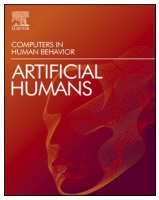
Computers in Human Behavior: Artificial Humans 2 (2024) 100055

Contents lists available at ScienceDirect

Computers in Human Behavior: Artificial Humans

journal homepage: www.journals.elsevier.com/computers-in-human-behavior-artificial-humans



User-driven prioritization of ethical principles for artificial intelligence systems

Yannick Fernholz [a,*], Tatiana Ermakova [b], B. Fabian [c,d], P. Buxmann [e]

[a] Weizenbaum-Institut e.V., Hardenbergstraße 32, 10632, Berlin, Germany
[b] HTW Berlin, 10313, Berlin, Germany
[c] Humboldt University, Information Systems, 10178, Berlin, Germany
[d] Technical University of Applied Sciences Wildau, EDIH pro Digital, 15745, Wildau, Germany
[e] Technical University Darmstadt, 64289, Darmstadt, Germany

ARTICLE INFO

Index Terms:
Artificial intelligence
Ethics
Ethical guidelines
Trustworthy AI
Requirements prioritization

ABSTRACT

Despite the progress of Artificial Intelligence (AI) and its contribution to the advancement of human society, the prioritization of ethical principles from the viewpoint of its users has not yet received much attention and empirical investigations. This is important to develop appropriate safeguards and increase the acceptance of AI-mediated technologies among all members of society.

In this research, we collected, integrated, and prioritized ethical principles for AI systems with respect to their relevance in different real-life application scenarios.

First, an overview of ethical principles for AI was systematically derived from various academic and non-academic sources. Our results clearly show that transparency, justice and fairness, non-maleficence, responsibility, and privacy are most frequently mentioned in this corpus of documents.

Next, an empirical survey to systematically identify users' priorities was designed and conducted in the context of selected scenarios: AI-mediated recruitment (human resources), predictive policing, autonomous vehicles, and hospital robots.

We anticipate that the resulting ranking can serve as a valuable basis for formulating requirements for AI-mediated solutions and creating AI algorithms that prioritize user's needs. Our target audience includes everyone who will be affected by AI systems, e.g., policy makers, algorithm developers, and system managers as our ranking clearly depicts user's awareness regarding AI ethics.

1. Introduction

Artificial intelligence (AI) systems nowadays contribute to people's daily lives and have both positive as well as negative consequences for individuals and society as a whole (Mirbabaie et al., 2022), (Sengupta et al., 2020), (Bingley et al., 2023), (Oppermann et al., 2019). Concrete application examples encompass various fields. For example, AI systems are used for fraud detection, to support predictive and prescriptive analytics, for image processing in medicine, for recommender systems, connecting people and providing entertainment in many ways (Sengupta et al., 2020), (Willis, 2018), (Plummer). Additionally, talking devices, digital assistants, and autonomous driving vehicles have become possible and widespread (Mirbabaie et al., 2022). However, these contributions are also accompanied by a growing number of negative examples where harm has resulted from technology that is not advanced (e.g., bias and discrimination, credit denial and medical misdiagnosis) (Burrell & Fourcade, 2021), (Scheuerman et al., 2020), (Svaldi), (Soper) and from the misuse of technology (e.g., user manipulation, facial recognition surveillance, mass data collection without consent, etc.) (Kazim & Koshiyama, 2021). This gives rise to a variety of new ethical, legal, and social challenges that can have serious negative consequences if not handled appropriately (El Khattabi et al., 2018), (King et al., 2020), (Floridi & Cowls, 2019), (Thiebes et al., 2021).

Civil organizations, research centers, private companies and governmental agencies made their commitments and insights public and contributed to the formulation of overarching ethical principle for the

* Corresponding author.
E-mail addresses: yannick.fernholz@weizenbaum-institut.de (Y. Fernholz), Tatiana.ermakova@htw-berlin.de (T. Ermakova), bfabian@wiwi.hu-berlin.de, benjamin.fabian@th-wildau.de (B. Fabian), peter.buxmann@tu-darmstadt.de (P. Buxmann).

https://doi.org/10.1016/j.chbah.2024.100055
Received 4 September 2023; Received in revised form 2 February 2024; Accepted 3 February 2024
Available online 10 February 2024

development and implementation of AI. As a result, a whole body of ethical guidelines has been developed in recent years, which are collecting principles to control the development of AI applications (Hagendorff, 2020). Institutions that use ethical AI principles are expected to gain the support and loyalty of the majority of users (Kaur et al., 2022). However, these ethical guidelines tend to focus on abstract concepts and thus can be difficult to implement in practice (Shneiderman, 2021), (Kiritchenko & Mohammad, 2017). Thus, this work sets out to shine light on that issue from the viewpoint of prioritization. The pertinent question arises as to what ethical principles should be prioritized when development resources are limited in the operating environment. Therefore, this work focuses on two research questions: (1) What is the current development of ethical principles for AI applications? (2) How do (potential) users prioritize these principles in the context of various application scenarios?

Our first contribution is a systematic and structured analysis of current ethical principles and guidelines in soft law and academic publications. Then, the results were used to create an empirical survey to prioritize these principles in various application scenarios to get a better understanding of which ethical aspects need to be integrated into AI algorithms in the future as to serve user's needs appropriately.

This article is organized as follows. Section two reviews the current literature on ethics and AI. Section three presents the methodological steps underlying this work. Section four presents the results, and section five provides a discussion, implications for future research, and limitations of this work.

2. Ethics and AI

The technological development and growing application of AI systems are making a significant contribution to many people's lives all over the world in various situations ranging from personal to professional settings (Li et al., 2019). The complexity and capability of what AI is already capable of make its usage unique and controversial at the same time. AI may cause mass unemployment, make independent decisions that people cannot control or understand, lead to wealth redistribution, and replace unique human tasks eventually (Wang & Siau, 2018).

Since the concept of machine ethics was proposed (Anderson et al., 2006), the ethical issues of AI have mostly been discussed by scholars. Compared to its global attention and the investment in AI technology, the consideration of AI ethics and morality is just at its beginning. Some argue that there is no rush to consider ethical problems related to technology since there still is a long way for AI to be comparable to humans and have consciousness. Yet many researchers believe that ethics and morality issues must be considered as soon as possible before these issues related to AI become importunate (Wang & Siau, 2018), since within half a decade machine learning applications have progressively spread their roots into most aspects of our daily lives (Prates et al., 2018).

However, while most studies concur that AI brings many benefits, there are also many examples that highlight ethical concerns which need to be dealt with. AI could pose risks to personal data protection and privacy issues, a risk of discrimination when algorithms are used for profiling purposes or to resolve situations in criminal justice (Madiega, 2019). For example, facial recognition software has gained popularity in the last years and is nowadays commonplace, from organizing the pictures on our phones to predicting criminal suspects. The ethical validity of these technologies was questioned by the recent discovery of the phenomenon of machine bias: the process by which personal preconceptions of AI engineers find their way into projects in which they are involved The list of ethical concerns and resulting challenges of immediate or future relevance faced by AI researchers is extensive (Prates et al., 2018).

Ethical dilemmas refer to situations in which any available choice leads to the infringement of some ethical principle while a decision must still be made (Kirkpatrick, 2015). The AI research community realizes that machine ethics is a determining factor to the extent autonomous systems are permitted to interact with human beings (Yu et al., 2018). Ethical issues emerge whenever a decision or an action may affect the well-being of an individual or a group of people. Dilemma situations arise because competing moral values or conflicting factors become relevant due to the absence of universally accepted decision-making criteria or outcome preferences (Martinsons & Ma, 2009).

Ethics is commonly referred to as the study of morality. In this work, morality is understood as a system of rules and values for guiding human behavior, actions, and principles for evaluating those rules. Consequently, ethical behavior does not necessarily mean good behavior, but it indicates compliance with specific values. These values are commonly accepted as being part of human nature (protection of human life, freedom and dignity) or as a moral expectation characterizing beliefs and convictions of specific groups of people (e.g., religious rules) (Walz & Firth-Butterfield, 2019).

Moral expectations may be of individual nature and therefore differ among two people, regardless of if they share the same cultural and moral values or not. This broad definition is used here as the intention of this work is not to approach AI from a specific normative perspective and therefore to contribute to the discussion on the determination of appropriate regulatory means to implement ethics into AI. In addition, the benefit of this neutral definition of ethics is that it addresses the issue of ethical diversity from a regulatory and policy-making perspective (Walz & Firth-Butterfield, 2019).

As has been acknowledged by (Whittlestone et al., 2019), principles should be seen as a starting point from which standards and eventually regulations can be developed, but it is necessary to look at specific use cases of AI and evaluate tensions between values. For principles to be practical useful, they need to be able to guide action in concrete situations. Already back in 2004, (Dancy, 2004) has argued to focus on specific cases entirely. (Jakesch et al., 2022) argue that a more diverse ethical judgement must be incorporated into the AI development process and that little is known about the priorities of values of different stakeholders.

3. Methodology

3.1. Introduction: elicitation, analysis, and prioritization

The methodology of this study is divided into three separately performed phases to ensure a structural and analytical approach. The steps build onto each other and in combination comprise the methodological base of this research. Phases one and two control each other's results and thus ensure that the data used in phase three, where the empirical study is performed, is comprehensive.

The first phase consists of a scoping review of documents containing soft-law or non-legal norms issued by organizations and institutions, including grey literature containing principles and guidelines for ethics in AI, with academic and legal sources excluded. Due to the absence of a database consisting of ethical principles for AI, (Jobin et al., 2019) invented the modified PRISMA approach, which was slightly adapted and applied in this research step. This protocol was adapted from the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework, which was developed to ensure transparent and complete reporting of systematic reviews (Liberati et al., 2009). Because of the vast number of non-academic sources containing ethical principles and guidelines for AI it was necessary to systematically review these documents and compare them with the results coming from academic and legal sources.

The second phase consists of a comprehensive review of literature on ethical principles in AI applications and settings. A systematic and structured literature review approach was applied, following the guidelines of (Brocke et al., 2009), (Webster & Watson, 2002). The approach presented by (Webster & Watson, 2002) is regarded by many

as the quasi-standard approach for a systematic literature review in information systems research (Müller-Bloch & Kranz, 2015), (Wolfswinkel et al., 2013). Additionally, (Brocke et al., 2009) extensively highlighted the importance of applying rigor in documenting the literature search process. In combination, these guidelines provide a systematic and transparent approach to performing a comprehensive literature review.

The third phase aims at the prioritization of the ethical principles retrieved in phases one and two. An online survey was conducted and evaluated based on the BWS method (Finn & Louviere, 1992).

1) Best-worst-scaling (BWS):

The best-worst scaling (BWS) method or maximum difference scaling method can be described as an extension of Random Utility Theory (Thurstone, 1927) for comparison judgements of pairs. It was first introduced by (Finn & Louviere, 1992) to measure preferences of individuals and thus allow for a comprehensive ranking of measured options (Cohen, 2009), (Marley & Louviere, 2005). It attempts to create trade-off choice situations which are easy to handle for respondents. Within multiple rounds the respondents are required to choose the best and the worst option for various choice sets, including combinations of options from a larger list of choices (Finn & Louviere, 1992)– (Marley & Louviere, 2005).

Best-worst-scaling (BWS) enables the disclosure of participant's preferences regarding the ethical principles and does not require fine-grained distinctions (Finn & Louviere, 1992)– (Cohen, 2003). It gives a more informative and accurate prioritization, in contrast to a numerical assignment technique (Cohen, 2003). The BWS-method is a straightforward and effective approach to find out how respondents rank a list of items by having to choose the top and bottom items for a given question. It aims at better understanding choice processes of participants.

Two underlying models of the best-worst choices are the MaxDiff model and the sequential choice model (Marley & Louviere, 2005), (Flynn et al., 2007). In the MaxDiff model, the differences in utilities for all possible best-worst pairs per set are calculated. The best-worst pair which provides the biggest difference in the utility or other continuum of interest will then be chosen (Finn & Louviere, 1992), (Flynn et al., 2007). In the sequential choice model, the most preferred option will be chosen first. Then, the least preferred alternative in the subset will be pointed out (Marley & Louviere, 2005). Both models provide a notion about how the B–W discrete choices of the respondents are made and provide consistent results.

Initially, the BWS method was applied by (Finn & Louviere, 1992) to investigate public concerns with respect to food safety. (Auger & Devinney, 2007) used BWS to examine people's attitude towards social and ethical issues. Further, (Cohen, 2009) conducted a survey based on BWS to investigate the customers' preferences with respect to wine attributes. In cloud computing, BWS was applied to evaluate the relative importance of certifications. However, the present study is the first one to apply the BWS to collect data on ethical principles for AI.

BWS is expected to provide better and more reliable estimates, since it is less vulnerable to potential biases, changes in means and variances (Lee et al., 2007), (Baumgartner & Steenkamp, 2001). Since the BWS approach is scale-free, respondents are not able to constantly use certain parts of the rating scale (Lee et al., 2007), (Cohen, 2003). Furthermore, people must make trade-offs and discriminate between options, because respondents need to consider only the best and respectively worst choices only.

In the field of ethical principles for AI, this is particularly helpful. People may easily tend to identify all ethical principles as most important. Moreover, the BWS method has certain advantages to be used for cross-cultural studies. It avoids translation problems of verbal scales to various languages, and it does not need to face the problems that are typical for numerical scales, e.g., other meanings of numbers in different cultures. Additionally, the data collected within the BWS framework is expected to deliver the maximum amount of information including the most preferred requirement, the least preferred requirement, individual and aggregated preference rankings of requirements (Marley & Louviere, 2005). Additionally, collected data is easy to analyze and a full rank of requirements can be easily obtained by calculating simple best-worst scores, choice probabilities or estimations based on conditional logistic regression (Cohen, 2009), (Marley & Louviere, 2005), (Louviere et al., 2015).

Out of several alternatives, the BWS method has evolved as the most useful and straightforward ranking methodology to find out about ethical principle's prioritization for AI.

In the remainder of this chapter, the applied systematic steps of the PRISMA framework, the literature review process, as well as the BWS-framework to prioritize the resulting principles will be explained in more detail.

3.2. *First phase: PRISMA framework*

The PRISMA framework (Jobin et al., 2019) was used to structurally find, screen, and analyze non-academic documents dealing with ethical principles for AI. So called soft-law efforts by governments and organizations contributing with advanced research on AI principles have risen drastically in recent years and create a need for these research efforts to be analyzed and taken seriously by the academic research community (Zhang et al., 2021). The following chapters explain the review and search process.

1) Scope of the review

The focus of the review was to find the relevant non-academic documents dealing with ethical principles for AI and its applications. The main goal is to screen the available corpus of literature, synthesize it and subsequently map and summarize it. The challenge was the necessity of including the mostly available grey literature containing guidelines for ethics in AI, e.g., government reports or policy statements by organizations. A scoping review is considered particularly suitable for heterogenous areas of research (Arksey & O'Malley, 2005), (Pham et al., 2014). The absence of a unified database for ethical principles for AI led to the creation of the adapted PRISMA framework (Jobin et al., 2019), which was adapted and subsequently applied in this research.

2) Search steps

Before data collection, the protocol was pilot tested on various search terms and calibrated to function properly. By following well working practices for grey literature retrieval, a multi-stage screening strategy was applied that guaranteed a systematic and transparent approach. First, inductive screening via the search engine Google was conducted. The second step consisted of deductive identification of relevant entities with associated websites and online collections containing ethical principles for AI. To ensure comprehensiveness, relevant documents were selected by applying three sequential search strategies.

To begin with, a manual search of six link hub webpages was performed. 43 sources were retrieved, out of which 21 were eligible (15 after removing duplicates). As a second step, a keyword-based web search on the Google search engine was performed in private-browsing mode, including logging out from all personal accounts and subsequently erasing web cookies and search history. The search was performed using the following keywords: 'AI principles', 'artificial intelligence principles', 'AI guidelines', 'artificial intelligence guidelines', 'ethical AI', 'ethical principles AI', 'ethical principles artificial intelligence' and 'ethical artificial intelligence'. Each link in the first 30 search results was followed and screened for ethical AI principles, resulting in 8 more sources after removing duplicates, as well as for articles mentioning AI principles. This led to the identification of 4 additional non-duplicate sources. The remaining results up to the 200th

hit for each Google search were followed and screened for ethical principles for AI only.

These additional 615 links resulted in 18 new non-duplicate documents. After the identification of relevant documents through the two prior described processes, citation chaining was performed to manually screen the full document texts and references of all eligible sources. Seventeen additionally relevant documents were identified in this step. The literature monitoring was continued until December 31, 2020 in parallel with the data analysis of the existing corpus of documents. Twelve new sources resulted out of this extended time frame search. The process of citation chaining was performed until theoretical saturation with the given time frame.

Based on specified inclusion/exclusion criteria, documents (including principles, guidelines, and institutional reports) included in the final synthesis were published in English or German. These were issued by institutions and organizations from both the private and the public sectors. The final list of relevant documents consisted of 43 documents that were included in the final synthesis.

3) Screening and coding

The manual screening procedure was performed based on the codes adapted from the analysis performed by (Jobin et al., 2019). A thorough content analysis of the 43 sources was conducted manually and the results were synthesized with the relevant academic literature found in phase two. During the manual coding of the sources, all relevant texts were analyzed based on whether they contained the prior defined codes. To ensure eligible results, the mentioned processes were performed independently. The results of both individually performed tasks were compared and consolidated. Mentions of ethical principles through their respective codes were counted and aggregated based on each principle.

The PRISMA analysis resulted in a list of 43 documents containing ethical principles or guidelines for AI. Combined with the results from the previous study performed by (Jobin et al., 2019). 127 documents were taken into consideration. Based on the previous study, eleven high-level ethical principles have emerged from the research that were ready to be synthesized with the relevant academic findings resulting from phase two.

3.3. Second phase: literature review

1) Scope of the review

The focus of the literature review lies on studies in the broad context of ethical principles for AI and its applications. Information Systems, Computer Science as well as studies coming from a more philosophical or respectively ethical background were taken into consideration. The available academic studies searched for have been published in 2010 or after since the topic of ethics in AI has gained wider attention only in recent times. Following the work of (Levy & Ellis, 2006) from 2006, relevant search queries as well as publication outlets and databases were identified which are going to be mentioned in the following chapters. During the review process, a neutral position was taken, and only academic results were derived. Relevant non-academic sources were taken into consideration during the subsequent PRISMA framework search.

2) Literature search and analysis

The literature search was conducted systematically and in line with the framework by (Brocke et al., 2009) and (Brendel et al., 2020). As an initial step, publications included in relevant previous literature review studies such as the one from (Jobin et al., 2019) or from (Hagendorff, 2020) provided the first set of publications. Subsequently, the list of references of previous relevant literature review publications gave the first indications of appropriate literature and keywords as proposed in (Larsen et al., 2018).

During the first review of publications, the first 30 sources for each search term were screened based on their title. The resulting list of publications for each search term was subsequently analyzed based on their abstracts. The lists for every search term were now consolidated and checked for duplicates. After duplicate removal, the content of the remaining documents has been analyzed in completion (Brendel et al., 2020), (Braccini & Federici, 2013).

The search process covered two main areas: database search and backward/forward search. The journal selection was based on the AIS World MIS Journal Ranking which covers many management journals, information systems journals and conference proceedings. For the sake of completeness, additional sources have been searched for in the following databases: IEEE Xplore, Emerald, ScienceDirect, AISeL, Springer, ACM Digital Library and Proquest. To make sure that the review results are relevant and up to date, the time frame was limited to the last ten years (2010–2020) as proposed by (Leukel et al., 2014).

The keywords used during the literature search are the main terms in the context of ethical principles for AI. The following terms and their possible combinations were used: AI principles, artificial intelligence principles, AI ethical guidelines, ethical principles AI, artificial intelligence guidelines, artificial intelligence ethical guidelines, ethical AI, ethical artificial intelligence. Subsequently, a backward and forward search was conducted based on the approach mentioned by (Webster & Watson, 2002).

Finally, out of 24.134 hits, 26 studies were temporarily classified as relevant for further search and analysis. 24.008 studies were rejected because they were either off topic or did not clearly specify ethical principles for AI. Additionally, to the 26 relevant studies from the database search, 9 sources based on the backward and forward search were added. In total, a final list of 35 studies was formed and these contributed to the ethical principles' elicitation for AI. The synthesis between the results from phases one and two served as the basis for the construction of the empirical survey. The AI-related scenarios as well as the final explanatory sentences for each principle were based on the research conducted in both steps.

3.4. Third phase: prioritization of principles

The third step of the methodology consists of the BWS approach to prioritize the requirements defined in the prior phases. Additionally, a description of the design of the online survey based on the BWS approach will be presented.

3.5. Empirical part

3.5.1. AI-related scenario elicitation

To enable respondents to express their true preferences in relation to real rather than fictional situations, we derived scenarios in which AI is already being used or can be used. In doing so, the scenarios were carefully reviewed in terms of their societal relevance. Furthermore, based on the generally accepted concepts of weak and strong AI (Bostrom & Yudkowsky, 2018), we decided to use a mixture of both forms of AI algorithms for this purpose. The chosen scenarios are generally agreed to raise several ethical issues and are based on current literature in each field.

3.5.1.1. Usage of AI in human resources (Karnouskos, 2018). One might use machine learning to find "good" employees to hire, but the meaning of "good" is not self-evident. Given the difficulty of answering this question, employers might prefer a definition that focuses on sales numbers, as these are easier for them to monitor. In this way, the hiring problem is framed as being about predicting the sales numbers of applicants, rather than simply identifying "good" employees (Passi & Barocas, 2019). Amazon's hiring algorithm has been found to be sexist (Dastin, 2022).

3.5.1.2. Usage of AI for predictive police work (Asaro, 2019), (Ferguson, 2017). Predicting when and where criminal events are more likely to occur can help law enforcement agencies appropriately plan and allocate resources (Borges et al., 2018)– (Wang & Yuan, 2019) and help citizens and travelers avoid unsafe locations (Wang & Yuan, 2019). This applies to criminal events of varying severity, including crime in cities, terrorist attacks, cybercrime, and fraud (Saidi & Zeki, 2019). Previous research has proposed several sophisticated predictive policing models (Borges et al., 2018), (Boldt et al., 2018)– (Zheng et al., 2011) and implementations such as Crime Prevention Decision Support System (CreP-DSS) (Retnowardhani & Triana, 2016) and web-based crime analysis toolkits (Calhoun et al., 2008)– (Molcho et al., 2014), for example, WebCAT (Calhoun et al., 2008), CAPER (Molcho et al., 2014), and PAVED (Godé et al., 2020).

Indeed, crime prevention strategies implemented on the basis of predictive policing can lead to a remarkable decrease in crime rates (Vandeviver & Bernasco, 2017). However, the advent of predictive policing technology should ensure that its use enables police departments to adequately serve the public, not the other way around (Hirsh, 2016).

Indeed, the CalGang database, a criminal database used to predict violent gang-related crimes, was extremely skewed and rife with errors leading to bias and injustice (Crawford, 2021). The recidivism algorithm being used in U.S. courts to predict the likelihood of recidivism was biased against blacks (Angwin et al., 2016). Over the past year, a debate has erupted over the use of data science in criminal justice, where courts rely on risk assessments to decide who should be released from prison while awaiting trial. The stakes are high: Those released on bail are more likely to keep their jobs, homes, children and spouses (Barocas & Boyd, 2017).

3.5.1.3. Usage of AI in autonomous vehicles. The American Association of Electrical and Electronic Engineering (IEEE) estimates that 75% of vehicles will be self-driving by 2040 (Feng et al. 2019). The five levels of automation in autonomous driving vehicles defined by the Society of Automotive Engineers (SAE) and the German Association of the Automotive Industry (VDA) include (Luetge, 2017): Level 1 (assisted driving) and Level 2 (partially automated driving), where the driver is assisted with steering or controlling braking and acceleration; Level 3 (highly automated driving), where the driver is assisted in monitoring the vehicle or the surroundings. Level 4 systems (fully automated driving) steer the vehicle and control braking and acceleration, while still allowing the driver to take control of the vehicle. A level 5 system (driverless driving) has complete control of the vehicle.

In this paper, we consider the highest level of automation.

Autonomous driving vehicles pose challenges to the automotive industry (Brenner et al., 2018) and to transportation policy makers (Bagloee et al., 2016). Crucial to their successful introduction are technological maturity, regulation (Lackes et al., 2020), (Sternberg et al., 2020), and road safety (Wiefel, 2021).

Indeed, passengers' concerns relate to the physical and psychological risks of riding in an autonomous driving vehicle. Owners' concerns relate to liability when acquiring an autonomous driving vehicle. They are also unsure who to allow to use their vehicle and are contemplating legal consequences if something happens to go wrong (David et al., 2019). Trust plays an integral role in forming individual attitudes toward autonomous driving vehicles (Lackes et al., 2020), intention to use them (Bruckes et al., 2019), and their adoption (Lackes et al., 2020).

Companies involved in autonomous driving vehicles are aware of and committed to ethical aspects (Martinho et al., 2021).

3.5.1.4. Usage of AI in automated delivery robots for hospitals. Demand for professional caregivers far exceeds supply. The aging population worldwide will further increase the demand for help with elder care. Artificial intelligence (AI), particularly through the use of robots, could help, especially in caring for the elderly, increasing their independence and potentially reducing the harmful effects of social isolation (2019 Edelman AI survey, 2019).

3.5.2. Structure and design of the survey

The online survey consisted of the following parts.

3.5.2.1. Introduction. The introductory part is intended to create a common understanding of the survey's purpose and the topic of AI in general. Participants were given a brief explanation of how AI algorithms work to ensure that each participant was able to understand the importance of ethical principles for AI.

3.5.2.2. AI scenario introduction. Next, participants were randomly assigned one of four AI application scenarios and presented with a short textual as well as graphical explanation of the application of AI in the given scenario.

3.5.2.3. The survey: BWS ranking. Participants were briefly told how to answer the best or worst choice question. They were instructed to always select one most and one least preferred principle per set of two questions. To avoid errors (e.g., missing answers or indicating more than one requirement as most or least preferred), the online survey was programmed so that respondents could not select more than one answer for each question. The survey portion consisted of 22 questions related to 11 question sets. After respondents completed all best/worst answers, they were asked to freely express their views on ethical principles for AI.

3.5.3. Preparation

3.5.3.1. Pre-testing and calibration. To arrive at the final version displayed to respondents, two separate pre-tests were performed with a small section of 8 participants each that did not take part in the public survey. Each scenario was pre-tested with 4 respondents. The resulting feedback was then collected in one-on-one interviews and summarized accordingly. To receive structured feedback, the test respondents were asked to fill out a short questionnaire about comprehension of principles as well as scenarios and demographical questions.

3.5.3.2. Principle items. The preparation of the final policy statements presented to survey participants had to be done carefully and considerately due to the complexity of explaining ethical principles in a short and concise manner. Based on the ethical principles from the first part of the methodology, related original items from the IS literature were analyzed and then categorized based on related constructs. Items were clustered for each underlying ethical principle. Then, the list of items for each cluster was consistently narrowed down in a stepwise process to obtain a list of candidate items. These candidate items were then reanalyzed and compared with each other to obtain the final list of items for the empirical survey.

Following the recommendations (Rupp et al., 2005), (Sommerville & Sawyer, 1997), the principle definitions were formulated informally and in a natural language to be easily understandable.

The elicitation of the final items used for the survey is exemplified for the ethical principles of transparency and responsibility in the following.

3.5.3.3. Transparency.

1) Cluster items (exemplary):
   - The reasoning process of an AI system should be transparent.
   - The criteria for decision making of an AI system should be transparent.
   - The reasoning process of an AI system should be understandable.
   - The input data to an AI system should be transparent, valid, correctly labeled and interpreted.

- The data transformations and hyper parameters used in training the AI system should be transparent.

2) Candidate item
   - The criteria for decision making of an AI system should be transparent and understandable. An AI system should provide transparency of current and future usage of personal data and provide users with control over their data.

3) Final item
   - The decision-making process of an AI system should be transparent and understandable.

3.5.3.4. Responsibility.

1) Cluster items (exemplary):
   - There should be an audit process and governance of an AI system to identify and mitigate errors, assess the impact of the system on its environment and to address unexpected results.
   - An AI system should be reliable.
   - Developers of an AI system need to understand ethical norms and best practices and should be aware of their ethical responsibility.
   - There should be clear responsibility and accountability for all roles in the design and implementation lifecycle of an AI system.
   - An advanced AI system should be treated as a moral agent, and it should adopt responsibility for its acts.
   - There should be human control over the development and the decision of an AI system.

2) Candidate item:
   - There should be clear responsibility and accountability for all roles in the design and implementation lifecycle of an AI system. Furthermore, there should be human control over the development and the decision of an AI system.

3) Final item:
   - Responsibility for an AI system should be borne by its vendors and implementers, and they should bear the consequences in case of a mistake.

3.5.3.5. Balanced incomplete block design (BIBD). During the main part of the survey, the respondents faced the same best-worst choices, but were introduced to various scenarios, as mentioned before. The considered principles were organized into choice sets according to the balanced incomplete block design (BIBD) that is commonly used to design BWS experiments. Following the textbook on experimental design by (Cochran & Cox, 1992), BIBD based on eleven choice sets was applied (11 ethical principles resulted from phases one and two). In line with best practice, every choice set consisted of five items and each of them appeared only twice with another one.

To account for order effects, the order appearance of the principles in the framing part, within the blocks, as well as of the blocks themselves was randomized.

A balanced incomplete block design is a design with v treatment labels, each occurring r times, and with bk experimental units grouped into b blocks of size $k < v$ in such a way that the units within a block are alike and units in blocks are substantially different.

3.5.3.6. The plan of the design satisfies the following conditions.

(i) Each treatment label appears either once or not at all in a block (that is, the design is binary).

(ii) Each pair of labels appears together in λ blocks, where λ is a fixed integer.

3.5.3.7. There are three necessary conditions for the existence of a balanced incomplete block design.

1. $vr = bk$,
2. $r(k-1) = \lambda(v-1)$,
3. $b \geq v$.

The order of the BIBD is explained in Table 1.

Best-Worst-Scaling has been applied because it provides consistent and high-quality results for data annotation and has been shown to outperform the widely used method of rating scales (Kiritchenko & Mohammad, 2017). Further, it has been successfully applied to assess user's ethical beliefs (Auger et al., 2007).

4. Results

This chapter describes the results of the elicitation and prioritization process conducted in this study. The literature-driven elicitation delivered a total set of eleven principles for AI regarding ethical aspects, e.g., transparency, fairness, or privacy. The empirical analysis of the survey data delivered interesting insights about the ethical priorities of how to incorporate ethical behavior into algorithms of respondents with respect to various AI-related scenarios.

4.1. Literature review results

The performed PRISMA analysis resulted in a list of 43 documents containing ethical principles or guidelines for AI published in 2019 and 2020.

Combined with the results from the previous study performed by (Jobin et al., 2019), 127 documents were taken into consideration. Based on the main study, eleven high-level ethical principles combining various ethical codes have emerged.

Each document has been coded subsequently based on whether it explicitly mentions and explains each ethical principle. The results of the coding process are shown in Table 2.

The individual ethical principles were mentioned with varying frequency in the two studies. In the current study, each principle was mentioned more frequently overall. Compared to the 2019 literature base, trust was mentioned 46% more often in the literature base of the current study, followed by sustainability with 41% growth. The middle field is composed of privacy (25%), non-maleficence (22%), freedom and autonomy (18%), dignity (18%), responsibility (17%), beneficence (16%). Transparency (8%), solidarity (7%), and justice and fairness (5%) show a rather slight increase in the percentage of mentions. Interestingly, there was no principle commonly mentioned in all the documents analyzed, neither in the main study, nor in the performed research of this work.

Transparency was mentioned most often in both individually

Table 1
Balanced incomplete block design (BIBD).

Block/Question	Ethical Principles				
1	1	2	3	5	8
2	8	9	10	1	4
3	4	5	6	8	11
4	10	11	1	3	6
5	2	3	4	6	9
6	9	10	11	2	5
7	11	1	2	4	7
8	6	7	8	10	2
9	7	8	9	11	3
10	5	6	7	9	1
11	3	4	5	7	10

Table 2
Ethical Principles based on (Jobin et al., 2019).

Ethical principle	Jobin et al. (2019) (Jobin et al., 2019)	This research	Aggregated	Exemplary References
1. Transparency	87%	95%	90%	(Li et al., 2019), (Balakrishnan et al., 2019)– (Sokol, 2019)
2. Justice and fairness	81%	86%	83%	(Balakrishnan et al., 2019), (Rothenberger et al., 2019)– (Yapo & Weiss, 2018)
3. Non-maleficence	71%	93%	79%	(Gómez-González et al., 2020), (Rothenberger et al., 2019), (Siau & Wang, 2020), (LaBrie & Steinke, 2019), (Dobbe et al., 2020)– (Maas, 2018)
4. Responsibility	71%	88%	77%	(Feng et al., 2019), (Balakrishnan et al., 2019), (Gómez-González et al., 2020), (Rothenberger et al., 2019), (Siau & Wang, 2020), (Seymour, 2018)
5. Privacy	56%	81%	65%	(Li et al., 2019), (Balakrishnan et al., 2019), (Gómez-González et al., 2020), (Rothenberger et al., 2019), (Siau & Wang, 2020), (Li & Zhang, 2017), (Weibel et al., 2017)
6. Beneficence	49%	65%	54%	(Hooker & Kim, 2018), (Siau & Wang, 2020)
7. Freedom & autonomy	40%	58%	46%	(Hooker & Kim, 2018), (Siau & Wang, 2020), (Susser, 2019)
8. Trust	33%	79%	49%	(Gómez-González et al., 2020), (Siau & Wang, 2020), (Li & Zhang, 2017), (Susser, 2019), (Giattino et al., 2019)
9. Sustainability	17%	58%	31%	(Li et al., 2019), (Rothenberger et al., 2019)
10. Dignity	15%	33%	21%	(Rothenberger et al., 2019), (Siau & Wang, 2020)
11. Solidarity	7%	14%	9%	(Li et al., 2019), (Gómez-González et al., 2020), (Siau & Wang, 2020), (Cruz, 2019)

performed studies, in 87% and 95% of sources, respectively. The overall results clearly indicate that transparency, justice and fairness, non-maleficence, responsibility, and privacy were mentioned most often in the overall corpus of documents, with over 55% and 80% of mentions, respectively, in the sources of the two studies. We decided to concentrate on them in the present study.

The synthesis of the 35 additionally identified academic literature review results with the above-mentioned sources from the applied PRISMA approach needed to be done in a strictly structured and concise way. As pointed out in the methodology section, IS literature-based items were searched for according to the codes from the research of (Jobin et al., 2019). It took 9 individual steps to arrive at the final list of principle items representing the underlying ethical guidelines found prior.

In step three, the constructed items have been consistently categorized into seven clusters, namely transparency, justice & fairness, responsibility, non-maleficence, privacy, dignity, and trust. The subsequent two steps narrowed these seven clusters down into 5 mutually exclusive ones. They were based on the most important high-level principles groups being mentioned among the 127 documents analyzed according to the PRISMA framework.

The final clusters identified based on both reviews were transparency, justice and fairness, responsibility, non-maleficence, and privacy. In the remaining steps, the items in each cluster have been refined and rephrased, as shown in Table 3.

Table 3
Item overview.

Item for the survey	Ethical principle/ cluster
An AI system should respect human safety; in particular, it should not harm human beings, or watch human beings suffer danger and ignore it.	Non-maleficence
An AI system should be secure; in particular, attacks on the system and unauthorized system use should be prevented.	Non-maleficence
The decision of an AI system should be fair, unbiased and free of discrimination against individuals and groups.	Justice & Fairness
There should be audit processes applied for identifying and mitigating limitations and errors of an AI system.	Responsibility
There should be clear traceability of errors and accountability for the lifecycle of an AI system.	Transparency
An AI system should respect human privacy; in particular, humans should be able to control their personal data used in the system, and misuse of personal data should be prevented.	Privacy
The data used for training the model and the decision-making process of an AI system should be correct and complete.	Transparency
A human being should have the control over executing the decision of an AI system.	Responsibility
The decision-making process of an AI system should be transparent and understandable.	Transparency
Responsibility for an AI system should be borne by its vendors and implementers, and they should bear the consequences in case of a mistake.	Responsibility
The fact that an AI system is applied should be made clear to the end user.	Transparency

4.2. Empirical prioritization results

This section presents the findings of the survey data analysis on the respondents' prioritization of ethical principles for AI applications in life.

4.2.1. Data collection and sample

4.2.1.1. Demographical data. The survey data was collected in Germany during two separate periods of four months and respectively 3 months from January 2021 until the end of April 2021, as well as from September until November 2021. The data is depicted in Table 4. Respondents who participated in the survey were recruited via social networks such as LinkedIn and Facebook, as well as through university networks mainly at Humboldt University of Berlin. We were able to recruit 457 people to start the survey, and 225 participants completed it in full.

First, respondents were given a clear and concise overview the research topic and its importance, among other things, to encourage them to participate in the survey. Next, participants were randomly assigned to one of four possible scenarios. Scenario one deals with AI in human resources. Scenario two is about predictive policing. Scenario three addresses autonomous vehicles and scenario four delivery robots in hospitals. Most participants that finished the survey were assigned to scenario three (30.67%), while the remaining participants were similarly equally distributed across the remaining three scenarios one, two, and four (20.44%, 25.33%, and 23.56, respectively).

Of the 225 participants, 52% were male and 41% female. Six respondents preferred not to indicate their gender, 2% of individuals indicated "other" as their gender, and 5% chose not to indicate their gender. Most of the participants (71%) are between 18 and 39 years old, which is not surprising given that the study is mainly aimed at students and people of the same age.

Most respondents (56%) named a university degree (bachelor and/or master's degree) as their highest educational qualification. Another third (32%) of participants have a high school diploma and around 7% do have a doctoral degree or habilitation. In line with this, most participants are students in their current employment status (57%),

Table 4
Demographics.

Scenario	1	2	3	4	Total
Number	46	57	69	53	225
Gender					
Male	48%	51%	61%	45%	52%
Female	48%	44%	32%	43%	41%
N/A	4%	2%	6%	7%	5%
Other	0%	3%	1%	4%	2%
Age					
0–17	0%	0%	2%	0%	1%
18–24	37%	46%	36%	32%	38%
25–39	39%	37%	36%	43%	39%
40–59	13%	12%	16%	17%	15%
60–79	7%	2%	3%	4%	4%
N/A	4%	2%	7%	4%	4%
Highest educational degree					
Secondary school diploma	2%	2%	3%	2%	2%
High school diploma	37%	28%	33%	28%	32%
Master's degree	26%	33%	22%	24%	26%
Bachelor's degree	21%	32%	36%	30%	30%
Doctoral degree (Ph.D.)	4%	2%	3%	7%	4%
Habilitation	4%	2%	3%	4%	3%
N/A	4%	2%	0%	4%	2%
Current employment status					
Student	46%	67%	62%	49%	57%
Employee	35%	21%	15%	23%	22%
Official	6%	2%	7%	13%	7%
Unemployed	0%	4%	4%	0%	2%
N/A	4%	2%	6%	4%	4%
Self-employed	2%	2%	0%	6%	2%
Retired	0%	2%	1%	2%	1%
Other	0%	2%	1%	2%	1%
CEO	7%	0%	3%	0%	2%
Apprentice	0%	0%	0%	2%	0%
Data Science experience					
Less than 1 year	71%	68%	58%	59%	64%
1 year–5 years	24%	21%	28%	26%	25%
6 years–10 years	0%	2%	3%	2%	2%
11 years–15 years	0%	0%	1%	6%	2%
More than 15 years	0%	5%	6%	2%	4%
N/A	4%	3%	4%	5%	4%

followed by around 22% employees.

In terms of experience with data science or information systems, most respondents (64%) had less than one year of relevant experience, while 27% reported having between one and ten years of experience. 25% had between one and five years of relevant experience.

4.2.1.2. Rank orders. The BWS method offers several procedures to analyze respondents' preferences and provides a tool to obtain a rank order of principles in terms of their relative importance (Cohen, 2009), (Marley & Louviere, 2005). Even though the procedure based on counting provides results that are a close approximation of a conditional logistic regression (Marley & Louviere, 2005), both methods are applied for the sake of completeness. The second method is chosen to verify whether it delivers consistent results in comparison with the main method. In addition, the square root method is applied to allow for further analysis. Finally, the level of agreement among the respondents on the obtained rankings is calculated. For the numerical and statistical analysis of the survey data, an Excel spreadsheet has been applied.

4.2.1.3. Overall rankings based on counting analysis. In contrast to a conditional logistic regression, the counting procedure applied here is less complicated and easier to perform no matter which MNL model (e.g., MaxDiff or sequential model) better fits the data (Marley & Louviere, 2005). By counting the number of times, a principle was chosen to be the most (best) and least (worst) important, based on the proposed methodology by (Finn & Louviere, 1992), the M-L scores are calculated at the aggregate level by subtracting the number of times a principle was identified as least important from the number of times it was chosen as most important. The average M-L scores are calculated for every requirement by dividing the aggregate M-L scores by the number of overall respondents (225) and the number of times every single principle appeared in the choice sets in total (5). Finally, the M-L scores and the average M-L scores allow to create a rank order of the ethical principles (Cohen, 2009). The obtained first results given in Table 5 provide first insights on the ethical principles priorities of the survey respondents.

The first two columns include the underlying ethical principle or cluster and the item used in the survey. Columns three and four show the number of times an item is chosen as most or least important. The item containing human safety (non-maleficence) is most frequently chosen as most important, while the item about transparency to the end-user (transparency) most often identified as least important. The positive values of the M-L scores (column 5) indicate that the considered requirement was chosen more frequently as most than least important.

The negative M-L scores indicate that these requirements are more often marked as least than most important. The average M-L scores (column 6) take on values between −1 and 1 and provide equivalent results to the simple M-L scores. Thus, a higher average M-L score level indicates higher level of importance (Cohen, 2009). Based on the M-L scores and the average M-L scores, the requirements are assigned to ranks from 1 to 11 (last column). While rank 1 is most important, rank 11 is least important at the aggregate level.

The overall ranking clearly indicates that respondents rate high-level ethical principles such as non-maleficence and justice & fairness as most important. Human safety, non-bias and the prevention of unauthorized usage are most important to most people participating in the survey. Transparency and responsibility are of great importance to respondents as well, reflected in items describing accountability in case of a mistake and traceability of errors in such cases. Privacy in terms of personal data usage is seen as important as well. Items describing the internal aspects behind an applied AI, as well as more theoretical descriptions of how the underlying algorithms work, receive less ratings as important by the survey participants. Overall, it can be concluded that those items describing actual applications and practical usage of AI have been rated as most important more often than those items describing the more theoretical side of AI.

4.2.1.4. Rankings per scenario based on counting analysis. To analyze the data further, the applied counting method has been broken down into the data for each of the four scenarios. For clarification, the scenario contents will be repeated. Scenario one deals with human resources, scenario two focuses on predictive policing, while scenarios three and four deal with autonomous vehicles and hospital robots respectively. As for the items rated as most important, there is no substantial difference to the overall ranking (for comparative reasons displayed in the last column). The same counts for the items rated as least important.

Overall, both ends (most and least) of the scale have been consistently chosen as most or least important overall, regardless of the various applied scenarios. Items ranked as important overall, such as personal privacy, differ among the individual scenarios. Respondents rated it as less important in the predictive policing scenario than in the overall results. Respectively, privacy has been rated as far more important in the human resources scenario than in the overall results. Respondents have shown a deeper understanding of the functionality of AI algorithms regarding the item describing that training data should be correct and complete by rating it significantly higher in the predictive policing scenario than overall. Interestingly, the individual results also show that it does not seem to bother respondents whether they know if AI is applied or not, regardless of the four use cases (see Table 6).

A square root analysis is applied to allow for further insights into respondents' preferences with respect to ethical principles for AI (see Table 7).

$$\frac{M}{(L+0.5)}$$

Table 5
M-L ranking.

Ethical principle/ Cluster	Item	Most counts	Least counts	M-L score	Avg. M-L score	Rank
Responsibility	Responsibility for an AI system should be borne by its vendors and implementers, and they should bear the consequences in case of a mistake.	118	286	−168	−0,15	**10**
Transparency	The data used for training the model and the decision-making process of an AI system should be correct and complete.	201	209	−8	−0,01	**6**
Transparency	The decision-making process of an AI system should be transparent and understandable.	140	291	−151	−0,13	**9**
Transparency	The fact that an AI system is applied should be made clear to the end user.	64	652	−588	−0,52	**11**
Responsibility	There should be audit processes applied for identifying and mitigating limitations and errors of an AI system.	191	150	41	0,04	**4**
Non-maleficence	An AI system should be secure; in particular, attacks on the system and unauthorized system use should be prevented.	315	44	271	0,24	**2**
Non-maleficence	An AI system should respect human safety; in particular, it should not harm human beings, or watch human beings suffer danger and ignore it.	373	56	317	0,28	**1**
Justice & Fairness	The decision of an AI system should be fair, unbiased and free of discrimination against different individuals and groups.	261	141	120	0,11	**3**
Responsibility	A human being should have the control over executing the decision of an AI system.	224	255	−31	−0,03	**8**
Privacy	An AI system should respect human privacy; in particular, humans should be able to control their personal data used in the system, and misuse of personal data should be prevented.	134	143	−9	−0,01	**7**
Transparency	There should be clear traceability of errors and accountability for the lifecycle of an AI system.	186	155	31	0,03	**5**

Table 6
Ranking per each scenario.

Ethical principle/ Cluster	Item	Ranking (Scenario 1)	Ranking (Scenario 2)	Ranking (Scenario 3)	Ranking (Scenario 4)	Ranking (Overall)
Responsibility	Responsibility for an AI system should be borne by its vendors and implementers, and they should bear the consequences in case of a mistake.	10	10	7	9	10
Transparency	The data used for training the model and the decision-making process of an AI system should be correct and complete.	9	6	5	6	6
Transparency	The decision-making process of an AI system should be transparent and understandable.	6	9	9	10	9
Transparency	The fact that an AI system is applied should be made clear to the end user.	11	11	11	11	11
Responsibility	There should be audit processes applied for identifying and mitigating limitations and errors of an AI system.	7	5	4	4	4
Non-maleficence	An AI system should be secure; in particular, attacks on the system and unauthorized system use should be prevented.	1	2	2	2	1
Non-maleficence	An AI system should respect human safety; in particular, it should not harm human beings, or watch human beings suffer danger and ignore it.	2	1	1	1	2
Justice & Fairness	The decision of an AI system should be fair, unbiased and free of discrimination against different individuals and groups.	3	3	6	3	3
Responsibility	A human being should have the control over executing the decision of an AI system.	5	4	10	5	8
Privacy	An AI system should respect human privacy; in particular, humans should be able to control their personal data used in the system, and misuse of personal data should be prevented.	4	8	8	7	7
Transparency	There should be clear traceability of errors and accountability for the lifecycle of an AI system.	8	7	3	8	5

The calculated square roots estimate the choice probability of each principle in per cent and put them in relation to the most important one (i.e., with the highest sq root) (Cohen et al., 2009).

Table 8 compares the square root for each scenario with the overall result. In total, the results of the square root analysis support the counting analysis results. The rankings are similar with only the most important two items being switched in the square root results. As for each individual scenario the findings support the counting analysis' outcome. Table 9 shows the overall ranking.

5. Discussion

This research applied a systematic reviewing procedure and subsequent prioritization of AI ethical principles by means of an empirical study. It contributes to the evaluation of how AI should be designed and applied among four areas of societal life to serve its users' needs.

The systematic approach has been conducted according to a rigorous procedure that is documented and fully reproducible for further analysis in the same or other contexts (cultural background, different scenarios).

The first objective of this work was to apply a structured systematic literature review for both non-academic and academic resources to identify relevant ethical guidelines or principles for AI applications. So called soft-law efforts by governments and organizations contributing with advanced research on AI principles have risen drastically in recent years and created a need for these research efforts to be analyzed and taken seriously by the academic research community (Liberati et al., 2009). By following the adapted PRISMA approach introduced by (Jobin et al., 2019), non-academic resources have been collected, analyzed, coded, and synthesized with the academic findings resulting from the second phase of this research as well as the findings from the previous performed study in 2019. The overall results clearly indicate that transparency, justice and fairness, non-maleficence, responsibility, and privacy were mentioned most often in the overall corpus of documents, with over 55% and 80% of mentions, respectively, in the sources of the two studies.

The PRISMA research method resulted in a comprehensive list of 43 documents that were fully analyzed and coded based on eleven ethical principles clusters that emerged from the previous study.

Table 7
Rank (BWS).

Ethical principle/ Cluster	Item	Sqrt (M/ L)	Relative importance	Rank (Sqrt)	Rank (BWS)
Responsibility	Responsibility for an AI system should be borne by its vendors and implementers, and they should bear the consequences in case of a mistake.	0,64	24%	10	10
Transparency	The data used for training the model and the decision-making process of an AI system should be correct and complete.	0,98	37%	6	6
Transparency	The decision-making process of an AI system should be transparent and understandable.	0,69	26%	9	9
Transparency	The fact that an AI system is applied should be made clear to the end user.	0,31	12%	11	11
Responsibility	There should be audit processes applied for identifying and mitigating limitations and errors of an AI system.	1,13	42%	4	4
Non-maleficence	An AI system should be secure; in particular, attacks on the system and unauthorized system use should be prevented.	2,66	100%	1	2
Non-maleficence	An AI system should respect human safety; in particular, it should not harm human beings, or watch human beings suffer danger and ignore it.	2,57	97%	2	1
Justice & Fairness	The decision of an AI system should be fair, unbiased and free of discrimination against different individuals and groups.	1,36	51%	3	3
Responsibility	A human being should have the control over executing the decision of an AI system.	0,94	35%	8	8
Privacy	An AI system should respect human privacy; in particular, humans should be able to control their personal data used in the system, and misuse of personal data should be prevented.	0,97	36%	7	7
Transparency	There should be clear traceability of errors and accountability for the lifecycle of an AI system.	1,09	41%	5	5

Subsequently, a structured literature review combining the approaches of (Brocke et al., 2009) and (Webster & Watson, 2002) has been conducted separately to synthesize the softlaw findings with academic resources. 35 publications have been retrieved and fully analyzed based on the codes from (Jobin et al., 2019). The body of literature was then used to create items for the empirical survey, as well as provide insights into the construction of AI application scenarios. The item construction has been a difficult and time intensive task due to the complexity of explaining ethical principles in a short, easy-to-understand, and concise manner. The eleven ethical principles have been narrowed down into the five most important overall clusters based on the results from the literature review. Based on the clusters, eleven items (i.e., descriptions) have been constructed that comprise the most important ethical principles and were used subsequently to identify participants' priorities about ethical AI applications.

(Whittlestone et al., 2019) pointed out that principles need to be ultimately formalized through certain standards or regulations. These can only be effective if clear guidance exists on how underlying values should be prioritized in different scenarios or cases. In our work, we set out to explore stakeholders' priorities in detail. Principles alone are naturally limited because they are likely to be interpreted differently by different groups of people or stakeholders. Additionally, they are highly general and theoretical. In practice, they come into conflict with one another. Prioritizing ethical values in different scenarios is an important step among others, such as clearly highlighting value tensions and finding ways how to resolve them (Whittlestone et al., 2019). Sanderson et al. (Sanderson et al., 2023) highlighted tensions between different ethical aspects as they interact with one another. Different to (Whittlestone et al., 2019), they focus on three-sided interactions instead of the more common two-sided point of view to look at it. Typical AI ethical principles have a common set of underlying aspects and to operationalize AI ethics, it is important to study the interactions between these aspects and prioritize them. With our work, we hope to contribute to raising awareness about the interaction of ethical aspects and foster the operationalization of AI ethics. Jakesch et al. (Jakesch et al., 2022) explain that different ethical guidelines for responsible AI focus on a specific set of values, while little is known about the priorities of a representative public. They conducted an empirical study similar to our survey but focused on three different groups and highlighted the differences between the public and AI practitioners. More diverse ethical judgement is needed because priorities of different groups differ greatly. It is important to pay attention to who gets to define what ethical AI means. Our results aim to increase the sensitivity for context when trying to define a set of AI ethical principles.

5.1. Prioritization findings

In the prioritization part it has been differentiated between principles' items that were selected as most important more frequently than as least important and those principles that were chosen as least important more frequently than as most important.

On the principles description level, the first group contained non-maleficence in terms of human safety and security of the AI system,

Table 8
Relative choice probabilities based on sqrt per each scenario.

Ethical principle/ Cluster	Item	Sqrt (Sc. 1)	Sqrt (Sc. 2)	Sqrt (Sc. 3)	Sqrt (Sc. 4)	Sqrt (Overall)	Rank Overall (Sqrt)	Rank (BWS)
Responsibility	Responsibility for an AI system should be borne by its vendors and implementers, and they should bear the consequences in case of a mistake.	0,47	0,61	0,84	0,65	0,64	10	10
Transparency	The data used for training the model and the decision-making process of an AI system should be correct and complete.	0,62	1,18	1,03	1,05	0,98	6	6
Transparency	The decision-making process of an AI system should be transparent and understandable.	0,98	0,65	0,65	0,54	0,69	9	9
Transparency	The fact that an AI system is applied should be made clear to the end user.	0,58	0,21	0,27	0,21	0,31	11	11
Responsibility	There should be audit processes applied for identifying and mitigating limitations and errors of an AI system.	0,92	1,23	1,14	1,16	1,13	4	4
Non-maleficence	An AI system should be secure; in particular, attacks on the system and unauthorized system use should be prevented.	2,04	2,34	3,06	2,95	2,66	1	2
Non-maleficence	An AI system should respect human safety; in particular, it should not harm human beings, or watch human beings suffer danger and ignore it.	1,61	2,60	2,91	3,37	2,57	2	1
Justice & Fairness	The decision of an AI system should be fair, unbiased and free of discrimination against different individuals and groups.	1,39	1,51	0,99	1,81	1,36	3	3
Responsibility	A human being should have the control over executing the decision of an AI system.	1,00	1,16	0,63	1,09	0,94	8	8
Privacy	An AI system should respect human privacy; in particular, humans should be able to control their personal data used in the system, and misuse of personal data should be prevented.	1,68	0,90	0,77	0,97	0,97	7	7
Transparency	There should be clear traceability of errors and accountability for the lifecycle of an AI system.	0,89	1,10	1,32	0,94	1,09	5	5

justice, and fairness (no bias or discrimination), responsibility, transparency (accountability, traceability), and personal data privacy.

The second group contained correct and complete training data (transparency), human control (responsibility), understandability of the underlying algorithm, responsibility in case of a mistake, and whether it is made clear to the end user if AI is being applied.

The overall results of the survey clearly show that respondents view the safety of AI applications ("*An AI system should respect human safety; in particular, it should not harm human beings, or watch human beings suffer danger and ignore it.*" and "*An AI system should be secure; in particular, attacks on the system and unauthorized system use should be prevented.*") above all other ethical principles. The human bias towards the safety of AI systems is also evident in other studies. In one study, the majority of respondents favored cars with utilitarian controls (with the maximized aggregate total benefit, i.e., sum of the well-being of all concerned) for all other road users if possible, but they themselves would prefer a vehicle in which passengers are protected at all costs (Millard-Ball, 2018). In another series of studies, respondents favored autonomous driving vehicles that sacrifice their passengers for the greater good and that others should buy, but they themselves would ride in a vehicle in which passengers are safe at all costs (Bonnefon et al., 2016).

Fairness in terms of not being prone to potential bias ("*The decision of an AI system should be fair, unbiased and free of discrimination against individuals and groups.*") is ranked third for scenarios involving predictive policing and automated delivery robots for hospitals, fourth for the human resources scenario, and sixth for the autonomous vehicles scenario. In line with our findings, research in the context of the autonomous vehicles scenario shows that reaction time (Wolff et al., 2019), emotional state (van Berkel et al., 2022) including perceived stress (Wolff et al., 2019), moral preferences, individual characteristics (Awad et al., 2018), (Rhim et al., 2021), and culture (van Berkel et al., 2022), (Awad et al., 2018) can be crucial in such decisions.

Clear traceability and responsibility are rated as important as well. Participants have clearly focused on the outcomes of AI application and put the direct implications for human beings at the forefront of their prioritization. This can also be validated by analysing the results for each scenario independently. In general, they resemble the overall findings, but there are a few interesting insights that were noticed. In the predictive policing scenario, privacy, and the usage of correct data for training the algorithm were highlighted as far more important than overall. A general understanding of how AI algorithms work can be expected since respondents clearly prioritized the usage of correct and complete training data in the predictive policing scenario. This supports the overall prioritization of non-maleficence (human safety). In the AI introductory part of the survey, participants were informed about how AI algorithms learn and improve their performance in an easy-to-understand description. The individual results for the human resources scenario show that privacy and the correct usage of personal data have been rated as more important than overall, which makes sense in the context.

The principle of responsibility in case of a mistake ("*Responsibility for an AI system should be borne by its vendors and implementers, and they should bear the consequences in case of a mistake.*") landed in one of the back places. In an empirically formed ranking of ethical guidelines for AI, responsibility got the most votes in terms of relevance, leaving privacy, transparency, robustness, minimizing bias, and usefulness of AI in the background. The participating experts and the participants in the online survey found it most important to have someone with whom responsibility for actions of an AI system and especially for its consequences should be sought (Rothenberger et al., 2019). The slight contradiction may be due to the lack of consensus on the distribution of responsibility (Martin, 2017), (Rochel & Evéquoz, 2021), as noted in a recent literature review (Selter et al., 2022).

The overall results indicate that participants do not place greater importance on whether they are informed in the given scenarios that they are dealing with an AI application and not a human ("*The fact that an AI system is applied should be made clear to the end user.*"). The corresponding item was consistently rated as least important in all four scenarios. This may well be positive. It has been shown that moral judgment about the decision to donate a kidney to a patient can be influenced by AI feedback, even when that feedback is generated completely randomly (Chan et al., 2020). However, there is also the question of whether participants were able to understand the intent behind this item. This is also worth mentioning in the context that transparency was mentioned most frequently in both individual literature reviews.

The four scenarios were carefully chosen based on heterogeneity, but a close connection to the human being involved during the application (hospital robot for example). Interestingly, the survey results indicate that transparency and understandability regarding the decision-making process of the respective AI algorithm are not very important to the

Table 9
Overall ranking.

Ethical principle/ Cluster	Item/Scenario	1	2	3	4	All
Non-maleficence	An AI system should respect human safety; in particular, it should not harm human beings, or watch human beings suffer danger and ignore it.	2	1	1	1	1
Non-maleficence	An AI system should be secure; in particular, attacks on the system and unauthorized system use should be prevented.	1	2	2	2	2
Justice & Fairness	The decision of an AI system should be fair, unbiased and free of discrimination against individuals and groups.	3	3	6	3	3
Responsibility	There should be audit processes applied for identifying and mitigating limitations and errors of an AI system.	7	5	4	4	4
Transparency	There should be clear traceability of errors and accountability for the lifecycle of an AI system.	8	7	3	8	5
Transparency	The data used for training the model and the decision-making process of an AI system should be correct and complete.	9	6	5	6	6
Privacy	An AI system should respect human privacy; in particular, humans should be able to control their personal data used in the system, and misuse of personal data should be prevented.	4	8	8	7	7
Responsibility	A human being should have the control over executing the decision of an AI system.	5	4	10	5	8
Transparency	The decision-making process of an AI system should be transparent and understandable.	6	9	9	10	9
Responsibility	Responsibility for an AI system should be borne by its vendors and implementers, and they should bear the consequences in case of a mistake.	10	10	7	9	10
Transparency	The fact that an AI system is applied should be made clear to the end user.	11	11	11	11	11

overall group of participants. Furthermore, the overall results show that whether vendors and implementers of AI should bear the consequences in case of a mistake was almost consistently viewed as not important. Only in the scenario dealing with autonomous vehicles has it been rated as more important compared to the overall results.

Overall, both ends (most and least) of the scale have been consistently chosen as most or least important overall, regardless of the various applied scenarios. Items ranked as important overall, such as personal privacy, differ among the individual scenarios. Respondents rated it as less important in the predictive policing scenario than in the overall results. Respectively, privacy has been rated as far more important in the human resources scenario than in the overall results. Respondents have shown a deeper understanding of the functionality of AI algorithms regarding the item describing that training data should be correct and complete by rating it significantly higher in the predictive policing scenario than overall. Interestingly, the individual results also show that it does not seem to bother respondents whether they know if AI is applied or not, regardless of the four use cases.

5.2. Implications for research and practice

This work has several implications for research and practice. For research, a systematic and structured collection and ranking of ethical principles for AI applications has been conducted by means of prioritizing them from individual's perspective. The need for this stems from the current existence of high-level theoretical ethical principles without much practical validity (Hickok, 2021). According to (Paradice et al., 2018) and (Whittlestone et al., 2019), it is necessary to move from vague high-level ethics to more practical research, which this work set out to provide.

Our results indicate that the prioritization of ethical principles in terms of AI applications is rather homogenous among various practical use cases (four scenarios tested) at both ends of the spectrum (most important and least important). Regardless of the application, items consisting of human safety and/or security of the system have been consistently prioritized as most important. The same counts for justice and fairness in terms of eliminating bias and discrimination from applied AI. Being informed about whether a person is dealing with AI instead of a human being (e.g., human resources application) has been consistently prioritized as least important among the scenarios. Respondents have not placed greater importance on the understanding of how the algorithm performs its decision-making. Ethical principles descriptions on personal data usage (privacy), traceability, accountability in case of a mistake, and the usage of correct and complete data for training the algorithm have been prioritized more heterogeneously across all four scenarios. These principles' rating has been influenced greater by the respective scenario. This research combined non-academic resources (such as political or institutional guidelines) and academic works and synthesized the individual findings. Clarity seems to exist about what ethical principles are most and least important, regardless of most application scenarios, as well as academic or non-academic research (Hickok, 2021). Nevertheless, certain principles' prioritization clearly depends on its individual use case. To a certain extent, the findings of this work might have a degree of generalizability, but more practical research among various use cases is certainly necessary.

Overall, the analysis and prioritization framework developed in this research provides a foundation for a further stream of contributions aimed at its extension and refinement (e.g., various application scenarios, investigating rankings of stakeholder groups such as data scientists, or the extension to various cultural groups) over time with further related academic and non-academic theoretical and practical advancements (Taeihagh, 2021). The theoretical framework can be applied for various contexts as well.

From a practical perspective, this work provides essential guidelines for applying a structured theoretical research method, combining it with practical research efforts and thus establishing priorities and guidance for subsequent implementation.

In summary, human safety and security of the applied AI algorithms have been consistently rated as very important throughout the survey. Fairness and the elimination of various forms of bias or discrimination receive a high rating of importance as well. Respondents care for audit processes to limit mistakes and the results indicate that traceability and accountability are of high importance. Privacy and the protection of personal data have been valued as important overall, but as already mentioned, received greater attention in independent scenarios. Rated as least important were whether a human being has constant control over the AI application or whether the user is informed about the fact that he or she is dealing with an algorithm instead of a person.

According to the recommendations (Shneiderman, 2020), (1) the reliability of AI systems can be ensured through reliable software engineering practices in teams (e.g., audit logs for failure analysis, software engineering workflows, testing for verification and validation, bias testing for increased fairness, and explainable user interfaces). (2) The safety culture in organizations can be shaped by management strategies (e.g., management's commitment to safety, safety-focused hiring and

training, extensive reporting of failures and errors, internal review committees for issues and plans, and compliance with standard industry practices). (3) The trustworthiness certification can occur because of industry-wide efforts (e.g., state intervention and regulation, external auditing of accounting firms, compensation of insurance companies for failures, promotion of design principles by non-governmental and community organizations, creation of standards, policies, and new ideas by professional organizations and research institutes).

5.3. Limitations and future work

First, non-academic, or soft-law documents are grey literature and are not indexed in academic databases. The retrieval process is therefore less exactly replicable and unbiased in comparison to traditional academic systematic database searches. This work followed best practices of grey literature review and has been extensively tested prior to execution. Additionally, the PRISMA framework has been developed and applied by (Jobin et al., 2019) prior to this research.

The second limitation is due to the rapid pace of the publication of non-academic resources dealing with ethical AI. It is possible that important policy documents were published after the completion of this work. However, to minimize this risk, new literature has been manually checked, analyzed, and subsequently compared with the results until June 30, 2021. As with most empirical studies, this work relied on a limited sample size and a study population mainly from one country, i.e., Germany. Future work should take other countries or respectively cultures into consideration (Floridi & Cowls, 2019).

Third, by applying the PRISMA framework and building on the research foundation created by (Jobin et al., 2019), the literature review process has been performed on an overarching level. By building onto the here proposed methodology, further research could be done on a more detailed level, for example by investigating only certain ethical principles in specific scenarios, e.g., human safety in predictive policing scenarios, since it was the single most often mentioned principle across all four scenarios.

Some limitations arose because of the applied BWS methodology for prioritization. During the careful item construction phase, five overarching clusters have been created and items have been subsequently chosen based on these clusters. The five clusters were transparency, justice & fairness, responsibility, non-maleficence, and privacy. Great attention has been paid to the creation of heterogenous items that are as independent as possible from each other.

Having the guiding research questions in mind, future research needs to further explicate the foundation built here by deriving specific guidelines' rankings in cultural contexts, such as China (Hickok, 2021). The theoretical foundation has been investigated and presented in this work, while the conduction of practical research needs to be further investigated in future research.

With a sample size of 225 valid responses, we acknowledge that our findings may be limited, in part because they involve four scenarios and may include participants from various cultures. The number of participants is still sufficient to derive initial significant findings in the given research area.

Moreover, most of our respondents were young individuals between the ages of 18–39 (71%) and students in their current employment status (57%). However, this has the advantage that younger generations are, as expected, the intensive users of the technologies, thus proactively informing themselves and well considering their attitudes and behaviors (Dinev et al., 2009).

Another limitation of the study is that most participants are either German or at least live in Germany. Therefore, our work so far mainly reflects the German and Western perspective on the ethics of AI. The next important step would therefore be to extend our research to other countries to support the development of international and cross-cultural standards for AI ethics (van Berkel et al., 2022), (Awad et al., 2018).

It should be clarified that a quantitative approach like it was conducted in our study is not sufficient on its own. It would benefit greatly from being enhanced through qualitative small-n investigations such as expert interviews or specific focus group sessions. In a similar work to ours with a quantitative focus as well it has been stated that no normative "should" can be based on a descriptive "is" (Jakesch et al., 2022), (Musschenga, 2005). Our study should be understood as one part in a bigger puzzle that highlights the importance of being sensitive to specific contexts when talking about ethical AI decision-making.

6. Conclusion

In this paper we approached the evaluation of ethical guidelines for artificial intelligence systems from the viewpoint of prioritization. First, we reviewed the theoretical foundation (ethical guidelines and principles) and used the results to create an empirical survey to assess user's views in terms of prioritization. For the theoretical part, we employed a rigorous literature review procedure by using a PRISMA framework for both non-academic and academic sources and based our study on prior work by (Jobin et al., 2019) and (Brocke et al., 2009). In the empirical survey, participants were confronted with selected AI application scenarios that involve ethical challenges. Subsequently, respondents had to choose from a list of ethical principles which ones they considered as most and least important respectively.

The outcomes clearly indicate that participants view human safety and security of the AI system as most important regardless of one of the four scenarios they were randomly presented. The same holds true for items that can be summarized as the overarching principles of justice and fairness, i.e., elimination of bias and non-discrimination. Privacy and personal data usage have been overall rated as important as well, yet there are differences based on the respective scenarios (predictive policing and human resources). Respondents consistently answered that whether they are informed about dealing with AI or a human being is of less importance to them. This also applies for whether the decision-making process of AI is transparent and understandable as well. The results of our study support the incorporation of ethical values into AI applications and provide insightful contributions regarding the prioritization of ethical values in various real-life scenarios.

There is a need for more empirical research in the future that can build on the foundations built here to better understand the priorities of various individuals as well as societal or cultural groups. We argue that the creation of ethical guidelines must be accompanied by a deeper understanding of user needs. It is important to mention that most works up to date, including ours, only reflect a western perspective on AI ethics and it would be beneficial to extend this research to other perspectives and cultural contexts to get a more precise picture of how AI algorithms should be developed to serve humanity.

CRediT authorship contribution statement

Yannick Fernholz: Conceptualization, Data curation, Formal analysis, Methodology, Project administration, Validation, Visualization, Writing – original draft, Writing – review & editing. **Tatiana Ermakova:** Conceptualization, Investigation, Methodology, Supervision, Writing – original draft, Writing – review & editing. **B. Fabian:** Conceptualization, Data curation, Funding acquisition, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **P. Buxmann:** Supervision.

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.

References

[ERROR!: PLEASE CHECK REFERENCE STYLE] *"2019 Edelman AI survey: Survey of technology executives and the general population shows excitement and curiosity yet uncertainty and worries that artificial intelligence could be a tool of division.*(2019) [Online]. Available: https://www.edelman.com/sites/g/files/aatuss191/files/2019-03/2019_Edelman_AI_Survey_Whitepaper.pdf.

Anderson, M., Anderson, S. L., & Armen, C. (2006). An approach to computing ethics. *IEEE Intelligent Systems, 21*(4), 56–63. https://doi.org/10.1109/MIS.2006.64

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. 23* pp. 77–91). ProPublica.

Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology, 8*(1), 19–32. https://doi.org/10.1080/1364557032000119616

Asaro, P. M. (2019). AI ethics in predictive policing: From models of threat to an ethics of care. *IEEE Technology and Society Magazine, 38*(2), 40–53. https://doi.org/10.1109/MTS.2019.2915154

Auger, P., & Devinney, T. M. (2007). Do what consumers say matter? The misalignment of preferences with unconstrained ethical intentions. *Journal of Business Ethics, 76*(4), 361–383. https://doi.org/10.1007/s10551-006-9287-y

Auger, P., Devinney, T. M., & Louviere, J. J. (2007). Using best–worst scaling methodology to investigate consumer ethical beliefs across countries. *Journal of Business Ethics, 70*, 299–326. https://doi.org/10.1007/s10551-006-9112-7

Awad, E., et al. (2018). The moral machine experiment. *Nature, 563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Bagloee, S. A., Tavana, M., Asadi, M., & Oliver, T. (2016). Autonomous vehicles: Challenges, opportunities, and future implications for transportation policies. *J. Mod. Transport., 24*(4), 284–303. https://doi.org/10.1007/s40534-016-0117-3

Balakrishnan, M., Bansal, G., Cagigal, D., Mehta, R., & Thiel, T. (2019). Panel discussion: CIO panel on ethical framework for AI & big data. In *MWAIS 2019 proceedings* [Online]. Available: https://aisel.aisnet.org/mwais2019/6.

Barocas, S., & Boyd, D. (2017). Engaging the ethics of data science in practice. *Communications of the ACM, 60*(11), 23–25.

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38*(2), 143–156. https://doi.org/10.1509/jmkr.38.2.143.18840

Bingley, W. J., et al. (2023). Where is the human in human-centered AI? Insights from developer priorities and user experiences. *Computers in Human Behavior, 141*, Article 107617. https://doi.org/10.1016/j.chb.2022.107617

Boldt, M., Boeva, V., & Borg, A. (2018). Multi-expert estimations of burglars' risk exposure and level of pre-crime preparation using coded crime scene data: Work in progress. In *2018 European intelligence and security informatics conference (EISIC)* (pp. 77–80). https://doi.org/10.1109/EISIC.2018.00021

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science, 352*(6293), 1573–1576. https://doi.org/10.1126/science.aaf2654

Borges, J., et al. (2018). Time-series features for predictive policing. In *2018 IEEE international smart cities conference (ISC2)* (pp. 1–8). https://doi.org/10.1109/ISC2.2018.8656731

Bostrom, N., & Yudkowsky, E. (2018). The ethics of artificial intelligence. In *Artificial intelligence safety and security* (pp. 57–69). Chapman and Hall/CRC.

Braccini, A., & Federici, T. (2013). *"A Measurement Model for investigating digital natives and their organizational behavior," ICIS 2013 Proceedings* [Online]. Available: https://aisel.aisnet.org/icis2013/proceedings/ResearchInProgress/72.

Brendel, A. B., Trang, S., Marrone, M., Lichtenberg, S., & Kolbe, L. M. (2020). What to do for a literature review? – a synthesis of literature review practices. In *AMCIS 2020 proceedings* [Online]. Available: https://aisel.aisnet.org/amcis2020/meta_research_is/meta_research_is/2.

Brenner, W., & Herrmann, A. (2018). An overview of technology, benefits and impact of automated and autonomous driving on the automotive industry. In C. Linnhoff-Popien, R. Schneider, & M. Zaddach (Eds.), *Digital marketplaces unleashed* (pp. 427–442). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-662-49275-8_39.

Brocke, J. V., Simons, A., Niehaves, B., Niehaves, B., Reimer, K., Plattfaut, R., & Cleven, A. (2009). *Reconstructing the giant: On the importance of rigour in documenting the literature search process*.

Bruckes, M., Grotenhermen, J.-G., Cramer, F., & Schewe, G. (2019). "Paving the way for the adoption of autonomous driving: Institution-based trust as a critical success factor,". In *Presented at the European conference on information systems (ECIS), Stockholm & uppsala, Sweden, jun* [Online]. Available: https://aisel.aisnet.org/ecis2019_rp/87.

Burrell, J., & Fourcade, M. (2021). The society of algorithms. *Annual Review of Sociology, 47*(1), 213–237. https://doi.org/10.1146/annurev-soc-090820-020800

Calhoun, C. C., Stobbart, C. E., Thomas, D. M., Villarrubia, J. A., Brown, D. E., & Conklin, J. H. (2008). Improving crime data sharing and analysis tools for a web-based crime analysis toolkit: WebCAT 2.2. In *2008 IEEE systems and information engineering design symposium* (pp. 40–45). https://doi.org/10.1109/SIEDS.2008.4559682

Chan, L., et al. (2020). Artificial artificial intelligence: Measuring influence of AI 'assessments' on moral decision-making. In *Proceedings of the AAAI/ACM conference on AI* (pp. 214–220). New York, NY, USA: Ethics, and Society. https://doi.org/10.1145/3375627.3375870.

Cochran, W. G., & Cox, G. M. (1992). *Experimental designs* (2nd ed.). Wiley. Apr. 13, 2022. [Online]. Available: https://www.wiley.com/en-us/Experimental+Designs%2C+2nd+Edition-p-9780471545675.

Cohen, S. (2003). Maximum difference scaling: Improved measures of importance and preference for segmentation. In *, Vol. 530. Sawtooth software conference proceedings* (pp. 61–74). Fir St., Sequim, WA: Sawtooth Software, Inc.

Cohen, E., Goodman, S., & Cohen, E. (2009). Applying best-worst scaling to wine marketing. *International Journal of Wine Business Research, 21*, 8–23.

Cohen, E. (2009). Applying best-worst scaling to wine marketing. *International Journal of Wine Business Research, 21*(1), 8–23. https://doi.org/10.1108/17511060910948008

Crawford, K. (2021). *Atlas of AI: The real worlds of artificial intelligence*. New Haven: Yale University Press.

Cruz, J. (2019). Shared moral foundations of embodied artificial intelligence. In *Proceedings of the 2019 AAAI/ACM conference on AI* (pp. 139–146). Honolulu, HI, USA: Ethics, and Society. https://doi.org/10.1145/3306618.3314280.

Dancy, J. (2004). *Ethics without principles*. Oxford University Press on Demand.

Dastin, J. (2022). Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of data and analytics* (pp. 296–299). Auerbach Publications.

David, A., Mamun, M. R. A., & Peak, D. (2019). "Risk and liability in autonomous vehicle technology,". In *Presented at the 25th americas conference on information systems (AMCIS), cancún, Mexico, aug. 15* [Online]. Available: https://aisel.aisnet.org/amcis2019/treo/treos/43.

Dinev, T., Goo, J., Hu, Q., & Nam, K. (2009). User behaviour towards protective information technologies: The role of national cultural differences. *Information Systems Journal, 19*(4), 391–412. https://doi.org/10.1111/j.1365-2575.2007.00289.x

Dobbe, R. I. J., Gilbert, T. K., & Mintz, Y. (2020). Hard choices in artificial intelligence: Addressing normative uncertainty through sociotechnical commitments. In *Proceedings of the AAAI/ACM conference on AI* (p. 242). New York, NY, USA: Ethics, and Society. https://doi.org/10.1145/3375627.3375861.

El Khattabi, G., Haij, O., Benelallam, I., & Bouyakhf, E. H. (2018). Detection of unethical intelligent agents in ethical distributed constraint satisfaction problems. In *Proceedings of the 2nd mediterranean conference on pattern recognition and artificial intelligence, rabat, Morocco* (pp. 52–57). https://doi.org/10.1145/3177148.3180083

Feng, C., Yuguo, J., Guiqin, L., & Zhiyuan, G. (2019). Ethical dilemma and countermeasure in artificial intelligence engineering. In *Proceedings of the 2019 international conference on modern educational technology, nanjing, China* (pp. 111–114). https://doi.org/10.1145/3341042.3341061

Ferguson, A. G. (2017). *The rise of big data policing: Surveillance, race, and the future of law enforcement*. New York: New York University Press.

Finn, A., & Louviere, J. J. (1992). Determining the appropriate response to evidence of public concern: The case of food safety. *Journal of Public Policy and Marketing, 11*(2), 12–25.

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review, 1*(1). https://doi.org/10.1162/99608f92.8cd550d1

Flynn, T. N., Louviere, J. J., Peters, T. J., & Coast, J. (2007). Best–worst scaling: What it can do for health care research and how to do it. *Journal of Health Economics, 26*(1), 171–189. https://doi.org/10.1016/j.jhealeco.2006.04.002

Giattino, C. M., Kwong, L., Rafetto, C., & Farahany, N. A. (2019). The seductive allure of artificial intelligence-powered neurotechnology. In *Proceedings of the 2019 AAAI/ACM conference on AI* (pp. 397–402). Honolulu, HI, USA: Ethics, and Society. https://doi.org/10.1145/3306618.3314269.

Godé, C., Brion, S., & Bohas, A. (2020). *The affordance-actualization process in a predictive policing context: Insights from the French military police*. ECIS. https://aisel.aisnet.org/ecis2020_rp/167.

Gómez-González, E., et al. (2020). *Artificial intelligence in medicine and healthcare: A review and classification of current and near-future applications and their ethical and social impact. arXiv:2001.09778 [cs]*. Aug. 06, 2020. [Online]. Available: http://arxiv.org/abs/2001.09778.

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines, 30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Hickok, M. (2021). Lessons learned from AI ethics principles for future actions. *AI Ethics, 1*(1), 41–47. https://doi.org/10.1007/s43681-020-00008-1

Hirsh, J. (2016). Predictive policing and civilian oversight: What will it take to get it right? *IEEE Potentials, 35*(5), 19–22. https://doi.org/10.1109/MPOT.2016.2569723

Hooker, J. N., & Kim, T. W. N. (2018). Toward non-intuition-based machine and artificial intelligence ethics: A deontological approach based on modal logic. In *Proceedings of the 2018 AAAI/ACM conference on AI* (pp. 130–136). New Orleans, LA, USA: Ethics, and Society. https://doi.org/10.1145/3278721.3278753.

Jakesch, M., Buçinca, Z., Amershi, S., & Olteanu, A. (2022). How different groups prioritize ethical values for responsible AI. In *2022 ACM conference on fairness, accountability, and transparency (FAccT '22), June 21–24, 2022, seoul, Republic of Korea* (p. 20). New York, NY, USA: ACM. https://doi.org/10.1145/3531146.3533097.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9). https://doi.org/10.1038/s42256-019-0088-2. Art. no. 9.

Karnouskos, S. (2018). Self-driving car acceptance and the role of ethics. *IEEE Transactions on Engineering Management, 67*(2), 252–265.

Kaur, D., Uslu, S., Rittichier J, K., & Durresi, A. (2022). Trustworthy artificial intelligence: A review. *ACM Computing Surveys*. https://doi.org/10.1145/3491209

Kazim, E., & Koshiyama, A. S. (2021). A high-level overview of AI ethics. *Patterns, 2*(9), Article 100314. https://doi.org/10.1016/j.patter.2021.100314

King, T. C., Aggarwal, N., Taddeo, M., & Floridi, L. (2020). Artificial intelligence crime: An interdisciplinary analysis of foreseeable threats and solutions. *Science and Engineering Ethics, 26*(1), 89–120. https://doi.org/10.1007/s11948-018-00081-0

Kiritchenko, S., & Mohammad, S. (2017). Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *, Vol. 2. Proceedings of the*

55th annual meeting of the association for computational linguistics (pp. 465–470). Vancouver, Canada: Association for Computational Linguistics. Short Papers.
Kirkpatrick, K. (2015). The moral challenges of driverless cars. *Communications of the ACM, 58*(8), 19–20. https://doi.org/10.1145/2788477
LaBrie, R., & Steinke, G. (2019). Towards a framework for ethical audits of AI algorithms. In *AMCIS 2019 proceedings* [Online]. Available: https://aisel.aisnet.org/amcis2019/data_science_analytics_for_decision_support/data_science_analytics_for_decision_support/24.
Lackes, R., Siepermann, M., & Vetter, G. (2020). *"Where can I take you? – The drivers of autonomous driving adoption," presented at the European conference on information systems (ECIS), an online AIS conference* [Online]. Available: https://aisel.aisnet.org/ecis2020_rp/159.
Larsen, K., Hovorka, D., Dennis, A., & West, J. (2018). *Understanding the elephant: The discourse approach to boundary identification and corpus construction for theory review articles*.
Lee, J. A., Soutar, G. N., & Louviere, J. (2007). Measuring values using best-worst scaling: The LOV example. *Psychology and Marketing, 24*(12), 1043–1058. https://doi.org/10.1002/mar.20197
Leukel, J., Mueller, M., & Sugumaran, V. (2014). *"The state of design science research within the BISE community: An empirical investigation," ICIS 2014 proceedings* [Online]. Available: https://aisel.aisnet.org/icis2014/proceedings/ISDesign/8.
Levy, Y., & Ellis, T. J. (2006). A systems approach to conduct an effective literature review in support of information systems research. *Informing Science: The International Journal of an Emerging Transdiscipline, 9*, 181–212.
Li, G., Deng, X., Gao, Z., & Chen, F. (2019). Analysis on ethical problems of artificial intelligence technology. In *Proceedings of the 2019 international conference on modern educational technology, nanjing, China, jun* (pp. 101–105). https://doi.org/10.1145/3341042.3341057
Li, X., & Zhang, T. (2017). An exploration on artificial intelligence application: From security, privacy and ethic perspective. In *2017 IEEE 2nd international conference on cloud computing and big data analysis (ICCCBDA)* (pp. 416–420). https://doi.org/10.1109/ICCCBDA.2017.7951949
Liberati, A., et al. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: Explanation and elaboration. *BMJ, 339*, b2700. https://doi.org/10.1136/bmj.b2700
Louviere, J. J., Flynn, T. N., & Marley, A. A. J. (2015). *Best-worst scaling: Theory, methods and applications*. Cambridge ; New York: Cambridge University Press.
Luetge, C. (2017). The German ethics code for automated and connected driving. *Philos. Technol., 30*(4), 547–558. https://doi.org/10.1007/s13347-017-0284-0
Maas, M. M. (2018). Regulating for 'normal AI accidents': Operational lessons for the responsible governance of artificial intelligence deployment. In *Proceedings of the 2018 AAAI/ACM conference on AI* (pp. 223–228). New Orleans, LA, USA: Ethics, and Society. https://doi.org/10.1145/3278721.3278766.
Madiega, T. A. (2019). *"EU guidelines on ethics in artificial intelligence: Context and implementation," Think Tank European Parliament*. Apr. 12, 2022. [Online]. Available: https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2019)640163.
Marley, A. A. J., & Louviere, J. J. (2005). Some probabilistic models of best, worst, and best–worst choices. *Journal of Mathematical Psychology, 49*(6), 464–480. https://doi.org/10.1016/j.jmp.2005.05.003
Martin, D. (2017). Who should decide how machines make morally laden decisions? *Science and Engineering Ethics, 23*(4), 951–967. https://doi.org/10.1007/s11948-016-9833-7
Martinho, A., Herber, N., Kroesen, M., & Chorus, C. (2021). Ethical issues in focus by the autonomous vehicles industry. *Transport Reviews, 41*(5), 556–577. https://doi.org/10.1080/01441647.2020.1862355
Martinsons, M., & Ma, D. (2009). Sub-cultural differences in information ethics across China: Focus on Chinese management generation gaps. *Journal of the Association for Information Systems, 10*(11). https://doi.org/10.17705/1jais.00213
Millard-Ball, A. (2018). Pedestrians, autonomous vehicles, and cities. *Journal of Planning Education and Research, 38*(1), 6–12. https://doi.org/10.1177/0739456X16675674
Mirbabaie, M., Brendel, A. B., & Hofeditz, L. (2022). Ethics and AI in information systems research. *Communications of the Association for Information Systems, 50*. https://doi.org/10.17705/1CAIS.05034
Molcho, G., Maier, S., Melero, F., & Aliprandi, C. (2014). Caper: Collaborative information, acquisition, processing, exploitation and reporting for the prevention of organised crime. In *2014 IEEE joint intelligence and security informatics conference* (p. 316). https://doi.org/10.1109/JISIC.2014.63, 316.
Müller-Bloch, C., & Kranz, J. (2015). *A framework for rigorously identifying research gaps in qualitative literature reviews*.
Musschenga, A. W. (2005). Empirical ethics, context-sensitivity, and contextualism. *Journal of Medicine and Philosophy, 30*(5), 467–490, 2005.
Oppermann, L., Boden, A., Hofmann, B., Prinz, W., & Decker, S. (2019). Beyond HCI and CSCW: Challenges and useful practices towards a human-centred vision of AI and IA. In *Proceed halfway to the future sympos*. https://doi.org/10.1145/3363384.3363481, 2019, 1–5.
Paradice, D., Freeman, D., Hao, J., Lee, J., & Hall, D. (2018). A review of ethical issue considerations in the information systems research literature. *Foundations and Trends® in Information Systems, 2*(2), 117–236.
Passi, S., & Barocas, S. (2019). Problem formulation and fairness. In *Proceedings of the conference on fairness, accountability, and transparency, New York, NY, USA* (pp. 39–48). https://doi.org/10.1145/3287560.3287567
Pham, M. T., Rajić, A., Greig, J. D., Sargeant, J. M., Papadopoulos, A., & McEwen, S. A. (2014). A scoping review of scoping reviews: Advancing the approach and enhancing the consistency. *Research Synthesis Methods, 5*(4), 371–385. https://doi.org/10.1002/jrsm.1123
Prates, M., Avelar, P., & Lamb, L. (2018). "On quantifying and understanding the role of ethics in AI research. In *, Vol. 55. A historical account of flagship conferences and journals," in EPiC Series in computing* (pp. 188–201). https://doi.org/10.29007/74gj
Retnowardhani, A., & Triana, Y. S. (2016). Classify interval range of crime forecasting for crime prevention decision making. In *2016 11th international Conference on knowledge, Information and creativity support systems (KICSS)* (pp. 1–6). https://doi.org/10.1109/KICSS.2016.7951409
Rhim, J., Lee, J.-H., Chen, M., & Lim, A. (2021). A deeper look at autonomous vehicle ethics: An integrative ethical decision-making framework to explain moral pluralism. *Frontiers in Robotics and AI, 8*. May 30, 2022. [Online]. Available: https://www.frontiersin.org/article/10.3389/frobt.2021.632394.
Rochel, J., & Evéquoz, F. (2021). Getting into the engine room: A blueprint to investigate the shadowy steps of AI ethics. *AI & Society, 36*(2), 609–622. https://doi.org/10.1007/s00146-020-01069-w
Rothenberger, L., Fabian, B., & Arunov, E. (2019). *Relevance of ethical guidelines for artificial intelligence – a survey and evaluation*. Sweden: Stockholm & Uppsala [Online]. Available: https://aisel.aisnet.org/ecis2019_rip/26.
Rupp, C., Hahn, J., Queins, S., Jeckle, M., & Zengler, B. (2005). *UML 2 glasklar: Praxiswissen für die UML-Modellierung und -zertifizierung, 2., überarbeitete und erweiterte edition*. München: Carl Hanser Verlag GmbH & Co. KG.
Saidi, W. A., & Zeki, A. M. (2019). The use of data mining techniques in crime prevention and prediction. In *2nd smart cities symposium (SCS 2019)* (pp. 1–4). https://doi.org/10.1049/cp.2019.0225
Sanderson, C., Douglas, D., & Lu, Q. (2023). *Implementing responsible AI: Tensions and trade-offs between ethics aspects*. arXiv preprint arXiv:2304.08275.
Scheuerman, M. K., Wade, K., Lustig, C., & Brubaker, J. R. (2020). How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis. *Proceed ACM on Human-Comput Inter., 4*(CSCW1), 58. https://doi.org/10.1145/3392866, 1–58:35.
Selter, J.-L., Wagner, K., & Schramm-Klein, H. (2022). *"Ethics and morality in AI - a systematic literature review and future research," presented at the European conference on information systems (ECIS), Timisoara, Romania* [Online]. Available: https://aisel.aisnet.org/ecis2022_rp/60.
Sengupta, S., et al. (2020). A review of deep learning with special emphasis on architectures, applications and recent trends. *Knowledge-Based Systems, 194*, Article 105596. https://doi.org/10.1016/j.knosys.2020.105596
Seymour, M. (2018). Artificial intelligence is No match for human stupidity: Ethical reflections on avatars and agents. In *ACIS 2018 proceedings* [Online]. Available: https://aisel.aisnet.org/acis2018/54.
Shneiderman, B. (2021). Human-centred AI. *Issues in Science & Technology, 37*(2), 56–61.
Shneiderman, B. (2020). Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Trans. Interact. Intell. Syst., 10*(4), 26. https://doi.org/10.1145/3419764, 1-26:31.
Siau, K., & Wang, W. (2020). Artificial intelligence (AI) ethics: Ethics of AI and ethical AI. *Journal of Database Management, 31*(2), 74–87. https://doi.org/10.4018/JDM.2020040105
Sokol, K. (2019). Fairness, accountability and transparency in artificial intelligence: A case study of logical predictive models. In *Proceedings of the 2019 AAAI/ACM conference on AI* (pp. 541–542). Honolulu, HI, USA: Ethics, and Society. https://doi.org/10.1145/3306618.3314316.
Sommerville, I., & Sawyer, P. (1997). *Requirements engineering: A good practice guide* (1st ed.). Chichester, Eng. ; New York: John Wiley & Sons.
Plummer, L. This is how Netflix's top-secret recommendation system works. Wired UK. https://www.wired.co.uk/article/how-do-netflixs-algorithms-work-machine-lea-rning-helps-to-predict-what-viewers-will-like. (Accessed 22 August 2017).
Sternberg, H. S., Chen, H., Hofmann, E., & Prockl, G. (2020). *Autonomous trucks: A supply chain adoption perspective*. Hawaii, USA: Maui [Online]. Available: https://aisel.aisnet.org/hicss-53/in/digital_supply_chain/5.
Susser, D. (2019). Invisible influence: Artificial intelligence and the ethics of adaptive choice architectures. In *Proceedings of the 2019 AAAI/ACM conference on AI* (pp. 403–408). Honolulu, HI, USA: Ethics, and Society. https://doi.org/10.1145/3306618.3314286.
Soper, S. Fired by bot at Amazon: "It's you against the machine." Star Tribune. https://www.startribune.com/fired-based-on-algorithms/600072977/. (Accessed 28 June 2021).
Svaldi, A. Unemployed Coloradans struggling with identity verification: "We are who we say we are." Denver Post. https://www.denverpost.com/2021/04/25/coloradoun employment-identity-verification-fraud/. (Accessed 25 April 2021).
Taeihagh, A. (2021). Governance of artificial intelligence. *Policy and Society, 40*(2), 137–157. https://doi.org/10.1080/14494035.2021.1928377
Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electronic Markets, 31*(2), 447–464. https://doi.org/10.1007/s12525-020-00441-4
Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*(4), 273–286. https://doi.org/10.1037/h0070288
van Berkel, N., Tag, B., Goncalves, J., & Hosio, S. (2022). Human-centred artificial intelligence: A contextual morality perspective. *Behaviour & Information Technology, 41*(3), 502–518. https://doi.org/10.1080/0144929X.2020.1818828
Vandeviver, C., & Bernasco, W. (2017). The geography of crime and crime control. *Applied Geography, 86*, 220–225. https://doi.org/10.1016/j.apgeog.2017.08.012
Walz, A., & Firth-Butterfield, K. (2019). *AI governance: A holistic approach to implement ethics into AI*. World Economic Forum.
Wang, W., & Siau, K. (2018). Ethical and moral issues with AI - a case study on healthcare robots. In *Proceedings of the 24th americas conference on information systems (2018, New Orleans, LA)* [Online]. Available: https://scholarsmine.mst.edu/bio_inftec_facwork/232.

Wang, S., & Yuan, K. (2019). Spatiotemporal analysis and prediction of crime events in Atlanta using deep learning. In *2019 IEEE 4th international conference on image, vision and computing (ICIVC)* (pp. 346–350). https://doi.org/10.1109/ICIVC47709.2019.8981090

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly, 26*(2). xiii–xxiii.

Weibel, N., Desai, P., Saul, L., Gupta, A., & Little, S. (2017). *HIV risk on twitter: The ethical dimension of social media evidence-based prevention for vulnerable populations*. Hawaii International Conference on System Sciences. *HICSS-50)*, Jan. 2017, [Online]. Available: https://aisel.aisnet.org/hicss-50/dsm/critical_and_ethical_studies/3.

Whittlestone, J., Nyrup, R., Anna, A., & Stephen, C. (2019). The role and limits of principles in AI ethics: Towards a focus on tensions. In *AAAI/ACM conference on AI, ethics, and society (AIES '19), january 27–28, 2019, Honolulu, HI, USA* (p. 6). New York, NY, USA: ACM. https://doi.org/10.1145/3306618.3314289.

Wiefel, J. (2021). *Required service characteristics for automated mobility as a service: A qualitative investigation* [Online]. Available: https://aisel.aisnet.org/ecis2021_rp/45.

Willis, O. (2018). *How social media is connecting people living with illness*. ABC News. February 26 https://www.abc.net.au/news/health/2018-02-27/how-social-media-is-connecting-people-living-with-illness/9484574.

Wolff, A., Gomez-Pilar, J., Nakao, T., & Northoff, G. (2019). Interindividual neural differences in moral decision-making are mediated by alpha power and delta/theta phase coherence. *Scientific Reports, 9*(1). https://doi.org/10.1038/s41598-019-40743-y. Art. no. 1.

Wolfswinkel, J. F., Furtmueller, E., & Wilderom, C. P. M. (2013). Using grounded theory as a method for rigorously reviewing literature. *European Journal of Information Systems, 22*(1), 45–55. https://doi.org/10.1057/ejis.2011.51

Yapo, A., & Weiss, J. (2018). *Ethical implications of bias in machine learning*. Hawaii International Conference on System Sciences. *HICSS-51)*, Jan. 2018, [Online]. Available: https://aisel.aisnet.org/hicss-51/os/topics_in_os/6.

Yu, H., Shen, Z., Miao, C., Leung, C., Lesser, V. R., & Yang, Q. (2018). Building ethics into artificial intelligence. In *Proceedings of the 27th international joint conference on artificial intelligence, Stockholm, Sweden* (pp. 5527–5533).

Zhang, D., et al. (2021). *The AI index 2021 annual report. arXiv:2103.06312 [cs]*, Mar. Apr. 12, 2022. [Online]. Available: http://arxiv.org/abs/2103.06312.

Zheng, X., Cao, Y., & Ma, Z. (2011). A mathematical modeling approach for geographical profiling and crime prediction. In *2011 IEEE 2nd international conference on software engineering and service science* (pp. 500–503). https://doi.org/10.1109/ICSESS.2011.5982362