarily on the transparency *that* AI systems were used to, for instance, manipulate images or videos ("deepfakes") whereas the question of *how* the manipulation was conducted technically is less relevant from an ethical perspective. After all, the ethical breach lies in political persuasion (table 2) where the goal of the AI is to convince people of certain beliefs without revealing the lack of authenticity involved in creating the message, thereby it is irrelevant to know from an ethical perspective e.g. which data is used to create such AI systems or showcase technical documentation. In healthcare, however, transparency refers to technical elements of AI such as data and methods that are used to construct the AI that could influence for instance care provision and medical analyses (table 2). Whereas the former breaches of transparency concern the lack of revealing *that* an AI was used, the latter refers to the lack of transparency involved where AI classifies things or comes to a certain decision.

Third, when intersecting the ethical value with the sectoral activity, it raises the question whether the phenomena are really new or whether it is rooted in a sectoral structure. As an example, the sector police is discussed in detail. Surveillance and privacy are ethical issues that could be seen as inherent to police work, since police work is a form of state governance that, with or without AI and machine learning, involves monitoring and identifying suspects [48]. This would require some form of gathering personal information from people. Also, when tracking, identifying or monitoring people, bias and discrimination is not a new phenomenon following the introduction of new technologies, but police work has previously been associated with racial bias [8]. Of course, the source of human bias and machine bias might be different. But the phenomenon itself in the police force is not new. In terms of transparency, it should equally be questioned whether law enforcement has thus far, i.e. without AI, been transparent in terms of how they collect their data and how much this differs when AI is being used. Finally, accuracy/reliability is an ethical theme that refers to the technicalities of the AI: if it works as it is intended. Thereby this theme by default does not discuss e.g. how accurately human police officers identify their suspects, but rather how well a machine performs this task. Nonetheless, one could still make the claim that also without AI police work has issues with accuracy/reliability, since making mistakes such as misidentification and making false estimations is by definition a human quality.

Of course, it is one-sided to claim that all of these ethical issues are inherent to the sector without intervention of AI systems, as if technologies do not introduce societal change and new ethical issues. To take the principle of surveillance/privacy as an example: one could for instance argue that the scalability of surveillance and privacy breaches in relation to tracking, identifying and monitoring people have the potential to increase or change form. To elaborate, Andrejevic and Gates argue that whereas prior, surveillance was targeted, data-driven surveillance techniques allow for a "collect-everything approach" [4]. However, this current paper does not deny that AI systems could trigger social change in form or intensity. Rather, the main argument is that the *very premise* of these ethical issues is sometimes inherent to the sector. For example, one of the core activities in the police sector is surveilling. It is thereby no surprise that ethical breaches occurred related to privacy, transparency and surveillance, when AI got deployed in this sector. In other words, the ethical problems with AI are arguably rooted in something rather stable and structural: specific sectoral routines and structures.

5 DISCUSSION

The results show that most ethical themes are recurring across sectors: accuracy/reliability, bias/discrimination, transparency, surveillance/privacy, security. This means that it makes sense to discuss ethical issues on a more general level as there is empirical evidence that some principles are repeatedly breached in across contexts. General ethical values and principles can and should be addressed when, for instance, discussing and conceptualizing ethics in policies, academic texts, or public communication. Furthermore, what this present paper also shows, is that additionally, knowing sectoral context can be helpful when understanding AI ethics in-depth as is explained next.

Taking sectoral context into consideration, one becomes aware they have their own dynamics and routines: the police surveils, teachers administer tests, physicians diagnose. Understanding these activities helps with understanding AI ethics better, as it is no surprise when AI gets deployed, e.g. issues of safety occur in the automotive industry, misinformation and transparency in politics, or

questionable theoretical premises are put forward in academic publishing because such principles and values are related to their respective sector and their specific activities. In other words, the results show that there is merit in understanding how ethical principles intersect with sectoral activities, as these reveal specific meaning of AI deployment in specific sectoral contexts. Therefore scholars, developers and operators, and other actors of AI systems ought to take sectoral context that an AI system is deployed in into account because each sector has its own quirks. Moving forward, applying a sector-based approach to AI ethics means studying the activities of that respective sector. Taking it one step further, one could even argue that domain specific knowledge is needed to assess sectors before AI deployment with e.g. a historical analysis. By doing so, one can understand *why* some ethical issues are more prevalent in a sector than others, even before AI systems are deployed.

Knowing that sectors have specific cultures and quirks, has several implications for the field of AI ethics. First, a sector-based approach argues for sector-specific sensitivity when discussing guidelines. A sector-based approach to AI ethics can address varying demands on and trade-offs to ethical values and principles. For instance, in policing, there is a legitimate interest for some level of secrecy to not hamper police investigations. Other sectors make different demands on trade-offs to ethical values and principles. Therefore to evaluate privacy or transparency-related incidents, requires to make sector-specific considerations. To give another example of such trade-offs, while some of the accuracy- and reliability-related issues of autonomous vehicles presumably can be best solved by advancing capabilities on the basis of providing more training data, such calls for ever more data are problematic in other sectors where data is often more personal and sensitive. For instance, the increased use of personal data in education is considered to be highly problematic due to privacy issues and problems regarding consent [41, 67]. In policing, the surveillance necessary to acquire data is a vividly discussed ethical issue itself [4]. The solutions that AI ethics guidelines suggest using to address specific ethical issues need to take these context-specific requirements for solutions to ethical issues into account. In contrast to non-contextualized general AI ethics guidelines, sector-specific guidelines with their much smaller scope can name and discuss sector-specific risks, and, in doing so, provide much more awareness for specific ethical issues. For instance, the historical issues in the political sector concerning attempts to persuade masses with certain beliefs and thereby not always being truthful or honest about their reporting (regardless of the use of AI or not), shows that the ethical value of accuracy/reliability of the system (as shown in the results, an often-found problem across the sectors) is less relevant to focus on compared with misinformation and transparency. In other words, a sector-based approach shows how certain issues are particularly relevant for some sectors, while less so for others. A sector-based approach to AI ethics can take these differences into account.

So far, contextuality is highlighted as one of the key aspects of a sector-based approach: try to understand each sector's activities, because ultimately, the technology gets embedded in this context and might reproduce and reinforce ethical issues that are already present. However, what this perspective does not offer, is an outlook on how AI technologies are able to *change* the dynamics of the sector. For instance, while the analysis shows that the automotive industry has safety in their breaches of ethical values, it does not show how autonomous vehicles could change the notion and perception of safety if a car is driven by non-human drivers compared with human drivers or the scalability of faked/inauthentic political messages in the field of political persuasion with deepfakes and twitter bots.

A second limitation concerns the data used in this study. This study shows representations of incidents, as they are represented in media reports – i.e., secondary data. This means, that 1) the narrative is framed by media outlets with their own media logic (i.e. the inner workings of the media sector), although it should be noted that the performed method of analysis is not on a semantic level, and 2) it might be that other kinds of incidents occurred post-deployment, but were not picked up by the sampled media reports. For instance, in politics, all cases except one relate to deepfakes that communicate political messages or ridicule political personas. However, the political sector concerns more than mediated messages. It is also an administrative institution in which AI technologies could be used. Similarly, in the automotive sector, media reports focused primarily on incidents with self-driving cars. Yet, AI systems might also be used for administrative purposes in the automotive sector and different ethical issues might arise there. Such potential blindspots could be related to having media reports as the unit of analysis. Future research on incidents could consider different types of data to understand human-computer interaction or human-robot interaction "in the wild", with e.g. an ethnography.

Third, the sampling strategy of this paper ended in 2021. Arguably, many other AI technologies have been introduced and deployed since then. The deployment of AI and its consequences are a moving target to study, and therefore it is important to study how the landscape of AI ethics has changed over time. Follow-up studies could thereby replicate this research to understand if the sheer increase in incidents also somehow diversifies the nature of the incidents in their breaches of ethical principles in particular sectoral contexts.

6 CONCLUSION

This article makes the case for a sector-based approach to AI ethics, in which sectoral context is regarded as relevant information to understand the ethics of AI deployment. To do so, it analyzes n=125 incidents from the AIAAIC repository [1] from the sectors police, education/academia, politics, healthcare, and automotive. The analysis shows that while certain ethical issues are recurring and their relevance spans across sectors, 1) other ethical issues are inherently related to specific sectors, 2) ethical issues appear to have different meanings and manifest differently in different social contexts 3) the problems with AI-deployment are related to pre-existing issues in the sector (i.e. prior to AI deployment). Instead of asking how AI ethics ought to look like, a sector-based approach argues to look at the activities and pre-existing social realities of such sectors, in order to understand the situated context of AI deployment. It serves as an addition to general AI ethics guidelines that have been described by the AI ethics community in terms of their vagueness, high level of abstraction, and ambiguity, as well as them being generic, difficult to apply, and vague [31, 43, 60]. While these

principles could be viewed as rather generic etc., they are empirically found breached across contexts. A sector-based approach serves as an additional view to AI ethics that enables scholars and practitioners to understand the relevance of sectoral cultures in AI deployment.

ACKNOWLEDGMENTS

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2002/1 "Science of Intelligence" – project number 390523135.

Funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 952026.

This work was partly done within etami.

REFERENCES

[1] AIAAIC. n.d. AIAAIC Repository. Retrieved from https://aiaaic.org

[2] algo.rules. 2019. *Regeln für die Gestaltung algorithmischer Systeme.* iRights.lab and Bertelsmann Stiftung. Retrieved from https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/Algo.Rules_DE.pdf

[3] Mike Ananny. 2016. Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness. *Sci. Technol. Hum. Values* 41, 1 (January 2016), 93–117. DOI:https://doi.org/10.1177/0162243915606523

[4] Mark Andrejevic and Kelly Gates. 2014. Big Data Surveillance: Introduction. *Surveill. Soc.* 12, 2 (2014), 185–196. DOI:https://doi.org/10.24908/ss.v12i2.5242

[5] Audrey Azoulay. 2019. Towards an ethics of artificial intelligence. *UN Chron.* 55, 4 (January 2019), 24–25. DOI:https://doi.org/10.18356/3a8f673a-en

[6] Yang Bao, Gilles Hilary, and Bin Ke. 2022. Artificial Intelligence and Fraud Detection. In *Innovative Technology at the Interface of Finance and Operations: Volume I*, Volodymyr Babich, John R. Birge and Gilles Hilary (eds.). Springer International Publishing, Cham, 223–247. DOI:https://doi.org/10.1007/978-3-030-75729-8_8

[7] Sarah Basford. 2020. Australian schools have been trialing facial recognition technology, despite serious concerns about children's data. *Gizmodo Australia.* Retrieved from https://www.gizmodo.com.au/2020/03/australian-schools-trial-facial-recognition-technology-looplearn/

[8] Sandra Bass. 2001. Policing space, policing race: Social control imperatives and police discretionary decisions. *Soc. Justice* 28, 1 (83) (2001), 156–176. Retrieved from https://www.jstor.org/stable/29768062

[9] BBC. 2020. Study finds quarter of climate change tweets from bots. *BBC.* Retrieved from https://www.bbc.co.uk/news/technology-51595285

[10] Max Bergen. 2016. Google's Self-Driving Car Hit Another Vehicle for the First Time. *Vox.* Retrieved from https://www.vox.com/2016/2/29/11588346/googles-self-drivingcar-hit-another-vehicle-for-the-first-time

[11] Sam Biddle. 2021. LAPD sought ring home security video related to black lives matter protests. *The Intercept.* Retrieved from https://theintercept.com/2021/02/16/lapd-ring-surveillance-black-lives-matter-protests/

[12] Pal Boza and Theodoros Evgeniou. 2021. Implementing AI principles: Frameworks, processes, and tools. *INSEAD Work. Pap.* 2021/04/DSC/TOM, (2021). DOI:http://dx.doi.org/10.2139/ssrn.3783124

[13] Thomas Brewster. 2019. Hackers use little stickers to trick tesla autopilot into the wrong lane. *Forbes Magazine.* Retrieved from https://www.forbes.com/sites/thomasbrewster/2019/04/01/hackers-use-little-stickers-to-trick-tesla-autopilot-into-the-wrong-lane/

[14] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, Proceedings of Machine Learning Research, 77–91. Retrieved from http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

[15] Dafna Burema. 2022. A critical analysis of the representations of older adults in the field of human–robot interaction. *AI Soc.* 37, 2 (June 2022), 455–465. DOI:https://doi.org/10.1007/s00146-021-01205-0

[16] Leyland Cecco. 2020. Tesla driver found asleep at wheel of self-driving car doing 150km/h. *The Guardian.* Retrieved from https://www.theguardian.com/world/2020/sep/17/canadatesla-driver-alberta-highway-speeding

[17] Tomás Chamorro-Premuzic and Reece Akhtar. 2019. Should Companies Use AI to Assess Job Candidates? *Harvard Business Review.* Retrieved from https://hbr.org/2019/05/should-companies-use-ai-to-assess-job-candidates

[18] Raja Chatila and John C. Havens. 2019. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. In *Robotics and Well-Being*, Maria Isabel Aldinhas Ferreira, João Silva Sequeira, Gurvinder Singh Virk, Mohammad Osman Tokhi and Endre E. Kadar (eds.). Springer International Publishing, Cham, 11–16. DOI:https://doi.org/10.1007/978-3-030-12524-0_2

[19] Monica Chin. 2020. These Students Figured Out Their Tests Were Graded by AI. *The Verge.* Retrieved from https://www.theverge.com/2020/9/2/21419012/edgenuity-online-class-ai-grading-keyword-mashing-students-school-cheating-algorithm-glitch

[20] Monica Chin. 2021. ExamSoft's proctoring software has a face-detection problem. Retrieved from https://www.theverge.com/2021/1/5/22215727/examsoft-online-exams-testing-facial-recognition-report

[21] Devin Coldewaey. 2019. Tesla explodes after crash on Russian highway. *techcrunch.* Retrieved from https://techcrunch.com/2019/08/11/tesla-explodes-after-crash-on-russianhighway/

[22] Rob Copeland. 2019. Google's 'Project Nightingale' Gathers Personal Health Data on Millions of Americans. *The Wall Street Journal.* Retrieved from https://www.wsj.com/articles/google-s-secret-project-nightingale-gathers-personal-health-data-on-millions-of-americans-11573496790

[23] Kate Crawford. 2021. *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence.* Yale University Press. Retrieved from https://doi.org/10.12987/9780300252392

[24] Kate Crawford and Ryan Calo. 2016. There is a blind spot in AI research. *Nature* 538, (2016), 311–313. DOI:https://doi.org/10.1038/538311a

[25] European Commission. Directorate General for Communications Networks, Content and Technology. 2020. *Sectoral Considerations on the Policy and Investment Recommendations for Trustworthy Artificial Intelligence.* Publications Office, LU. Retrieved February 22, 2023 from https://data.europa.eu/doi/10.2759/733662

[26] Theodoros Evgeniou, David R Hardoon, and Anton Ovchinnikov. 2020. What Happens When AI is Used to Set Grades. *Harvard Business Review.* Retrieved from https://hbr.org/2020/08/what-happens-when-ai-is-used-to-set-grades

[27] Todd Feathers. 2020. Facial Recognition Company Lied to School District About its Racist Tech. *Vice.* Retrieved from https://www.vice.com/en/article/qjpkmx/facrecognition-company-lied-to-school-district-about-its-racist-tech

[28] Todd Feathers. 2021. Major Universities Are Using Race as a "High Impact Predictor" of Student Success. *The Markup.* Retrieved from https://themarkup.org/machine-learning/2021/03/02/major-universities-are-using-race-as-a-high-impact-predictor-of-student-success

[29] Todd Feathers and Janus Rose. 2020. Students Are Rebelling Against Eye-Tracking Exam Surveillance Tools. *Vice.* Retrieved from https://www.vice.com/en/article/n7wxvd/students-are-rebelling-against-eye-tracking-exam-surveillance-tools

[30] Andrew Guthrie Ferguson. 2017. Policing predictive policing. *Wash. Univ. Law Rev.* 94, 5 (2017), 1115–1194. Retrieved from https://ssrn.com/abstract$=$2765525

[31] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. 2020. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Cent. Res. Publ.* 2020–1 (2020). DOI:http://dx.doi.org/10.2139/ssrn.3518482

[32] Fowler, Geoffrey and Kelly, Heather. 2020. Amazon's new health band is the most invasive tech we've ever tested. *The Washington Post.* Retrieved from https://www.washingtonpost.com/technology/2020/12/10/amazon-halo-band-review/

[33] Hamish Fraser, Enrico Coiera, and David Wong. 2018. Safety of patient-facing digital symptom checkers. *The Lancet* 392, 10161 (November 2018), 2263–2264. DOI:https://doi.org/10.1016/S0140-6736(18)32819-8

[34] Gary Gastelu. 2020. Tesla on autopilot hits police car as driver watches movie on cellphone. *Fox News.* Retrieved from https://www.foxnews.com/auto/tesla-on-autopilot-hits-police-car-as-driver-watches-movie-on-cellphone

[35] Google. 2018. Artificial Intelligence at Google: Our Principles.

[36] Andrew J Hawkins. 2019. Tesla didnt fix an autopilot problem for three years, and now another person is dead. *The Verge.* Retrieved from https://www.theverge.com/2019/5/17/18629214/tesla-autopilot-crash-death-josh-brown-jeremy-banner

[37] Yoni Heisler. 2020. Wild video shows a Tesla Model 3 on Autopilot crashing into a truck. *BGR.* Retrieved from https://bgr.com/tech/tesla-crash-model-3-autopilot-truck-taiwan/

[38] Jo He-Rim. 2020. Tesla accident: Faulty vehicle or bad driving? *The Korea Herald.* Retrieved from http://www.koreaherald.com/view.php?ud$=$20201213000152

[39] Merve Hickok. 2021. Lessons learned from AI ethics principles for future actions. *AI Ethics* 1, 1 (2021), 41–47.

[40] High-Level Expert Group on Artificial Intelligence. 2019. Ethics guidelines for trustworthy AI. Retrieved January 4, 2023 from https://ec.europa.eu/newsroom/dae/document.cfm?doc_id$=$60419

[41] Wayne Holmes, Kaska Porayska-Pomsta, Ken Holstein, Emma Sutherland, Toby Baker, Simon Buckingham Shum, Olga C. Santos, Mercedes T. Rodrigo, Mutlu Cukurova, Ig Ibert Bittencourt, and Kenneth R. Koedinger. 2022. Ethics of AI in Education: Towards a Community-Wide Framework. *Int. J. Artif. Intell. Educ.* 32, 3 (2022), 504–526. DOI:https://doi.org/10.1007/s40593-021-00239-1

[42] Chris Isidore. 2015. Self-driving cars from rivals Google, Delphi in close call. *CNN.* Retrieved from https://money.cnn.com/2015/06/26/autos/self-driving-car-near-accident/index.html

[43] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* 1, 9 (September 2019), 389–399. DOI:https://doi.org/10.1038/s42256-019-0088-2

[44] Jason Koebler. 2020. Detroit Police Chief: Facial Recognition Software Misidentifies 96% of the Time. *Vice.* Retrieved from https:

//www.vice.com/en/article/dyzykz/detroit-police-chief-facial-recognition-software-misidentifies-96-of-the-time
[45] Colin Lecher. 2020. Can a Robot Decide My Medical Treatment? *The Markup*. Retrieved from https://themarkup.org/the-breakdown/2020/03/03/healthcare-algorithms-robot-medicine
[46] Colin Lecher. 2020. Remote Exam Software Is Crashing When the Stakes Are the Highest. *The Markup*. Retrieved from https://themarkup.org/coronavirus/2020/10/13/remote-exam-software-failures-privacy
[47] Sam Levin. 2017. New AI can guess whether you're gay or straight from a photograph. *The Guardian*. Retrieved from https://www.theguardian.com/technology/2017/sep/07/new-artificial-intelligence-can-tell-whether-youre-gay-or-straight-from-a-photograph
[48] David Lyon, Kevin D Haggerty, and Kirstie Ball. 2012. Introducing surveillance studies. In *Routledge handbook of surveillance studies*. Routledge, 1–11.
[49] Ayang Macdonald. 2020. Privacy concerns greet adoption of facial recognition system by India's secondary education board. *biometricupdate*. Retrieved from https://www.biometricupdate.com/202010/privacy-concerns-greet-adoption-of-facial-recognition-system-by-indias-secondary-education-board
[50] Andrew McNamara, Justin Smith, and Emerson Murphy-Hill. 2018. Does ACM's code of ethics change ethical decision making in software development? In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ACM, Lake Buena Vista FL USA, 729–733. DOI:https://doi.org/10.1145/3236024.3264833
[51] Microsoft. 2017. Microsoft responsible AI principles. Retrieved January 3, 2023 from https://www.microsoft.com/en-us/ai/our-approach
[52] Luke Munn. 2022. The uselessness of AI ethics. *AI Ethics* (August 2022). DOI:https://doi.org/10.1007/s43681-022-00209-w
[53] Christopher Nilesh. 2020. We've Just Seen the First Use of Deepfakes in an Indian Election Campaign. *Vice*. Retrieved from https://www.vice.com/en/article/jgedjb/thefirst-use-of-deepfakes-in-indian-election-by-bjp
[54] OECD. 2019. Forty-Two Countries Adopt New Principles on Artificial Intelligence. Retrieved January 3, 2020 from https://www.oecd.org/science/forty-two-countries-adopt-new-oecd-principles-on-artificial-intelligence.htm
[55] Madison Park. 2017. Self-driving bus involved in accident on its first day. *CNN*. Retrieved from https://money.cnn.com/2017/11/09/technology/self-driving-bus-accident-las-vegas/index.html
[56] Beth Pearsall. 2010. Predictive policing: The future of law enforcement. *Natl. Inst. Justice J.* 266, 1 (2010), 16–19. Retrieved from https://mediaweb.saintleo.edu/courses/CRJ570/PredictivePolicing_Pearsall.pdf
[57] Dorian Peters, Karina Vold, Diana Robinson, and Rafael A. Calvo. 2020. Responsible AI—Two Frameworks for Ethical Design Practice. *IEEE Trans. Technol. Soc.* 1, 1 (March 2020), 34–47. DOI:https://doi.org/10.1109/TTS.2020.2974991
[58] Katyanna Quach. 2020. Uni revealed it killed off its PhD-applicant screening AI – just as its inventors gave a lecture about the tech. *The Register*. Retrieved from https://www.theregister.com/2020/12/08/texas_compsci_phd_ai/
[59] Anaïs Rességuier and Rowena Rodrigues. 2020. AI ethics should not remain toothless! A call to bring back the teeth of ethics. *Big Data Soc.* 7, 2 (2020), 1–5. DOI:https://doi.org/10.1177/2053951720942541
[60] Catharina Rudschies, Ingrid Schneider, and Judith Simon. 2021. Value Pluralism in the AI Ethics Debate – Different Actors, Different Priorities. *Int. Rev. Inf. Ethics* 29, (March 2021). DOI:https://doi.org/10.29173/irie419
[61] Mark Ryan and Bernd Carsten Stahl. 2021. Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. *J. Inf. Commun. Ethics Soc.* 19, 1 (2021), 61–86. DOI:https://doi.org/10.1108/JICES-12-2019-0138
[62] Daniel Schiff, Bogdana Rakova, Aladdin Ayesh, Anat Fanti, and Michael Lennon. 2020. Principles to Practices for Responsible AI: Closing the Gap. *ArXiv Prepr.* (2020). DOI:https://doi.org/10.48550/ARXIV.2006.04707
[63] Kaira Sekiguchi and Koichi Hori. 2020. Organic and dynamic tool for use with knowledge base of AI ethics for promoting engineers' practice of ethical AI design. *AI Soc.* 35, 1 (March 2020), 51–71. DOI:https://doi.org/10.1007/s00146-018-0867-z
[64] Sofia Serholt, Wolmet Barendregt, Asimina Vasalou, Patrícia Alves-Oliveira, Aidan Jones, Sofia Petisca, and Ana Paiva. 2017. The case of classroom robots: teachers' deliberations on the ethical tensions. *AI Soc.* 32, 4 (November 2017), 613–631. DOI:https://doi.org/10.1007/s00146-016-0667-2
[65] Tom Simonite. 2020. How an Algorithm Blocked Kidney Transplants to Black Patients. *Wired*. Retrieved from https://www.wired.com/story/how-algorithm-blocked-kidney-transplants-black-patients/
[66] José Antonio Siqueira De Cerqueira, Lucas Dos Santos Althoff, Paulo Santos De Almeida, and Edna Dias Canedo. 2021. Ethical perspectives in ai: A two-folded exploratory study from literature and active development projects. In *Proceedings of the 54th Hawaii International Conference on System Sciences*, University of Hawai'i at Manoa, Honolulu, 5240–5249. Retrieved from http://hdl.handle.net/10125/71257
[67] Sharon Slade and Paul Prinsloo. 2013. Learning Analytics: Ethical Issues and Dilemmas. *Am. Behav. Sci.* 57, 10 (October 2013), 1510–1529. DOI:https://doi.org/10.1177/0002764213479366
[68] Olivia Solon. 2016. Team of hackers take remote control of Tesla Model S from 12 miles away. *The Guardian*. Retrieved from https://www.theguardian.com/technology/2016/sep/20/tesla-model-s-chinese-hack-remote-control-brakes
[69] Amir Tal. 2021. Facebook suspends Israeli Prime Minister Benjamin Netanyahu-linked chatbot for breaking its privacy rules. *CNN*. Retrieved from https://edition.cnn.com/2021/01/25/middleeast/israel-facebook-netanyahu-chatbot-intl/index.html
[70] Li Tao. 2018. Jaywalkers under surveillance in Shenzhen soon to be punished via text messages. *South China Morning Post*. Retrieved from https://www.scmp.com/tech/china-tech/article/2138960/jaywalkers-under-surveillance-shenzhen-soon-be-punished-text
[71] The Economist. 2018. A faked video of Donald Trump points to a worrying future. *The Economist*. Retrieved from https://www.economist.com/leaders/2018/05/24/a-faked-video-of-donald-trump-points-to-a-worrying-future
[72] Alexandra Thompson. 2020. Coronavirus: Models predicting patient outcomes may be "flawed" and "based on weak evidence." *Yahoo!* Retrieved from https://sg.style.yahoo.com/style/coronavirus-covid19-models-patient-outcomes-flawed-153227342.html
[73] Ville Vakkuri, Kai-Kristian Kemell, Joni Kultanen, Mikko Siponen, and Pekka Abrahamsson. 2019. Ethically Aligned Design of Autonomous Systems: Industry viewpoint and an empirical study. (2019). DOI:https://doi.org/10.48550/arXiv.1906.07946
[74] Jackie Ward. 2018. Self-Driving Car Ticketed; Company Disputes Violation. *CBS Local San Francisco*. Retrieved January 3, 2021 from https://www.cbsnews.com/sanfrancisco/news/self-driving-car-ticketed-san-francisco/
[75] Jess; Nyrup Whittlestone Rune; Alexandrova, Anna, Rune Nyrup, Anna Alexandrova, Kanta Dihal, and Stephen Cave. 2019. *Ethical and Societal Implications of Data and Artificial Intelligence: a roadmap for research*. Nuffield Foundation, London. Retrieved from https://www.nuffieldfoundation.org/wp-content/uploads/2019/02/Ethical-and-Societal-Implications-of-Data-and-AI-report-Nuffield-Foundat.pdf
[76] Kyle Wiggers. 2020. Researchers find evidence of racial, gender, and socioeconomic bias in chest X-ray classifiers. *VentureBeat*. Retrieved from https://venturebeat.com/ai/researchers-find-evidence-of-racial-gender-and-socioeconomic-bias-in-chest-x-ray-classifiers/
[77] Kyle Wiggers. 2020. Google's breast cancer-predicting AI research is useless without transparency, critics say. *VentureBeat*. Retrieved from https://venturebeat.com/ai/googles-breast-cancer-predicting-ai-research-is-useless-without-transparency-critics-say/
[78] Kyle Wiggers. 2020. COVID-19 vaccine distribution algorithms may cement health care inequalities. *VentureBeat*. Retrieved from https://venturebeat.com/ai/covid-19-vaccine-distribution-algorithms-may-cement-health-care-inequalities/
[79] Kyle Wiggers. 2021. Outlandish Stanford facial recognition study claims there are links between facial features and political orientation. *VentureBeat*. Retrieved from https://venturebeat.com/ai/outlandish-stanford-facial-recognition-study-claims-there-are-links-between-facial-features-and-political-orientation/
[80] Cam Wilson. 2020. Australia's First Deepfake Political Ad is Here and it's Extremely Cursed. *Gizmodo Australia*. Retrieved from https://www.gizmodo.com.au/2020/11/australias-first-deepfake-political-ad-is-here-and-its-extremely-cursed/