Ethics and Information Technology (2024) 26:44
https://doi.org/10.1007/s10676-024-09773-7

ORIGINAL PAPER

Mapping the landscape of ethical considerations in explainable AI research

Luca Nannini[1,4] · Marta Marchiori Manerba[2,3] · Isacco Beretta[2]

Accepted: 4 May 2024

Abstract
With its potential to contribute to the ethical governance of AI, eXplainable AI (XAI) research frequently asserts its relevance to ethical considerations. Yet, the substantiation of these claims with rigorous ethical analysis and reflection remains largely unexamined. This contribution endeavors to scrutinize the relationship between XAI and ethical considerations. By systematically reviewing research papers mentioning ethical terms in XAI frameworks and tools, we investigate the extent and depth of ethical discussions in scholarly research. We observe a limited and often superficial engagement with ethical theories, with a tendency to acknowledge the importance of ethics, yet treating it as a monolithic and not contextualized concept. Our findings suggest a pressing need for a more nuanced and comprehensive integration of ethics in XAI research and practice. To support this, we propose to critically reconsider transparency and explainability in regards to ethical considerations during XAI systems design while accounting for ethical complexity in practice. As future research directions, we point to the promotion of interdisciplinary collaborations and education, also for underrepresented ethical perspectives. Such ethical grounding can guide the design of ethically robust XAI systems, aligning technical advancements with ethical considerations.

Keywords Explainable AI (XAI) · AI ethics · Ethical analysis · Bibliometric study

Introduction

As Artificial Intelligence (AI) system continues to be integrated, the ambiguity and opacity of these systems have stirred considerable concern, prompting an increase in the focus on eXplainable AI (XAI) research. The intention of XAI is to shed light on the internal workings of AI systems, thereby making them more transparent, comprehensible, and accountable (Gunning & Aha, 2019). This effort aligns closely with the broader endeavor towards ethical governance of AI. Indeed, the ethical implications of AI technologies have gained significant attention due to their potential to perpetuate existing inequalities, produce unintended negative consequences, and create new ethical dilemmas (Blasimme & Vayena, 2020; Buyl et al., 2022; Jobin et al., 2019; Pastaltzidis et al., 2022). However, the extent to which XAI research genuinely addresses ethical considerations, and effectively assimilates them into the design, development, and evaluation of AI systems, remains a topic of considerable debate (Balasubramaniam et al., 2023; van Otterlo & Atzmueller, 2020; Kaur et al., 2020; Alufaisan et al., 2021; Chazette et al., 2019).

Explanations can be misleading, oversimplified, or biased, and may not always align with human values and preferences (Bertrand et al., 2022; He et al., 2023; Bordt et al., 2022; Balagopalan et al., 2022). Moreover, the level of detail and complexity of explanations must be carefully calibrated to the needs and capacities of different stakeholders (Bhatt et al., 2020; Liao et al., 2020; Ehsan et al., 2021; Zhang et al., 2020; Bansal et al., 2021). This also to ultimately inform and adopt unambiguous policies around

✉ Luca Nannini
l.nannini@usc.es; lnannini@minsait.com

Marta Marchiori Manerba
marta.marchiori@phd.unipi.it

Isacco Beretta
isacco.beretta@phd.unipi.it

[1] Centro Singular de Investigación en Tecnoloxías Intelixentes, Universidade de Santiago de Compostela, Santiago de Compostela, Spain

[2] Computer Science Department, Università di Pisa, Pisa, Italy

[3] KDD Laboratory, ISTI, National Research Council, Pisa, Italy

[4] Minsait, Indra Sistemas, Madrid, Spain

Published online: 25 June 2024

AI explainability (Nannini et al., 2023). Designing XAI systems thus requires grappling with complex trade-offs between competing values. Navigating such ethical challenges requires a deep engagement with the principles and frameworks of moral philosophy. Ethical theories such as deontology, consequentialism, and virtue ethics offer valuable resources for evaluating different XAI techniques and approaches. At the same time, the novel and complex nature of XAI systems may require going beyond traditional ethical frameworks to develop new context-specific principles and guidelines (Floridi et al., 2018; Vainio-Pekka et al., 2023). The field of applied ethics, which focuses on translating moral theories into action-guiding principles for real-world decision-making, provides a useful lens for considering the responsible development and deployment of XAI systems (Beauchamp & Childress, 2001).

The primary research question driving this study is: *“What is the extent and depth of ethical discussions within XAI research, and how are ethical theories or frameworks applied in this domain?"*. To pursue this research direction, we critically investigate the relationship between XAI and ethical considerations by conducting a systematic review of (410) papers gathered from Scopus. Our research queries were formalised with the aim of identifying papers proposing contributions at the intersection of both XAI and ethics. We classify the papers according to their treatment of ethical aspects in the context of XAI, using a rigorous 5-point approach that takes into account the presence, depth, and focus of ethical discussions, as well as the application of ethical theories. The main contributions of this paper are: (1) a novel methodology and the taxonomy to conduct a bibliometric study on the current state of ethical discourse in the XAI field; (2) a classification of XAI research papers based on their treatment of ethical aspects; (3) an in-depth analysis of the findings.

The rest of the paper is organized as follows. Section 2 provides a detailed overview of the background knowledge and introduces the necessary terminology essential for comprehending our discussion. Section 3 presents the methodology employed and describes our bibliometric approach, with the classification further detailed within Appendix A & B. In Sects. 4 we report upon our findings pertaining to the extent of ethical considerations within the field of XAI research. Section 5 deliberates on such considerations for a more comprehensive integration of ethical considerations in XAI, accounting for limitations in Sect. 6 before concluding in Sect. 7.

Background

XAI techniques can favor AI-systems comprehension by illuminating the reasoning behind their decisions. To grapple with ethical complexities, XAI research must engage substantively with normative ethical theories and principles from the field of applied ethics. This background section provides an overview of the major ethical frameworks relevant to XAI, outlines key ethical challenges in operationalizing XAI principles, and reviews related work examining the treatment of ethics in XAI research to date.

Section 2.1 summarizes the core tenets of deontological, consequentialist, and virtue ethics perspectives, considering their potential implications for the design and governance of XAI systems. Section 2.2 then discusses the crucial role of applied ethics in translating abstract moral theories into context-sensitive guidance. There, Sect. 2.2.1 briefly outlines key philosophical debates underpinning these ethical theories, such as the nature of moral reasoning, the grounding of moral principles, and the scope of moral consideration. With this philosophical grounding established, Sect. 2.3 finally positions our research among recent systematic mapping studies and self-critical assessments examining the treatment of ethics within the XAI research landscape.

Major ethical theories and their relevance to XAI

To provide a solid foundation for our analysis, it is essential to briefly discuss the main normative ethical theories that have shaped moral reasoning and decision-making, and consider their potential implications for XAI. These theories, namely *deontology*, *consequentialism*, and *virtue ethics*, offer distinct perspectives on what constitutes ethical behavior and how to navigate moral dilemmas (Shafer-Landau, 2012; Copp, 2006).

Deontological theories—focus on the intrinsic rightness of actions based on moral rules or duties (Kant, 1959; Ross, 1930). The core commitment is to ground objective moral principles in the nature of rational agency itself. Kant argued that moral requirements are categorical imperatives—absolute, universal duties derived from pure practical reason that hold independent of contingent desires or social conventions (Kant, 1959; Hill, 1992). Actions have moral worth only if done from a “good will"—a stable disposition to act from duty rather than mere inclination (Kant, 1996). In the context of XAI, a deontological approach would emphasize the inherent rightness of designing AI systems that respect user autonomy and provide truthful, non-deceptive explanations as a matter of moral duty, regardless of the consequences. Neo-Kantians have developed this idea in terms of respecting the autonomy of persons as “ends-in-themselves" and acting only on principles that could consistently serve as universal laws (Korsgaard, 1996; Hill, 1992; O'Neill, 1975). The focus is on the formal principle used to assess maxims, not the consequences of individual acts—this would require ensuring that the underlying principles and decision-making processes of XAI systems are universalizable and can be made transparent to users without violating their autonomy or dignity. Some deontologists, such as Ross, propose a system of *prima facie* duties that can conflict in particular situations, requiring agents to weigh and balance competing moral considerations (Ross, 1930). This perspective aligns with the discussion of deliberative agency and the conditions for a right to explanation (Jongepier & Keymolen, 2022).

Consequentialist theories—in contrast, maintain that the rightness of an action depends on the value of its consequences (Bentham, 1961; Mill, 1979; Sidgwick, 1907). Classic utilitarianism is the paradigmatic example, holding that we should act to maximize overall welfare or well-being impartially considered (Bentham, 1961; Mill, 1979). From this perspective, the development and deployment of XAI systems would be evaluated based on their overall impact on human welfare, taking into account factors such as the benefits of increased transparency and understanding, the potential risks of gaming or misuse, and the trade-offs between explainability and other desirable outcomes like accuracy or efficiency. More recent consequentialist views incorporate a wider range of goods beyond just pleasure or preference-satisfaction, and that allow for agent-neutral reasons to favor certain individuals (Parfit, 1984; Railton, 1984; Sen, 1979). But the core idea likely remains that the ethical status of XAI methods depends on their outcomes rather than their intrinsic nature or the motives behind them. Moral rules are at best reliable heuristics that should be set aside when better results are achievable by violating them (Smart & Williams, 1973; Hare, 1981). The consequentialist and deontological perspectives on explainability discussed in Kempt et al. (2022) illustrate the diverse ethical considerations at play in XAI design and the potential tensions between them.

Virtue ethics—shifts focus from right action to good character, contending that virtues are stable dispositions or traits that reliably lead to human flourishing (Anscombe, 1958; MacIntyre, 1981; Slote, 1992; Hursthouse, 1999). The tradition draws heavily from Aristotle's conception of virtues as mean states between extreme vices, cultivated through proper upbringing and practical judgment (*phronesis*) rather than rigid rule-following (Aristotle, 1999; Sherman, 1989). While specific virtues may differ across cultures, the basic notion is that character and context are as ethically relevant as actions and their consequences (Nussbaum, 1988). Neo-Aristotelians argue that truly virtuous agents act for the right reasons and with appropriate emotions, not just in line with moral duties (Hursthouse, 1999; Foot, 1978; Oakley, 1996). Practical reasoning, akin to skill, is thus essential for translating virtuous dispositions into situationally-sensitive judgments (McDowell, 1979). In the context of XAI, a virtue ethics approach would prioritize the development of AI systems that embody and promote virtuous character traits, such as honesty and benevolence. This would require cultivating the practical wisdom necessary to discern when and how to provide explanations that are sensitive to the needs and situations of individual users, rather than simply following rigid rules or protocols.

Philosophical debates and challenges

The three main ethical frameworks raise important philosophical questions and debates that complicate the work of translating ethical principles into practice. One key debate concerns the relationship between motives, actions, and consequences in determining the moral status of an agent or decision. As seen, deontological theories emphasize the intrinsic rightness of actions based on universal duties, while consequentialist theories focus solely on outcomes. Virtue ethics, meanwhile, stresses the importance of character and moral perception in navigating context-specific challenges (Adams, 1976). These differences have implications for how XAI systems are designed and evaluated, and for how the decision-making of human agents interacting with these systems is understood and assessed. Related to this is the question of moral worth—whether right actions must flow from good will or virtuous character to be praiseworthy, or if accidental conformity to moral principles suffices (Arpaly, 2002).

Another consideration is the nature and grounding of moral principles. Deontologists argue that moral rules are grounded in the necessary requirements of rational agency, consequentialists justify them by their generally optimific results, and virtue ethicists see moral rules as heuristics to guide those still cultivating practical wisdom (Hursthouse, 1999). However, all three approaches recognize that there can be hard cases where moral rules conflict (Ross, 1930; Smart & Williams, 1973). The ethical frameworks also take different positions regarding the scope of moral consideration and the demands of impartiality, which have implications for how broadly we conceive of the moral status of AI systems and the ethical obligations we have towards them (Sidgwick, 1907; Scheffler, 1982). Finally, the broader philosophical question of the nature of intelligence shapes the way we conceive of the capacities and limitations of AI (Boden, 2006), which has significant implications for the standards of interpretability, robustness, and control that we demand from XAI systems.

Applied ethics in XAI

As seen, while major ethical theories offer valuable normative foundations yet they might not be always well-suited for the practical challenges of designing and governing XAI systems. This is where the field of applied ethics comes in, developing mid-level principles and context-sensitive guidance to address the moral, political and social implications of technologies in real-world settings (Felzmann et al., 2020; Beauchamp & Childress, 2001). A wealth of XAI surveys and reviews mention such applied ethics principles while identifying and reporting various explanation techniques and domain, shedding light on diverse domain applications (Samek et al., 2017; Adadi & Berrada, 2018; Arrieta et al., 2020; Cambria et al., 2023; Stepin et al., 2021; Saeed & Omlin, 2023a; Martins et al., 2024).

While the major ethical theories offer valuable normative foundations, they are not always well-suited for the practical challenges of designing and governing XAI systems. As seen in the history of bioethics, a purely deductive approach that seeks to derive practical guidance from overarching moral theories is often insufficient due to the gap between theoretical principles and the nuanced ethical dilemmas encountered in practice (Gert et al., 2006; Jonsen, 2012). Drawing from the lessons of bioethics, XAI ethics should strive for a reflective equilibrium between principles, contextual factors, stakeholder perspectives, and the actual challenges arising in the development and use of explainable AI (Loi & Spielkamp, 2021; Theodorou et al., 2017). This involves an iterative process of specifying principles in light of practical considerations, while also allowing on-the-ground insights to inform the interpretation and balancing of competing principles. Effective applied ethics in XAI requires close engagement with the technical, organizational, and social realities shaping the technology, as well as the needs and concerns of diverse stakeholders (Langer et al., 2021; Muralidharan et al., 2024). This requires a sociotechnical lens attentive to cognitive biases, power dynamics, and the distribution of authority between human and algorithmic agents (Zhang et al., 2020; Kitamura et al., 2021).

In this spirit, XAI can learn from other domains where applied ethics has addressed the responsible development of emerging technologies, such as bioethics, environmental ethics, and research ethics (Cohen et al., 2014; Morley et al., 2021; Mittelstadt, 2019). These fields provide valuable strategies for inclusive stakeholder engagement, contextual awareness, balancing principles and practice, and navigating trade-offs. For example, bioethics offers tools for ethical deliberation and oversight (Solomon, 2005; Dubler & Liebman, 2011), while environmental ethics provides insights on balancing competing values amid uncertainty (Brennan & Lo, 2022).

Ethical complexity in XAI

Challenges remain in translating both major ethical theories and applied ethics concepts into concrete XAI practices (Zicari et al., 2021; Morley et al., 2023). Especially for applied ethics, conceptual tensions must be resolved between competing desiderata like transparency and privacy or efficiency and user-friendliness (Loi & Spielkamp, 2021; Theodorou et al., 2017; Mittelstadt et al., 2019; Brey, 2010; Ehsan et al., 2021). This requires going beyond blanket imperatives to consider which stakeholders need what types of explanations for which aspects of AI systems in particular contexts (Felzmann et al., 2019; Bhatt et al., 2020; Tsamados et al., 2022; Nyrup & Robinson, 2022).

The Transparency-Explainability Dilemma – One key challenge is the complex relationship between transparency and explainability. While enhanced transparency is often heralded as desirable for facilitating explainability and contributing to ethical goals (do Prado Leite & Cappelli, 2010; Cysneiros, 2013), mere amplification of transparency does not inherently lead to superior explainability without clear guidelines on what and how to disclose information (Habibullah & Horkoff, 2021; Chazette et al., 2019; Köhl et al., 2019). The interplay between transparency and other requirements like trust, privacy, security, and accuracy must also be considered (Zerilli et al., 2019). Conflating explainability and transparency can stem from XAI designers lacking in-depth ethical understanding or researchers exploiting “ethics" rhetoric without genuine consideration of societal needs (Floridi, 2019; Bietti, 2020; Wagner, 2018a). Designers should be guided by how disclosed information will be processed and used, not just the need to disclose (Miller, 2023; Cabitza et al., 2024, 2023).

Enhancing Accountability—XAI techniques can facilitate auditing by illuminating decision processes, but accountability also requires pathways for recourse when problems are detected. Relying solely on “after-the-fact" explanations can instill false confidence without appropriate feedback channels and governance (Mökander & Axente, 2023; Bordt et al., 2022; Casper et al., 2024). Further, fairness and bias mitigation present challenges for XAI. Various mathematical definitions of fairness exist, sometimes encoding mutually exclusive criteria (Brun et al., 2018; Chouldechova, 2017). Identifying appropriate standards requires normative deliberation, not just computational evaluation, always cognizant that XAI techniques can perpetuate biases if not carefully designed (Bertrand et al., 2022; Chaudhuri & Salakhutdinov, 2019; Shamsabadi et al., 2022). Finally, still regarding fairness perspective, value pluralism poses issues as diverse stakeholders bring different ethical priorities. The same model may demand distinct explanations for different audiences (Markus et al., 2021). Trade-offs arise between competing goods in high-stakes applications, benefiting from ethical analysis and community input.

Trust and Reliance Dynamics—Engendering appropriate trust and reliance in AI remains a key XAI motivation, but real-world dynamics are fraught. More or better explanations do not automatically improve human judgment or error detection (Zhang et al., 2020; Kitamura et al., 2021; Bertrand et al., 2022). Overconfidence can lead to misplaced trust, while exposing flaws might foster undue skepticism. Levels of explainability should be based on realities of imperfect human reasoning to avoid unfairly holding AI to a “double standard," while potentially justifying a higher bar e.g., given physicians’ ability to take responsibility for their own heuristics but not an AI’s inscrutable reasoning (Kempt et al., 2022). As Loi argues, moving beyond post-hoc explanations to consider broader institutional contexts and “design publicity" is beneficial (Loi et al., 2021).

Related mapping studies in XAI and AI ethics

As shown, core ethical principles often conflict when operationalized in real-world XAI deployments. Purely technical approaches cannot resolve the inevitable value tensions and contextual particularities at play. Instead, grappling with the ethics of XAI requires critically examining the assumptions, methods and impacts of these systems through interdisciplinary collaboration and inclusive stakeholder engagement. The field of applied ethics offers conceptual frameworks and methodological tools well-suited to this challenge, by putting technical choices in dialogue with their social and institutional context.

To better position the current research, it is worth noting that scholarly discourse has moved towards self-critical approaches in the XAI field, with meta-surveys or analogous structural work inquiring over future research directions with also stronger ethical considerations (Löfström et al., 2022; Saeed & Omlin, 2023a; Ali et al., 2023; Schmid & Wrede, 2022; Brand & Nannini, 2023). In particular, a recent manifesto by Longo et al. (2024) outlined 28 key challenges and future directions for XAI research, organized into 9 high-level categories. While the article’s primary focus was not on ethics, it recognized XAI as a key component of responsible AI and highlighted various ethical challenges and considerations that the XAI community needs to grapple with moving forward. These included the need for human-centered explanations, mitigation of potential negative impacts, and the role of XAI in addressing societal issues like power imbalances and the “right to be forgotten”. The authors advocated for participatory design approaches involving impacted stakeholders as an ethically-minded way forward. Brand and Nannini (2023) offer a unique philosophical perspective on the ethical grounding of XAI, arguing that it should be viewed not merely as a universal right, but as a moral duty rooted in the principle of reciprocity. They contend that XAI plays a crucial role in maintaining reciprocal relationships between human agents in AI-assisted decision-making contexts by providing transparency and supporting genuine reason-sharing. By highlighting XAI’s instrumental value in upholding human agency and moral duties in the face of opaque AI systems, they proceed to map how such approach to XAI would benefit different communities, such as of XAI techniques developers, HCI designers, and policymakers. Similarly, Kasirzadeh (2021) contributes to this critical examination by systematically mapping the relationships between technical explanations, value judgments, and stakeholder perspectives in XAI systems, complementing and extending the typologies and challenges identified in other mapping studies of the XAI ethics landscape.

Yet, to the best our knowledge, the work closest to this research is the systematic mapping study by Vainio-Pekka et al. (2023), investigating the role of XAI in the field of AI ethics research. Their work provided valuable insights into the prevalence of XAI as a research focus within empirical AI ethics scholarship, the main themes and methodological approaches in this area, and potential research gaps. While their work shares some similarities with the present study in terms of the broad topic and the use of a systematic mapping methodology, there are important differences in scope and emphasis. Notably, their study focused specifically on the role of XAI within empirical AI ethics research, whereas the current analysis considers the engagement with ethical considerations across the broader landscape of XAI research, including both empirical and theoretical work. In addition to that, our study places greater emphasis on the depth and quality of ethical engagement in XAI research, using a novel classification scheme to assess the level of ethical analysis and the application of specific ethical theories and frameworks.

By providing a more comprehensive and fine-grained analysis of the ethical dimensions of XAI research, the present research aims to complement and extend findings of the aforementioned studies, offering new insights into the current state of the field and opportunities for future work at the intersection of ethics and XAI.

Methodology

This study employs a systematic review approach to investigate the landscape of ethical considerations in explainable AI (XAI) research. Our methodology consists of three key stages: (1) formulating research queries in Subsection 3.1; (2) applying a multi-stage filtering process (Sect. 3.2); and (3) developing a taxonomy for classifying depth and quality of ethical engagement in XAI literature (Sect. 3.3).

Research queries

Identifying relevant papers necessitated systematic searching on Scopus. Our search strings incorporated both XAI-specific and ethics-specific terms. The selection of XAI-related terms (i.e., "*Explainable AI*," "*XAI*," "*interpretable machine learning*," "*interpretability*," and "*AI explainability*") was straightforward given their direct relevance to the research focus. The choice of ethics-related terms, however, required careful consideration due to the complexity and diversity of ethical concepts applicable in XAI context. We adopted a twofold approach:

- *Major Ethical Theories:* We incorporated key terms related to the major normative ethical theories, including consequentialism, deontology, virtue ethics, and care ethics (Alexander & Moore, 2021; Hursthouse & Pettigrove, 2018; Held, 2005). These theories provide the philosophical underpinnings for many of the ethical principles and frameworks discussed in the context of AI and XAI.
- *Applied Ethics in XAI:* We dove into the specifics of ethics as they pertain to XAI. Principles like transparency, accountability, and fairness have unique connotations in this context (Jobin et al., 2019; Weller, 2019). For instance, transparency might refer to the explainability of AI systems, while accountability might involve mechanisms to hold AI systems and their creators responsible.

This approach resulted in an extensive list of ethics-related keywords, aiming to encompass the multifaceted ethical discussions within XAI research, fully detailed in Appendix B. By casting a wide net across both foundational ethical theories and XAI-specific ethical principles, these search terms aim to capture a broad range of ethical discussions within the XAI literature.

Filtering process

The initial search yielded a pool of 410 papers which underwent a multi-stage filtering process to ensure the relevance and quality of the included studies. The filtering process was conducted by three PhD students in XAI, two of whom had backgrounds in AI ethics and policy, while the third had a more technical focus. This diverse expertise allowed for a comprehensive and balanced assessment of the papers.[1] The filtering process involved the following steps:

1. *Initial Pool Screening*—We started with the preliminary full pool of papers as follows: we first removed duplicate entries to ensure that each paper is considered only once; excluded papers that were not written in English as well as papers produced before 2016, the year of DARPA's XAI program release (Gunning & Aha, 2019), to focus on the most recent and relevant developments in the field. We finally also excluded papers that were not peer-reviewed i.e., tutorials, workshop abstracts, white papers, and theoretical reviews, to ensure the inclusion of high-quality, original research that advances the state-of-the-art in XAI tools, applications, evaluations, or theoretical/framing contributions.

   Each paper was screened reading titles and abstracts using a three-reviewer system: each paper was independently assessed by two members of the research team to determine its relevance to both XAI and ethical considerations. Papers were classified as "relevant," "irrelevant," or "uncertain". In particular, disagreements and "uncertain" papers were resolved through discussion and consensus, where also the third reviewer—with a more technical background—was consulted if consensus could not be reached, in order to minimize individual bias and ensure a more reliable selection process (Cumpston et al., 2019; McDonald et al., 2019).
2. *Examined Papers Review*—After the preliminary screening, we obtained 237 papers that were analyzed by conducting a full-text review. In this phase, we identified and excluded not relevant papers i.e., works that appeared relevant based on their title and abstract but did not directly contribute to the study's focus upon closer examination. In particular it was assessed the quality and depth of ethical discussions in the remaining papers using a four-step review process:

   (a) *Identification of Ethical Discussions:* Searching for any sections or subsections addressing ethical concerns, considerations, or issues within the context of XAI. Papers that did not have any mention on ethics in XAI were further excluded.
   (b) *Evaluation of Discussion Depth:* Evaluating the depth of the ethical discussions within each paper, considering the complexity of the ethical issues addressed, the sophistication of the analysis, and the extent to which ethics was integrated.
   (c) *Examination of Ethical Theories:* Identifying and evaluating mentioned and/or application of ethical theories reported in Sect. 2.

[1] Disagreements were resolved through a structured discussion process, where team members referred to the predefined classification categories and their quantitative thresholds (Table 5 in the Appendix A). Each member presented their arguments, and the group critically evaluated the evidence supporting each classification until a consensus was reached. This structured approach aimed to minimize individual biases.

(d) *Focus Evaluation:* Determining the paper's primary focus based on the research question, objectives, and overall contribution to the field. P

3. *Final Pool*—With the final pool (= 77 entries) established, we assigned for each entry the most suitable category from A to E to each remaining paper according to our proposed taxonomy, which is outlined in the following paragraph and detailed in Appendix A. The taxonomy considers both the depth and extent of ethical discussions and the paper's overall focus on ethics within the XAI context. This process was also conducted by the reviewers independently, with disagreements resolved through discussion and consensus. To further improve transparency and reproducibility, we documented the reasons for exclusion at each stage of the filtering process and maintained a detailed record[2].

Proposed classification taxonomy

To analyze depth and quality of ethical engagement in XAI research, we developed a novel classification scheme comprising five categories (A–E). This taxonomy builds upon existing approaches to evaluating the integration of ethical considerations in technology design and development, while addressing their limitations in capturing the specific nuances of the XAI context.

The categories are differentiated based on three key dimensions: (i) the depth of ethical discussion, (ii) the application of specific ethical theories or frameworks, and (iii) the overall emphasis on ethical issues in relation to XAI. By considering these dimensions in combination, our taxonomy provides a more comprehensive and fine-grained assessment of the ethical landscape within XAI research. Each category is associated with a set of quantitative thresholds and qualitative criteria to ensure a systematic and replicable classification process (see Appendix A). These thresholds were iteratively refined through pilot testing and calibration among the research team to enhance inter-rater reliability.

Results

In the primary phase of our bibliometric study, an initial pool of 410 research papers was established. Following the application of our predefined inclusion criteria, we subsequently eliminated 173 of these articles, leaving a sample of 237 papers for further review. Within this remaining pool, each paper was thoroughly examined, with both abstract and body text read and analyzed. Prior to the final classification process, an additional elimination of papers deemed as not relevant was undertaken. These were primarily research articles that emerged as false positives in our methodology—papers not directly applicable to our study focus. These included a total of 143 papers that treated the subjects of XAI or ethical considerations independently, without a focus on their intersection. This category also encompassed 17 survey articles that were identified within our pool. The entire process of filtering and categorization is visually depicted in Fig. 1.

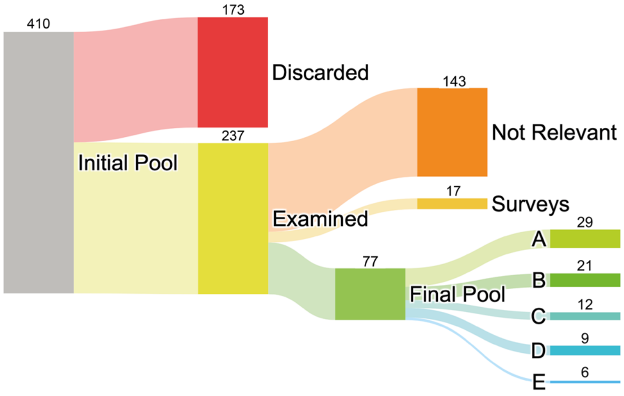

Fig. 1 Decision tree illustrating the distribution of papers at distinct stages of the process

Overview of paper distribution

Our multi-stage filtering process resulted in a final pool of 77 research papers for in-depth analysis. These papers were classified according to our pre-established five-tiered ranking system (A-E), which assessed the relevance and depth of ethical engagement in the context of XAI research.

Distribution across Categories—The distribution of papers across the five categories comprised 29 papers (37.66% of the pool) occurrences in Category **A**; 21–27.27% in Category **B**; 12–15.58% in Category **C**; 9–11.69% in Category **D**; 6–7.79% in Category **E**. Notably, over 60% of the papers fell into categories A and B, indicating a relatively superficial engagement with ethical considerations in a significant portion of XAI research. In contrast, only about 20% of the papers (categories D and E) demonstrated a deeper integration of ethical analysis into the design and development of XAI systems. Out of the 77 papers in the list, 39 (50.6%) were published in conference proceedings, 34 (44.2%) in journals, and 4 (5.2%) in workshops or other publication types. This distribution highlights the importance of both conferences and journals in advancing research on ethics in XAI.

Key Publication Venues—Several conferences and journals have emerged as key outlets for research on ethics in XAI, as reported in Table 1. The most prominent venue in

[2] Available at this CSV file.

Table 1 Key Publication Venues and References

Key Publication Venues	References
Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)	(Brunotte et al., 2022; Maruyama, 2021; Gulum et al., 2020; Benzmüller & Lomfeld, 2020; Meo et al., 2022; Zhang & Yu, 2022; Dexe et al., 2020; Lindner & Möllney, 2019; van Otterlo & Atzmueller, 2020)
AIES (AAAI/ACM Conference on AI, Ethics, and Society)	(Zhou et al., 2020; Slack et al., 2020; Lakkaraju & Bastani, 2020; Sullivan & Verreault-Julien, 2022)
Ethics and Information Technology	(Jongepier & Keymolen, 2022; Kempt et al., 2022; Theunissen & Browning, 2022)
FAccT (ex-FAT)	(Kasirzadeh & Smart, 2021; Hancox-Li, 2020)
Philosophy and Technology	(Baum et al., 2022; Herzog, 2022a)
Minds and Machines	(Narayanan & Tan, 2023; Robbins, 2019)
IEEE International Conference on Fuzzy Systems	(Hein et al., 2022; Alonso et al., 2020)
Advances in Intelligent Systems and Computing	(Gerdes, 2021; Alonso, 2020)
Frontiers in Artificial Intelligence and Applications	(Falomir & Costa, 2021)

Table 2 Disciplinary Domains and References among **C**, **D**, and **E** categories

Domain	Papers
Computer Science	(Hofeditz et al., 2022), (Baum et al., 2022), (El-Nasr & Kleinman, 2020), (Nicodeme, 2020), (Gerdes, 2021), (Dexe et al., 2020), (Lindner & Möllney, 2019), (Falomir & Costa, 2021), (van Otterlo & Atzmueller, 2020)
Ethics & Society	(Fleisher, 2022), (Larsson & Heintz, 2020), (Narayanan & Tan, 2023), (Kasirzadeh & Smart, 2021)
Medicine/Healthcare	(Martinho et al., 2021), (Morris et al., 2023), (Jongepier & Keymolen, 2022), (Kempt et al., 2022), (Heinrichs & Eickhoff, 2020), (Herzog, 2022a), (van der Waa et al., 2021), (Amann et al., 2020)
Law	(Graziani et al., 2023), (Sibai, 2020), (John-Mathews, 2021)
Business/Management	(Sullivan & Verreault-Julien, 2022)

the list is *Lecture Notes in Computer Science* (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), with 9 papers. This is followed by the conference *AIES* (AAAI/ACM Conference on AI, Ethics, and Society) with 4 papers; *Philosophy and Technology* with 2 papers and *Ethics and Information Technology* with 3 papers. Other notable venues include *Minds and Machines* with 2 papers, *IEEE International Conference on Fuzzy Systems* with 2 papers, and *Advances in Intelligent Systems and Computing* with 2 papers. These venues highlights the multidisciplinary nature of research on ethics in XAI.

In terms of disciplinary focus of publication venues, we report an excerpt of distribution across the categories that most consistenly engaged with ethical consideration in Table 2.

The full complete list seen a majority of papers (36%) coming from Computer Science outlets; Medicine/Healthcare venues account for the 32%; Ethics & Society outlets account for the 16%; Law venues are represented as the 12%; and Business/Management as the 4%. This distribution showcases the multidisciplinary nature of research on ethics in XAI, with significant contributions from computer science, medicine/healthcare, ethics & society, law, and business/management outlets. The strong representation of computer science and medicine/healthcare venues highlights technical and domain-specific considerations in the development and application of ethical XAI systems.

Depth and extent of ethical discussions

The majority of XAI articles in categories A-B make reference to ethical considerations regarding AI exclusively in the abstract or in the introduction section, without further developing the discussion. Ethics is often presented as a motivation for the work or used to contextualize the proposed XAI methods within the landscape of real-world applications. In practice, applications and ethical implications are almost always mentioned together, alongside legal issues. Furthermore, these considerations are typically used to introduce the term XAI in general as an AI ethics principle, rather than being concretely connected to the specific proposed method. Papers classified under categories C, D, and E demonstrated varying, yet more substantiated, levels of engagement with ethical theories and frameworks. A closer examination of the papers in each category reveals distinct patterns in the depth and quality of ethical engagement:

Table 3 References of papers that engage in a discourse regarding major ethical theories presented (not necessarily just one)

Ethical Theories	#	Refs.
Consequentialism	13	(Benzmüller & Lomfeld, 2020; Hein et al., 2022; Meo et al., 2022; Robbins, 2019; Kim & Routledge, 2022; Graziani et al., 2023; Morris et al., 2023; Jongepier & Keymolen, 2022; Kempt et al., 2022; Batliner et al., 2022; Sibai, 2020; Herzog, 2022b; Narayanan & Tan, 2023)
Deontological ethics	6	(Balasubramaniam et al., 2023; Narayanan & Tan, 2023; Lindner & Möllney, 2019; Benzmüller & Lomfeld, 2020; Batliner et al., 2022; Sibai, 2020)
Virtue ethics	10	(Lima et al., 2022; Zhang & Yu, 2022; El-Nasr & Kleinman, 2020; Dexe et al., 2020; Lindner & Möllney, 2019; Benzmüller & Lomfeld, 2020; Meo et al., 2022; Morris et al., 2023; Batliner et al., 2022; Sibai, 2020)

- *Category A* papers, which constituted the largest group (37.66%), typically mentioned ethics or ethical values in passing without engaging in any substantive ethical analysis. Many of these papers referred to ethics in the abstract or introduction as a general motivation for the work, but failed to connect these considerations to the specific XAI methods or applications being proposed.
- *Category B* papers (27.27%) went a step further by discussing ethical principles or values in the context of XAI, but still lacked a thorough or systematic ethical analysis. These papers often highlighted the importance of ethical considerations such as transparency, accountability, or fairness, but did not delve into the nuances of how these principles might be operationalized or navigated in practice.
- *Category C* papers (15.58%), such as (Graziani et al., 2023; Fleisher, 2022; Narayanan & Tan, 2023; El-Nasr & Kleinman, 2020; Nicodeme, 2020; Heinrichs & Eickhoff, 2020; Larsson & Heintz, 2020; Jongepier & Keymolen, 2022; Martinho et al., 2021; Waefler & Schmid, 2021; Morris et al., 2023; Löfström et al., 2022), present ethical analyses but do not explicitly link the ethical considerations to the design or development of specific XAI tools. Instead, they focus on critiquing existing approaches, highlighting ethical challenges, or proposing conceptual frameworks and guidelines for addressing ethical issues in XAI.
- *Category D* papers (11.69%) began to bridge this gap by proposing XAI tools or techniques that were informed by ethical considerations. Examples such (Hofeditz et al., 2022; Baum et al., 2022; Kempt et al., 2022; Lindner & Möllney, 2019; Sibai, 2020; Dexe et al., 2020; Kasirzadeh & Smart, 2021; Calegari et al., 2020; Gerdes, 2021; van Otterlo & Atzmueller, 2020), propose XAI tools or techniques informed by ethical considerations but do not thoroughly substantiate the connection between the ethical principles and the proposed solutions. These papers often focus on specific aspects of explainable AI, such as generating explanations for moral judgments (Lindner & Möllney, 2019), classifying AI crimes (Sibai, 2020), or designing human-agent collaboration protocols (van der Waa et al., 2021). However, the connection between the ethical principles invoked and the specific XAI solutions proposed was not always thoroughly substantiated or explored in depth.
- *Category E* papers, while representing the smallest proportion (7.79%), offered the most comprehensive and rigorous integration of ethical considerations into the design and development of XAI systems. Papers such as (John-Mathews, 2021; Falomir & Costa, 2021; van der Waa et al., 2021; Amann et al., 2020; Herzog, 2022a; Sullivan & Verreault-Julien, 2022), explicitly integrate ethical considerations into the design and development of XAI tools and provide comprehensive ethical analyses of the proposed solutions. These papers engage more deeply with ethical theories and frameworks, using them to guide the design and evaluation of explainable AI systems. For example, Amann et al. (2020) conducts an ethical assessment of explainability in AI-based clinical decision support using the “Principles of Biomedical Ethics," while (Sullivan & Verreault-Julien, 2022) proposes using the capability approach to provide ethical standards for algorithmic recourse. Notably, papers from philosophical, social science, and interdisciplinary backgrounds (e.g., (Amann et al., 2020)) often provide more extensive engagement with ethical theories and frameworks compared to papers from purely technical domains.

Across the papers reviewed, a diverse range of ethical theories and frameworks are applied to analyze the role of explainability in AI systems. Certain ethical theories and principles emerged as more prominent than others. *Consequentialism, Deontological Ethics, Virtue Ethics*) are relatively present. In Table 3, we report a list of works that refer to them more or less explicitly. Explicit mentions are also to be found to the “Principles of Biomedical Ethics" by Beauchamp & Childress (autonomy, beneficence, nonmaleficence, and justice) e.g., used as an analytical framework in Amann et al. (2020) to assess the ethical implications of explainability in AI-based clinical decision support systems. Similarly, Herzog (2022a) builds on the notion of “explicability" proposed by Floridi and Cowls (2019), which combines the demands for intelligibility and accountability of AI systems, to argue for the ethical and epistemological utility of explainable AI in medicine.

Several papers draw upon philosophical concepts and frameworks to examine the ethical dimensions of explainable AI. Narayanan and Tan (2023) explores the attitudinal tensions between explainability and trust in AI decision support tools, discussing the incompatible deliberative and unquestioning attitudes required for each. Kasirzadeh and Smart (2021) critiques the use of counterfactuals in algorithmic fairness and explainability, arguing that social categories may not admit counterfactual manipulation and proposing tenets for using counterfactuals in machine learning. John-Mathews (2021) introduces the concept of "denunciatory power" as an ethical desideratum for AI explanations, measuring their ability to reveal unethical decisions or behavior. Dexe et al. (2020) employs the Value Sensitive Design (VSD) method to facilitate transparency and the realization of ethical principles in AI and digital systems design. van der Waa et al. (2021) proposes three team design patterns with varying levels of agent autonomy and human involvement to enable moral decision-making in human-agent teams. Other papers engage with various ethical principles and concepts, such as informed consent, shared decision-making, accountability, fairness, and transparency (Jongepier & Keymolen, 2022; Kempt et al., 2022; Lindner & Möllney, 2019; Sullivan & Verreault-Julien, 2022).

Discussion

Our bibliometric analysis has revealed a complex landscape of ethical engagement within the field of XAI research. The quantitative findings expose a striking disparity between the high number of papers acknowledging the importance of ethics (categories A and B, >60%) and the limited number providing explicit theoretical ethical frameworks or substantively integrating ethical considerations into XAI design and development (categories D and E, <20%). This raises critical questions about the depth of ethical considerations and implications for XAI systems' application. We structure our discussion into three themes emerging from the observed patterns.

First, we discuss the prevalence of "ethics-acknowledging" research (Sect. 5.1), signaling ethics' importance but failing to substantively embed ethical complexity, arguing for rigorous engagement with ethical theories and frameworks. Second, we further advance explainability's inherent ethical tensions (Sect. 5.2), highlighted by the diverse ethical theories and principles applied, emphasizing the need for nuanced, context-specific guidelines navigating XAI's complex trade-offs and competing interests. Finally, we underscore ethical education and interdisciplinary collaborations' importance (Sect. 5.3) in advancing XAI's responsible development, drawing insights from the diverse disciplinary backgrounds represented and arguing for cross-disciplinary dialogue and incorporating underrepresented ethical perspectives.

From signaling to embedding ethical complexity

Ethics being mentioned as a general concept without substantive engagement—as reported in our coded categories **A** and **B**—suggests a trend of superficial treatment of ethical issues in XAI. We define such trend as the prevalence of "ethics-acknowledging" research. This approach risks oversimplifying the multifaceted nature of ethics and creating misalignment between the design of XAI systems and their intended ethical impacts.

As outlined in Sect. 2, the major ethical theories (deontology, consequentialism, and virtue ethics) and the field of applied ethics offer valuable frameworks for navigating the complex ethical challenges surrounding XAI systems (Shafer-Landau, 2012; Copp, 2006; Felzmann et al., 2020; Beauchamp & Childress, 2001). These theories provide the necessary grounding for substantive ethical engagement, enabling researchers to consider the specific implications of their XAI systems in light of established moral principles and context-specific guidelines (Floridi, 2019; Bietti, 2020). Yet our analysis reveals that while many XAI papers acknowledge the importance of ethics, there is often a lack of deep engagement with these theories and frameworks.

This failure to embed ethical considerations substantively in the research design, execution, or interpretation of XAI studies threatens to undermine the ethical grounding of these systems (Graziani et al., 2023; Floridi, 2019). We argue that such trend aligns with corporate ethics initiatives—also affecting XAI applications—that might lack both intrinsic value (as they are not undertaken out of genuine commitment to moral principles) and instrumental value (as they do not lead to beneficial outcomes for society) (Metcalf et al., 2019; Bietti, 2020). This dynamic risks perpetuating a superficial form of ethical engagement, where ethics is invoked to legitimize existing practices rather than to drive genuine transformation (Hu, 2021). Similarly, the lack of robust metrics for evaluating the ethical implications of XAI systems, as highlighted by Floridi's discussion of "ethics bluewashing," further compounds the risk of superficial ethical engagement (Floridi, 2019). Without clear, shared, and publicly accepted ethical standards, as well as metrics that capture not just the performance of XAI systems but also their potential adverse outcomes and adherence to ethical principles, the ethical claims made by XAI researchers may remain unsubstantiated and fail to drive genuine ethical progress in the field (Hu, 2021; Wagner, 2018b; Bietti, 2020).

To address these challenges, it would be beneficial for XAI researchers to be intentional about which ethical theories they apply and to consider the specific implications of their systems in light of these theories (Floridi, 2019; Wagner, 2018b). This necessitates asking critical questions such as: "*What ethical implications might arise due to the nature of my system, its users, or its context of use?"*. Such kind of questions move beyond generic ethical concerns to reflect over specific ethical paradigms that guide behavior and decision-making. As an example, by appealing to a consequentialist stance, the system should be evaluated on its ability to forecast and mitigate adverse outcomes. This would require metrics that not only measure the accuracy or performance of the system but also its potential implications (de Bruijn et al., 2022) while still being aware of difficulties in predicting all the negative possible consequences beforehand (Genus & Stirling, 2018). On the other hand, a deontological approach would prioritize fidelity to defined rules and principles (Alexander & Moore, 2021), being centered around regulatory compliance and integrity of operation, thus potentially being detrimental to more nuanced, contextually-grounded manner as advocated in the following subsection.

Inherent ethical tensions in explainability

Explainability in AI systems often intersects with deep-seated ethical dilemmas that arise from the very principles of our normative philosophical frameworks, as stressed in Sect. 2.2.1. For example, Narayanan and Tan (2023) discuss the attitudinal tensions between explainability and trust in AI decision support tools, arguing that the deliberative attitude required for meaningful engagement with explanations is incompatible with the unquestioning attitude implied by trust. Similarly, Kasirzadeh and Smart (2021) critique the use of counterfactuals in XAI, contending that social categories may not admit counterfactual manipulation. Addressing these tensions requires careful consideration of the specific context and stakeholders involved, as well as the development of nuanced ethical guidelines that can adapt to the unique challenges of different domains (Nyrup & Robinson, 2022).

Another challenge is ensuring meaningful stakeholder engagement throughout the XAI development process. The bibliometric analysis underscores the importance of involving domain experts, end-users, and affected communities in the design and evaluation of XAI systems (Langer et al., 2021; Muralidharan et al., 2024). The substantial contributions from computer science, philosophy, ethics, and interdisciplinary outlets highlight the need for continued cross-disciplinary dialogue and collaboration to address this gap. As echoed by van Otterlo and Atzmueller (2020); Kasirzadeh (2021); Amann et al. (2020), a multidisciplinary approach is crucial for balancing the various legitimate but potentially conflicting interests involved in XAI, such as transparency, privacy protection, and intellectual property rights (Langer et al., 2021; Muralidharan et al., 2024). However facilitating effective collaboration and communication between these diverse stakeholders can be difficult, particularly when there are differences in technical expertise, values, and priorities (Green, 2022; Kroll, 2021). As Metcalf et al. (2019) argue, the influence of corporate logics on the institutionalization of ethics in the tech industry can further complicate these efforts.

Nonetheless, these challenges also present valuable opportunities for advancing the integration of ethics into XAI. The development of standardized ethical frameworks and guidelines tailored to the specific needs of XAI can provide a common language and set of principles to guide the responsible development of explainable AI systems (Amann et al., 2020; Longo et al., 2024; Sokol & Flach, 2020). These frameworks should be informed by the insights gained from the diverse ethical theories and approaches identified in the bibliometric analysis, such as the "Principles of Biomedical Ethics" (Amann et al., 2020), the capability approach (Sullivan & Verreault-Julien, 2022), and the concept of "reflective equilibrium" between principles and practice (Loi & Spielkamp, 2021; Theodorou et al., 2017). Other recent frameworks, such as "Evaluative AI" (Miller, 2023), recognize the inherent tensions in XAI and aim to provide a more flexible and context-sensitive approach. By designing XAI systems that promote cognitive reflection, such frameworks can help developers and users navigate the ethical complexities of XAI in a more nuanced and contextually-grounded manner (Ehsan et al., 2022; Cabitza et al., 2024, 2023).

Educating to ethical theories and interdisciplinary collaborations

In line with stakeholders engagement, we finally underscore the value of cross-disciplinary dialogue in illuminating the multifaceted ethical landscape of XAI. Much can be learned from other domains of applied ethics, such as bioethics and environmental ethics, which have grappled with similar challenges of balancing competing values and interests in the face of uncertainty and high stakes (Beauchamp & Childress, 2001; Markus et al., 2021; Blasimme & Vayena, 2020).

The landscape of ethical theories is vast, encompassing not just the mainstream utilitarian or deontological approaches, but also less represented ones like virtue ethics, care ethics, and non-Western ethical traditions (Wu et al., 2023; Amugongo et al., 2023; Okolo et al., 2022). These lesser-known paths may offer valid perspectives, allowing to navigate ethical dilemmas in XAI through an unexplored lens (Okolo, 2023). In this vein, initiatives such as workshops, tutorials and courses designed to provide a robust understanding of ethical theories and their practical implications are instrumental in this endeavor. There are already promising steps in this direction, as evidenced by institutional initiatives like the NIST's effort to develop comprehensive reports on human psychology and tools for XAI implementation (Broniatowski, 2021; Phillips et al., 2021) or researches on the moral value of XAI for the public sector (Brand, 2023). In this spirit, future studies should further investigate how organizational constraints do influence XAI deployers' alignment with specific ethical stances and their willingness to express dissenting views (Hickok, 2021; Ibáñez & Olmeda, 2021; Kitamura et al., 2021).

Research limitations

Our study provides valuable insights into the ethical discourse within XAI research, but it is essential to consider the following key limitations:

1. *Scope and Framing*: It is important to note that our research queries were designed to capture a broad spectrum of ethical considerations in XAI research. As demonstrated by the search queries provided in the Appendix, we included both generic ethics terms (e.g., 'ethics,' 'ethical,' 'moral,' 'morality') and specific theories (e.g., 'deontology,' 'consequentialism,' 'virtue ethics'). This approach aimed to ensure that our analysis was not limited to papers explicitly mentioning ethical theories but also included those discussing ethical issues more broadly. By combining generic and specific ethical key terms, we sought to minimize the potential bias towards any particular ethical framework. Yet, focusing solely on works that explicitly discuss ethics in XAI may overlook articles that embed ethical considerations within alternative framings, such as "responsible AI" or "human-centered AI". Future research should explore these diverse conceptualizations to capture a more comprehensive understanding of the ethical landscape in XAI. In terms of linguistic and chronological constraints, we recognize that by concentrating on English-language articles published after 2016, we may have excluded valuable insights from non-English publications and pre-DARPA works (Gunning & Aha, 2019).
2. *Classification Complexity*: Despite our efforts to mitigate bias through double-coding, the inherent subjectivity in our research process remains a limitation. Researchers' shared backgrounds may influence their interpretations, emphasizing the importance of reflexivity, diverse research teams, and systematic approaches to managing subjectivity in future studies. Furthermore, our five-tier classification scheme, while useful for structured analysis, may oversimplify the intricate nature of ethical discussions. Future research could explore more nuanced or multi-dimensional classification approaches to better capture the complexity of ethical engagement in XAI.
3. *Academic Perspectives*: By focusing on the academic domain, our study does not fully capture the broader discourse on ethics in XAI that occurs in industry, policy-making, and societal contexts. These non-academic spaces may surface practical and societal considerations and misalignment that are less emphasized in scholarly publications but are critical for a holistic understanding of ethics in XAI (Nannini et al., 2023). Finally, while our study highlights the need for deeper engagement with ethical theories in XAI research, we acknowledge the constraints of scientific publishing. Not all AI journals may prioritize extensive discussions of philosophical works, which may contribute to the observed lack of depth in some papers. Future research could investigate these structural barriers and propose strategies for fostering more substantive ethical deliberation within the confines of academic publishing; similarly, research could greatly benefit from incorporating those non-academic perspectives while navigating the challenges of accessing and analyzing non-public or proprietary information.

Conclusion

Our study contributes to the growing body of research on the ethical dimensions of XAI by critically examining the depth and breadth of ethical engagement in the field. Our bibliometric analysis has revealed a complex landscape of ethical engagement within XAI research: while many studies acknowledge the importance of ethics, there is often a lack of depth in the application of ethical theories and frameworks. This superficial treatment risks oversimplifying the multifaceted nature of ethics and creating misalignments between the design of XAI systems and their intended ethical impacts. By acknowledging our limitations and identifying avenues for future research, we invite further exploration and discourse to advance a more comprehensive, nuanced, and inclusive understanding of ethics in XAI. Ultimately, our aim is to stimulate a reflective and actionable dialogue on the role of ethics in shaping the responsible development and deployment of explainable AI systems.

Appendix A: Protocol table for ethics classification

Recognizing the diverse ways in which ethical considerations can be integrated into XAI research, we have developed a systematic classification scheme to assess the depth and quality of ethical engagement across the analyzed literature as reported in Sect. 3. This classification protocol serves as a guide to examine the content and focus of each paper, and subsequently assign it to one of five categories (A-E). The categories are differentiated based on three key dimensions: (i) the depth of ethical discussion, (ii) the application of specific ethical theories or frameworks, and (iii) the overall emphasis on ethical issues in relation to XAI. The resulting classifications aims to reveal the prevalence of ethical discussions, alongside the extent to which these discussions are substantive, grounded in normative theories, and explicitly linked to the design and development of XAI tools and techniques.

To facilitate the assignment of each classification category (A-E), a rating system based on quantitative thresholds is established. These thresholds are based on key criteria and provide more precise and objective classification:

In this structured classification scheme, the depth of ethical discussion is evaluated through a Likert scale (Step 2), while the presence and application of ethical theories or frameworks are assessed separately (Step 3). The overall focus of the paper on ethical issues in XAI is also rated on a Likert scale (Step 4). These ratings are then summed to determine the quantitative thresholds for each category, as defined in Table 5. By combining a rigorous protocol with quantitative thresholds and illustrative examples reported in A.2, this classification approach aims to promote consistency, objectivity, and reproducibility in assessing the ethical dimensions of XAI research.

A.1 Guidance on applying quantitative thresholds

While the quantitative thresholds defined in Table 4 provide clear numerical ranges or values for determining the classification categories, consistent interpretation and application of these thresholds can be challenging, particularly when evaluating papers that may fall near the boundaries of a category. To aid in consistent application, we provide the following guidance:

1. For the "*Ethical Discussion Score*," annotators should consider not only the length of the ethical discussion but also its depth, complexity, and sophistication. A lengthy discussion that merely reiterates surface-level ethical principles without critical analysis or nuanced argumentation may not warrant a high score.

Table 4 Classification Protocol

Step	Action	Expected Outcome	Guidance
1	Identify explicit discussions of ethics in the context of XAI in the paper	Yes/No	If "Yes," the paper can be included in the classification
2	Evaluate the depth of ethical discussions. This should take into account the complexity, thoroughness, and sophistication of the ethical argumentation	Likert scale (1-5)	A rating of 1 indicates ethical considerations are only briefly mentioned, while a rating of 5 signifies an extensive ethical discussion with deep analysis of ethical principles in relation to XAI
3	Assess if the paper refers to any specific ethical theories or frameworks, and how they are applied to XAI	Yes/No + Description	If "Yes," specify the theory or framework and its application
4	Determine the main focus of the paper. This can be identified by understanding the research question, objectives, and the contribution the paper is making to the field of XAI	Likert scale (1-5)	A rating of 1 indicates the paper is not focused on ethics, while a rating of 5 signifies an extreme focus on ethics in relation to XAI
5	Assign the most appropriate category (A-E) to the paper based on steps 1-4 (Refer to Table 2a below)	A, B, C, D, E	The assigned category should reflect the overall depth in ethical discussion, the application of ethical theories/frameworks, and the focus on ethical issues in XAI

Table 5 Quantitative Thresholds for Classification Categories

Cat.	Definition	Quantitative Threshold	Guidance
A	Papers that merely mention ethics or ethical values but do not engage in any ethical analysis	Ethical Discussion Score: 1-2	The paper makes only passing reference to ethics or ethical values, with no meaningful analysis or discussion
B	Papers that discuss ethical values or principles in the context of XAI without providing a thorough ethical analysis	Ethical Discussion Score: 3-4	The paper includes a discussion of ethics, but the analysis is largely surface level, lacking in depth and sophistication
C	Papers that present a systematic ethical analysis, but the ethical considerations are not explicitly linked to the design or development of XAI tools	Ethical Discussion Score: 4-5 Ethical Theory Mentioned: Yes Primary Focus Score: 1-3	The paper engages in a systematic ethical analysis, but does not link these ethical considerations to XAI design or development
D	Papers that propose XAI tools or techniques that are informed by ethical considerations, but the connection between the ethical principles and the XAI solutions is not thoroughly substantiated	Ethical Discussion Score: 4-5 Ethical Theory Mentioned: Yes Primary Focus Score: 3-4	The paper proposes XAI tools or techniques that are informed by ethical considerations, but the connection between the ethical principles and the XAI solutions is not thoroughly substantiated
E	Papers that explicitly integrate ethical considerations into the design and development of XAI tools, and provide a comprehensive and rigorous ethical analysis of the proposed solutions	Ethical Discussion Score: 5 Ethical Theory Mentioned: Yes Primary Focus Score: 5	The paper explicitly integrates ethical considerations into the design and development of XAI tools, and provides a comprehensive and rigorous ethical analysis of the proposed solutions

2. When assessing whether an “*Ethical Theory*" is mentioned, annotators should look for explicit references to specific ethical frameworks (e.g., utilitarianism, deontology, virtue ethics) or their core principles. Vague allusions to ethical concepts or values may not qualify as mentioning a theory.
3. The “*Primary Focus Score*" should be based on the overall emphasis and centrality of ethical considerations in the paper, as evidenced by the research questions, objectives, and contribution to the field. A paper that primarily focuses on technical aspects of XAI, with ethical considerations as a secondary or peripheral concern, would receive a lower score.
4. In cases where a paper’s scores or characteristics straddle the boundaries of two categories, annotators should carefully consider the overall balance and alignment with the category definitions. If a clear determination cannot be made, the paper may be assigned to the lower category to maintain a conservative approach.
5. Annotators are encouraged to document and discuss any particularly challenging or ambiguous cases during the annotation process, as these instances may inform future refinements or clarifications to the classification scheme and guidance.

By adhering to this supplementary guidance and maintaining open communication among annotators, we aimed to promote consistent and reliable application of the quantitative thresholds, while acknowledging the inherent complexities involved in such evaluations.

A.2 Justification and structure of the classification scheme (A-E)

To further clarify the distinctions between categories, we provide illustrative examples from the analyzed literature:

- *Category A:* A paper that merely states e.g., “Ethical issues are important in XAI development" without any further analysis would fall into this category.
- *Category B:* A paper discussing the need for transparency and fairness in XAI systems, but not delving into a deeper examination of these ethical principles, would be classified as Category B.
- *Category C:* A paper that systematically analyzes the application of deontological ethics (e.g., Kantian ethics) to XAI, but does not explicitly link this analysis to the design or development of XAI tools, would be considered Category C.
- *Category D:* A paper proposing an XAI technique for enhancing fairness, citing ethical principles of non-discrimination, but without thoroughly substantiating the

connection between the proposed technique and the ethical principles, would fall under Category D.

- *Category E:* A paper that explicitly grounds the development of an XAI tool in the ethical framework of care ethics, providing a rigorous analysis of how the tool's design and implementation uphold the principles of attentiveness, responsibility, competence, and responsiveness, would be classified as Category E.

While the classification scheme aims to capture distinct levels of ethical integration, it is important to acknowledge its inherent limitations and potential biases. The assessment of the depth of ethical discussion and the determination of a paper's primary focus inevitably involve some degree of subjectivity, despite the efforts to establish clear criteria and quantitative thresholds. Additionally, annotator biases may persist despite the training and conflict resolution measures employed.

Appendix B: Research queries on scopus

The research queries employed in this study were carefully crafted to encompass the diverse ethical considerations relevant to the XAI field, as established in the Background Sect. 2. The selection of search terms was grounded in the key ethical theories, principles, and debates identified as pertinent to the design, development, and deployment of XAI systems. As outlined in the Methodology 3, our primary research question aimed to assess the extent and depth of ethical discussions within XAI research and the application of ethical theories or frameworks in this domain. To address this question comprehensively, we adopted a two-pronged approach in constructing our search queries:

- *Foundational Ethical Theories:* We incorporated terms related to the major normative ethical theories, such as deontology, consequentialism, virtue ethics, and care ethics 2.1. These theories provide the philosophical underpinnings for many of the ethical principles and frameworks discussed in the context of AI and XAI.
- *Applied Ethics in XAI:* We included terms specific to ethical principles and concepts relevant to XAI, such as transparency, accountability, fairness, and responsible AI design 2.2. These principles capture the unique ethical challenges and considerations that arise in the development and deployment of explainable AI systems.

The following Scopus search queries were used in June 2023, reflecting this comprehensive approach:

```
TITLE-ABS-KEY ( "explainable AI" OR "interpretable AI" AND "ethical theories"
    ↪ AND "application" ) = 0
TITLE-ABS-KEY("Explainable AI" OR "XAI" OR "interpretable machine learning" OR
    ↪  "interpretability" OR "AI explainability") AND TITLE-ABS-KEY("ethics"
    ↪ OR "ethical" OR "moral" OR "morality" ) = 409
TITLE-ABS-KEY("Explainable AI" OR "XAI" OR "interpretable machine learning" OR
    ↪  "interpretability" OR "AI explainability") AND TITLE-ABS-KEY("
    ↪ deontology" OR "consequentialism" OR "virtue ethics" OR "care ethics"
    ↪ OR "ethics of care" OR "utilitarianism" OR "rights-based ethics" OR "
    ↪ contractualism" OR "social contract theory" OR "relational ethics" OR "
    ↪ distributive justice") = 4
TITLE-ABS-KEY("Explainable AI" OR "XAI" OR "interpretable machine learning" OR
    ↪  "interpretability" OR "AI explainability") AND TITLE-ABS-KEY("ethics"
    ↪ OR "ethical" OR "moral" OR "morality" AND "responsible AI" OR "ethical
    ↪ design" OR "ethical impact assessment") = 25
```

Acknowledgements Funding contribution from the ITN project NL4XAI Natural Language for Explainable AI. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860621. This document reflects the views of the author(s) and does not necessarily reflect the views or policy of the European Commission. The REA cannot be held responsible for any use that may be made of the information this document contains.

This work has been partially supported by the European Community Horizon 2020 programme under the funding scheme ERC-2018-ADG G.A. 834756 XAI: Science and technology for the eXplanation of AI decision making.

Author contributions I, Luca Nannini, am writing to clarify the authorship of our recent submission. All authors listed have agreed with the content and provided explicit consent for the submission. We have also obtained the necessary ethical consent from the responsible authorities (i.e., the University of Santiago de Compostela) where the research has been conducted. We understand and appreciate the stance of Springer Ethics and Information Technology in guiding our submission. However, please be assured that all listed authors have adhered to the authorship guidelines applicable in our specific research field. 1. All secondary authors whose names appear on the submission have made substantial contributions to the conception, design of the work, or the acquisition, analysis, or interpretation of data (references consulted); 2. Authors has contributed to drafting the work or revising it critically for important intellectual content; 3. All authors have approved the version to be published upon conditions to allow its acceptance. 4. All authors agree to be accountable for all aspects of the work, ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. 5. The nature of the work submitted is not made available in any public form elsewhere and it is submitted just to this Special Issue; 6. The nature of the work was checked to be devoid of plagiarized contribution; reported ideas and concepts are adequately attributed to third authors through references mentioned in the body text. In summary, all authors included in the submission have contributed significantly to the research and preparation of the manuscript. Please note that I personally conceptualized the study design, as well as that me and Ms. Marchiori Manerba are the main ones responsible for the funding acquisition that allowed this research. The research reported is the result of a concerted effort by all authors in terms of writing, editing, and collaborative supervision. Thank you for your understanding and your consideration of our work.

Data availability The bibliometric information extracted from Scopus is accessible via the provided hyperlink to the corresponding CSV file.

Declarations

Conflict of interest The authors have no Conflict of interest to declare that are relevant to the content of the submitted draft.

Human Participants No experiment was conducted with human participants or animals for this study.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Adams, R. M. (1976). Motive utilitarianism. *The Journal of Philosophy, 73*(14), 467–481.

Alexander, L., & Moore, M. (2021). Deontological Ethics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy, Winter* (2021st ed.). Metaphysics Research Lab: Stanford University.

Ali, S., Abuhmed, T., El-Sappagh, S., et al. (2023). Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion* p 101805. https://doi.org/10.1016/j.inffus.2023.101805, https://www.sciencedirect.com/science/article/pii/S1566253523001148

Alonso, J.M., Toja-Alamancos, J., & Bugarín, A. (2020). Experimental study on generating multi-modal explanations of black-box classifiers in terms of gray-box classifiers. In *29th IEEE International Conference on Fuzzy Systems*, FUZZ-IEEE 2020, Glasgow, UK, July 19-24, 2020. IEEE, pp 1–8, https://doi.org/10.1109/FUZZ48607.2020.9177770

Alonso, R.S. (2020). Deep symbolic learning and semantics for an explainable and ethical artificial intelligence. In: Novais, P., Vercelli, G.V., Larriba-Pey, J.L., et al. (eds) Ambient Intelligence - Software and Applications—11th International Symposium on Ambient Intelligence, ISAmI 2020, L'Aquila, Italy, October 7 - 9, 2020, Advances in Intelligent Systems and Computing, vol 1239. Springer, pp 272–278, https://doi.org/10.1007/978-3-030-58356-9_30

Alufaisan, Y., Marusich, L.R., Bakdash, J.Z., et al. (2021). Does explainable artificial intelligence improve human decision-making? In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021. AAAI Press, Virtual Event, February 2-9,2021, pp 6618–6626, https://ojs.aaai.org/index.php/AAAI/article/view/16819

Amann, J., Blasimme, A., Vayena, E., et al. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics Decis Mak, 20*(1), 310. https://doi.org/10.1186/S12911-020-01332-6

Amugongo, L.M., Bidwell, N.J., & Corrigan, C.C. (2023). Invigorating ubuntu ethics in AI for healthcare: Enabling equitable care. In *Proceedings of the 2023 ACM Conference on Fairness*, Accountability, and Transparency, FAccT 2023, Chicago, IL, USA, June 12-15, 2023. ACM, pp 583–592, https://doi.org/10.1145/3593013.3594024

Anscombe, G. E. M. (1958). Modern moral philosophy. *Philosophy, 33*(124), 1–19.

Aristotle. (1999). *Nicomachean ethics*. Hackett Publishing.

Arpaly, N. (2002). Moral worth. *The Journal of Philosophy, 99*(5), 223–245.

Arrieta, A. B., Rodríguez, N. D., Ser, J. D., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion, 58*, 82–115. https://doi.org/10.1016/J.INFFUS.2019.12.012

Balagopalan, A., Zhang, H., Hamidieh, K., et al. (2022). The road to explainability is paved with bias: Measuring the fairness of explanations. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea*, June 21 - 24, 2022. ACM, pp 1194–1206, https://doi.org/10.1145/3531146.3533179

Balasubramaniam, N., Kauppinen, M., Rannisto, A., et al. (2023). Transparency and explainability of AI systems: From ethical guidelines to requirements. *Inf Softw Technol, 159*, 107197. https://doi.org/10.1016/j.infsof.2023.107197

Bansal, G., Wu, T., Zhou, J., et al. (2021). Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery*, New York, NY, USA, CHI '21, https://doi.org/10.1145/3411764.3445717

Batliner, A., Hantke, S., & Schuller, B. W. (2022). Ethics and good practice in computational paralinguistics. *IEEE Trans Affect Comput, 13*(3), 1236–1253. https://doi.org/10.1109/TAFFC.2020.3021015

Baum, K., Mantel, S., Speith, T., et al. (2022). From responsibility to reason-giving explainable artificial intelligence. *Philosophy and Technology, 35*(1), 1–30. https://doi.org/10.1007/s13347-022-00510-w

Beauchamp, T. L., & Childress, J. F. (2001). *Principles of Biomedical Ethics*. USA: Oxford University Press.

Bentham, J. (1961). *An introduction to the principles of morals and legislation*. Clarendon Press.

Benzmüller, C., & Lomfeld, B. (2020). Reasonable machines: A research manifesto. In: Schmid, U., Klügl, F., & Wolter, D. (eds) *KI 2020: Advances in Artificial Intelligence - 43rd German Conference on AI*, Bamberg, Germany, September 21-25, 2020, Proceedings, Lecture Notes in Computer Science, vol 12325. Springer, pp 251–258, https://doi.org/10.1007/978-3-030-58285-2_20

Bertrand, A., Belloum, R., Eagan, J.R., et al. (2022). How Cognitive Biases Affect XAI-Assisted Decision-Making: A Systematic Review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society. Association for Computing Machinery*, New York, NY, USA, AIES '22, p 78-91,https://doi.org/10.1145/3514094.3534164

Bhatt, U., Xiang, A., Sharma, S., et al. (2020). Explainable machine learning in deployment. In [81], pp 648–657. https://doi.org/10.1145/3351095.3375624

Bietti, E. (2020), From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy. In: Hildebrandt, M., Castillo, C., Celis, L.E., et al. (eds) *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain*, January 27-30, 2020. ACM, pp 210–219, https://doi.org/10.1145/3351095.3372860

Blasimme, A., & Vayena, E. (2020). The ethics of ai in biomedical research, patient care, and public health. In *The Oxford Handbook of Ethics of AI*. Oxford University Press, UK, https://doi.org/10.1093/oxfordhb/9780190067397.013.45, https://academic.oup.com/book/0/chapter/290676282/chapter-ag-pdf/44521915/book_34287_section_290676282.ag.pdf

Boden, M. A. (2006). *Mind as Machine: A History of Cognitive Science*. Oxford: Oxford University Press.

Bordt, S., Finck, M., Raidl, E., et al. (2022). Post-hoc explanations fail to achieve their purpose in adversarial contexts. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea*, June 21 - 24, 2022. ACM, pp 891–905, https://doi.org/10.1145/3531146.3533153

Brand, J. (2023). Exploring the moral value of explainable artificial intelligence through public service postal banks. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society. Association for Computing Machinery*, New York, NY, USA, AIES '23, p 990-992, https://doi.org/10.1145/3600211.3604741

Brand, J.L.M., & Nannini, L. (2023). Does explainable ai have moral value? arXiv:2311.14687

Brennan, A., & Lo, N. Y. S. (2022). Environmental Ethics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy, Summer* (2022nd ed.). Metaphysics Research Lab: Stanford University.

Brey, P. (2010). Values in technology and disclosive computer ethics. In L. Floridi (Ed.), *The Cambridge Handbook of Information and Computer Ethics* (pp. 41–58). United Kingdom: Cambridge University Press.

Broniatowski, D. (2021). Psychological foundations of explainability and interpretability in artificial intelligence. Tech. rep., NIST, https://doi.org/10.6028/NIST.IR.8367. https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=931426

de Bruijn, H., Warnier, M., & Janssen, M. (2022). The perils and pitfalls of explainable AI: strategies for explaining algorithmic decision-making. *Gov Inf Q, 39*(2), 101666. https://doi.org/10.1016/J.GIQ.2021.101666

Brun, Y., Johnson, B., & Meliou, A. (2018). Fairness definitions explained. *ACM, 10*(1145/3194770), 3194776.

Brunotte, W., Chazette, L., Klös, V., et al. (2022). Quo vadis, explainability? - A research roadmap for explainability engineering. In Gervasi, V., & Vogelsang, A. (eds) *Requirements Engineering: Foundation for Software Quality - 28th International Working Conference*, REFSQ 2022, Birmingham, UK, March 21-24, 2022, Proceedings, Lecture Notes in Computer Science, vol 13216. Springer, pp 26–32, https://doi.org/10.1007/978-3-030-98464-9_3

Buijsman, S., Klenk, M., & van den Hoven, J. (forthcoming). Ethics of artificial intelligence. In Smuha, N. (ed) Cambridge Handbook on the Law, Ethics and Policy of AI. Cambridge University Press.

Buyl, M., Cociancig, C., Frattone, C., et al. (2022). Tackling algorithmic disability discrimination in the hiring process: An ethical, legal and technical analysis. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, New York, NY, USA, FAccT '22, p 1071-1082,https://doi.org/10.1145/3531146.3533169

Cabitza, F., Campagner, A., Famiglini, L., et al. (2023). Let me think! investigating the effect of explanations feeding doubts about the AI advice. In Holzinger, A., Kieseberg, P., Cabitza, F., et al. (eds) *Machine Learning and Knowledge Extraction - 7th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference*, CD-MAKE 2023, Benevento, Italy, August 29 - September 1, 2023, Proceedings, Lecture Notes in Computer Science, vol 14065. Springer, pp 155–169, https://doi.org/10.1007/978-3-031-40837-3_10

Cabitza, F., Natali, C., Famiglini, L., et al. (2024). Never tell me the odds: Investigating pro-hoc explanations in medical decision making. *Artificial Intelligence in Medicine, 150*, https://doi.org/10.1016/j.artmed.2024.102819. https://www.sciencedirect.com/science/article/pii/S0933365724000617

Calegari, R., Omicini, A., & Sartor, G. (2020). Argumentation and logic programming for explainable and ethical AI. In Musto, C., Magazzeni, D., Ruggieri, S., et al. (eds) *Proceedings of the Italian Workshop on Explainable Artificial Intelligence co-located with 19th International Conference of the Italian Association for Artificial Intelligence*, XAI.it@AIxIA 2020, Online Event, November 25-26, 2020, CEUR Workshop Proceedings, vol 2742. CEUR-WS.org, pp 55–68, https://ceur-ws.org/Vol-2742/paper5.pdf

Cambria, E., Malandri, L., Mercorio, F., et al. (2023). A survey on XAI and natural language explanations. *Information Processing and Management, 60*(1), 103111. https://doi.org/10.1016/J.IPM.2022.103111

Casper, S., Ezell, C., Siegmann, C., et al. (2024). Black-box access is insufficient for rigorous ai audits. arXiv:2401.14446

Chaudhuri, K., & Salakhutdinov, R. (eds) (2019). Fairwashing: the risk of rationalization, Proceedings of Machine Learning Research, vol 97, PMLR, http://proceedings.mlr.press/v97/aivodji19a.html

Chazette, L., Karras, O., & Schneider, K. (2019). Do end-users want explanations? analyzing the role of explainability as an emerging aspect of non-functional requirements. In Damian, D.E., Perini, A., Lee, S. (eds) *27th IEEE International Requirements Engineering Conference*, RE 2019, Jeju Island, Korea (South), September 23-27, 2019. IEEE, Jeju Island, Korea (South), pp 223–233, https://doi.org/10.1109/RE.2019.00032

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153–163. https://doi.org/10.1089/BIG.2016.0047

Cohen, I. G., Amarasingham, R., Shah, A., et al. (2014). The legal and ethical concerns that arise from using complex predictive analytics in health care. *Health Affairs, 33*(7), 1139–1147. https://doi.org/10.1377/hlthaff.2014.0048

Copp, D. (Ed.). (2006). *The Oxford Handbook of Ethical Theory*. New York: Oxford University Press.

Cumpston, M., Li, T., Page, M. J., et al. (2019). Updated guidance for trusted systematic reviews: a new edition of the cochrane handbook for systematic reviews of interventions. *The Cochrane database of systematic reviews, 2019*(10).

Cysneiros, L.M. (2013). Using i* to elicit and model transparency in the presence of other non-functional requirements: A position paper. In Castro, J., Horkoff, J., Maiden, N.A.M., et al. (eds) Proceedings of the 6th International *i** Workshop 2013, Valencia, Spain, June 17-18, 2013, CEUR Workshop Proceedings, vol 978. CEUR-WS.org, Spain, pp 19–24, https://ceur-ws.org/Vol-978/paper_4.pdf

Dexe, J., Franke, U., Nöu, A.A., et al. (2020). Towards increased transparency with value sensitive design. In Degen, H., Reinerman-Jones, L. (eds) *Artificial Intelligence in HCI - First International Conference, AI-HCI 2020, Held as Part of the 22nd HCI International Conference*, HCII 2020, Copenhagen, Denmark, July 19-24, 2020, Proceedings, Lecture Notes in Computer Science, vol 12217. Springer, Denmark, pp 3–15, https://doi.org/10.1007/978-3-030-50334-5_1

Dubler, N. N., & Liebman, C. B. (2011). *Bioethics mediation: A guide to shaping shared solutions*. Vanderbilt University Press.

Ehsan U, Passi S, Liao QV, et al. (2021) The who in explainable AI: how AI background shapes perceptions of AI explanations. CoRR arXiv:2107.13509.

Ehsan, U., Wintersberger, P., Liao, Q.V., et al. (2022). Human-centered explainable ai (hcxai): Beyond opening the black-box of ai. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery*, New York, NY, USA, CHI EA '22,https://doi.org/10.1145/3491101.3503727

El-Nasr, M.S., & Kleinman, E. (2020). Data-driven game development: Ethical considerations. In Yannakakis, G.N., Liapis, A., Kyburz, P., et al. (eds) *FDG '20: International Conference on the Foundations of Digital Games*, Bugibba, Malta, September 15-18, 2020. ACM, Malta, pp 64:1–64:10, https://doi.org/10.1145/3402942.3402964

Elish, M.C., Isaac, W., & Zemel, R.S. (eds) (2021). *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto*, Canada, March 3-10, 2021, ACM, https://doi.org/10.1145/3442188

Falomir, Z., & Costa, V. (2021). On the rationality of explanations in classification algorithms. In Villaret, M., Alsinet, T., Fernández, C., et al. (eds) *Artificial Intelligence Research and Development - Proceedings of the 23rd International Conference of the Catalan Association for Artificial Intelligence*, CCIA 2021, Virtual Event, 20-22 October, 2021, Frontiers in Artificial Intelligence and Applications, vol 339. IOS Press, pp 445–454, https://doi.org/10.3233/FAIA210165

Felzmann, H., Fosch-Villaronga, E., Lutz, C., et al. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data Soc, 6*(1), 205395171986054. https://doi.org/10.1177/2053951719860542

Felzmann, H., Fosch-Villaronga, E., Lutz, C., et al. (2020). Towards transparency by design for artificial intelligence. *Science and Engineering Ethics, 26*(6), 3333–3361. https://doi.org/10.1007/S11948-020-00276-4

Fleisher, W. (2022). Understanding, idealization, and explainable ai. *Episteme, 19*(4), 534–560. https://doi.org/10.1017/epi.2022.39

Floridi, L. (2019). Translating principles into practices of digital ethics: Five risks of being unethical. *Philosophy & Technology, 32*(2), 185–193. https://doi.org/10.1007/s13347-019-00354-x

Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*, *1*(1). URL: Https://hdsr.mitpress.mit.edu/pub/l0jsh9d1.

Floridi, L., Cowls, J., Beltrametti, M., et al. (2018). Ai4people—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds Mach, 28*(4), 689–707. https://doi.org/10.1007/S11023-018-9482-5

Foot, P. (1978). *Virtues and vices and other essays in moral philosophy*. University of California Press.

Genus, A., & Stirling, A. (2018). Collingridge and the dilemma of control: Towards responsible and accountable innovation. *Research Policy, 47*(1), 61–69. https://doi.org/10.1016/j.respol.2017.09.012. https://www.sciencedirect.com/science/article/pii/S0048733317301622

Gerdes, A. (2021). Dialogical guidelines aided by knowledge acquisition: Enhancing the design of explainable interfaces and algorithmic accuracy. In Arai, K., Kapoor, S., & Bhatia, R. (eds) *Proceedings of the Future Technologies Conference (FTC) 2020*, Volume 1, Advances in Intelligent Systems and Computing, vol 1288. Springer, Cham, https://doi.org/10.1007/978-3-030-63128-4_19

Gert, B., Culver, C. M., & Clouser, K. D. (2006). *Bioethics: a return to fundamentals*. Oxford University Press.

Graziani, M., Dutkiewicz, L., Calvaresi, D., et al. (2023). A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences. *Artificial Intelligence Review, 56*(4), 3473–3504. https://doi.org/10.1007/s10462-022-10256-8

Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law and Security Review, 45*, 105681. https://doi.org/10.1016/J.CLSR.2022.105681

Gulum, M.A., Trombley, C.M., & Kantardzic, M.M. (2020). Multiple interpretations improve deep learning transparency for prostate lesion detection. In Gadepally, V., Mattson, T.G., Stonebraker, M., et al. (eds) *Heterogeneous Data Management, Polystores, and Analytics for Healthcare - VLDB Workshops, Poly 2020 and DMAH 2020*, Virtual Event, August 31 and September 4, 2020, Revised Selected Papers, Lecture Notes in Computer Science, vol 12633. Springer, pp 120–137, https://doi.org/10.1007/978-3-030-71055-2_11

Gunning, D., & Aha, D. W. (2019). Darpa's explainable artificial intelligence (XAI) program. *AI Mag, 40*(2), 44–58. https://doi.org/10.1609/aimag.v40i2.2850

Habibullah, K.M., & Horkoff, J. (2021). Non-functional requirements for machine learning: Understanding current use and challenges in industry. In *29th IEEE International Requirements Engineering Conference*, RE 2021, Notre Dame, IN, USA, September 20-24, 2021. IEEE, USA, pp 13–23, https://doi.org/10.1109/RE51729.2021.00009

Hancox-Li, L. (2020). Robustness in machine learning explanations: does it matter? In [81], pp 640–647, https://doi.org/10.1145/3351095.3372836

Hare, R. M. (1981). *Moral thinking: Its levels, method, and point*. Oxford University Press.

He, G., Kuiper, L., & Gadiraju, U. (2023). Knowing about knowing: An illusion of human competence can hinder appropriate reliance on AI systems. In Schmidt, A., Väänänen, K., Goyal, T., et al. (eds) *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI 2023, Hamburg, Germany, April 23-28, 2023. ACM, pp 113:1–113:18, https://doi.org/10.1145/3544548.3581025

Hein, A., Meier, L.J., Buyx, A., et al. (2022). A fuzzy-cognitive-maps approach to decision-making in medical ethics. In *IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2022*, Padua, Italy, July 18-23, 2022. IEEE, Padua, Italy, July 18-23, 2022, pp 1–8, https://doi.org/10.1109/FUZZ-IEEE55066.2022.9882615

Heinrichs, B., & Eickhoff, S. (2020). Your evidence? machine learning algorithms for medical diagnosis and prediction. *Human Brain Mapping, 41*(6), 1435–1444. https://doi.org/10.1002/hbm.24886

Held, V. (2005). *The Ethics of Care: Personal, Political, and Global*. Oxford: Oxford University Press. https://doi.org/10.1093/0195180992.001.0001

Herzog, C. (2022). On the ethical and epistemological utility of explicable ai in medicine. *Philosophy and Technology, 35*(2), 1–31. https://doi.org/10.1007/s13347-022-00546-y

Herzog, C. (2022). On the ethical and epistemological utility of explicable ai in medicine. *Philosophy & Technology, 35*(2), 50.

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

Hildebrandt, M., Castillo, C., Celis, L.E., et al. (eds) (2020). FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020, ACM, https://doi.org/10.1145/3351095

Hill JThomas, E. (1992). *Dignity and practical reason in Kant's moral theory*. Cornell University Press.

Hofeditz, L., Clausen, S., Rieß, A., et al. (2022). Applying XAI to an ai-based system for candidate management to mitigate bias and discrimination in hiring. *Electron Mark, 32*(4), 2207–2233. https://doi.org/10.1007/S12525-022-00600-9

Hu, L. (2021). Tech ethics: Speaking ethics to power, or power speaking ethics? *Journal of Social Computing, 2*(3), 238–248. https://doi.org/10.23919/JSC.2021.0033. https://www.sciopen.com/article/10.23919/JSC.2021.0033

Hursthouse, R. (1999). *On virtue ethics*. Oxford University Press.

Hursthouse, R., & Pettigrove, G. (2018) Virtue ethics in en zalta (ed.) the stanford encyclopedia of philosophy.

Ibáñez, J. C., & Olmeda, M. V. (2021). *Operationalising AI ethics: how are companies bridging the gap between practice and principles?* An exploratory study: AI & Soc. https://doi.org/10.1007/s00146-021-01267-0

Information Commissioner's Office (ICO) of the United Kingdom, The Alan Turing Institute (2019) Project explain - interim report. https://ico.org.uk/media/about-the-ico/documents/2615039/project-explain-20190603.pdf

International Standards Association (ISO) SAIS (2023) Iso/iec awi ts 6254 -information technology - artificial intelligence - objectives and approaches for explainability of ml models and ai systems. https://www.iso.org/standard/82148.html

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399. https://doi.org/10.1038/S42256-019-0088-2

John-Mathews, J. (2021). Some critical and ethical perspectives on the empirical turn of AI interpretability. CoRR arXiv:2109.09586.

Jongepier, F., & Keymolen, E. (2022). Explanation and agency: exploring the normative-epistemic landscape of the "right to explanation''. *Ethics and Information Technology, 24*(4), 49. https://doi.org/10.1007/S10676-022-09654-X

Jonsen, A. R. (2012). The ethics of organ transplantation: a brief history. *AMA Journal of Ethics, 14*(3), 264–268. https://doi.org/10.1001/virtualmentor.2012.14.3.mhst1-1203

Kant, I. (1959). *Foundations of the metaphysics of morals*. Bobbs-Merrill.

Kant, I. (1996). *The metaphysics of morals*. Cambridge University Press.

Kasirzadeh, A. (2021). Reasons, values, stakeholders: A philosophical framework for explainable artificial intelligence. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery*, New York, NY, USA, FAccT '21, p 14, https://doi.org/10.1145/3442188.3445866

Kasirzadeh, A., & Smart, A. (2021). The use and misuse of counterfactuals in ethical machine learning. In [55], pp 228–236, https://doi.org/10.1145/3442188.3445886

Kaur, H., Nori, H., Jenkins, S., et al. (2020). Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In Bernhaupt, R., Mueller, F.F., Verweij, D., et al. (eds) *CHI '20: CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu, HI, USA, April 25-30, 2020, pp 1–14, https://doi.org/10.1145/3313831.3376219

Kempt, H., Heilinger, J., & Nagel, S. K. (2022). Relative explainability and double standards in medical decision-making. *Ethics and Information Technology, 24*(2), 20. https://doi.org/10.1007/S10676-022-09646-X

Kim, T. W., & Routledge, B. R. (2022). Why a right to an explanation of algorithmic decision-making should exist: A trust-based approach. *Business Ethics Quarterly, 32*(1), 75–102. https://doi.org/10.1017/beq.2021.3

Kitamura, Y., Quigley, A., Isbister, K., et al. (2021). Does the Whole Exceed its Parts? *The Effect of AI Explanations on Complementary Team Performance, ACM, 10*(1145/3411764), 3445717.

Köhl, M.A., Baum, K., Langer, M., et al. (2019). Explainability as a non-functional requirement. In Damian, D.E., Perini, A., Lee, S. (eds) *27th IEEE International Requirements Engineering Conference*, RE 2019, Jeju Island, Korea (South), September 23-27, 2019. IEEE, Jeju Island, Korea (South), pp 363–368, https://doi.org/10.1109/RE.2019.00046

Korsgaard, C. M. (1996). *Creating the kingdom of ends*. Cambridge University Press.

Kroll, J.A. (2021). Outlining traceability: A principle for operationalizing accountability in computing systems. In [55], pp 758–771. https://doi.org/10.1145/3442188.3445937

Lakkaraju, H., & Bastani, O. (2020). "how do I fool you?": Manipulating user trust via misleading black box explanations. In [116], pp 79–85, https://doi.org/10.1145/3375627.3375833

Langer, M., Oster, D., Speith, T., et al. (2021). What do we want from explainable artificial intelligence (xai)? - A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial intelligence, 296*, 103473. https://doi.org/10.1016/J.ARTINT.2021.103473

Larsson, S., & Heintz, F. (2020). Transparency in artificial intelligence. *Internet Policy Rev*, *9*(2). https://doi.org/10.14763/2020.2.1469

Liao, Q.V., Gruen, D.M., & Miller, S. (2020). Questioning the AI: informing design practices for explainable AI user experiences. In Bernhaupt, R., Mueller, F.F., Verweij, D., et al. (eds) *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA*, April 25-30, 2020. ACM, pp 1–15, https://doi.org/10.1145/3313831.3376590

Lima, G., Grgic-Hlaca, N., Jeong, J.K., et al. (2022). The conflict between explainable and accountable decision-making algorithms. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea*, June 21 - 24, 2022. ACM, Jeju Island, Korea (South), pp 2103–2113, https://doi.org/10.1145/3531146.3534628

Lindner, F., & Möllney, K. (2019). Extracting reasons for moral judgments under various ethical principles. In Benzmüller, C., & Stuckenschmidt, H. (eds) *KI 2019: Advances in Artificial Intelligence - 42nd German Conference on AI, Kassel, Germany*, September 23-26, 2019, Proceedings, Lecture Notes in Computer Science, vol 11793. Springer, Germany, pp 216–229, https://doi.org/10.1007/978-3-030-30179-8_18

Löfström, H., Hammar, K., & Johansson, U. (2022). A meta survey of quality evaluation criteria in explanation methods. In Weerdt, J.D., & Polyvyanyy, A. (eds) *Intelligent Information*

Systems - CAiSE Forum 2022, Leuven, Belgium, June 6-10, 2022, Proceedings, Lecture Notes in Business Information Processing, vol 452. Springer, pp 55–63,https://doi.org/10.1007/978-3-031-07481-3_7

Loi, M., & Spielkamp, M. (2021). Towards accountability in the use of artificial intelligence for public administrations. In Fourcade, M., Kuipers, B., Lazar, S., et al. (eds) *AIES '21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA*, May 19-21, 2021. ACM, pp 757–766,https://doi.org/10.1145/3461702.3462631

Loi, M., Ferrario, A., & Viganò, E. (2021). Transparency as design publicity: explaining and justifying inscrutable algorithms. *Ethics and Information Technology, 23*(3), 253–263. https://doi.org/10.1007/S10676-020-09564-W

Longo, L., Brcic, M., Cabitza, F., et al. (2024). Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion, 106*, 102301. https://doi.org/10.1016/j.inffus.2024.102301. https://www.sciencedirect.com/science/article/pii/S1566253524000794

MacIntyre, A. (1981). *After virtue: A study in moral theory*. University of Notre Dame Press.

Markham, A.N., Powles, J., Walsh, T., et al. (eds) (2020). *AIES '20: AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA*, February 7-8, 2020, ACM, https://doi.org/10.1145/3375627

Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics, 113*, 103655. https://doi.org/10.1016/J.JBI.2020.103655

Martinho, A., Kroesen, M., & Chorus, C. G. (2021). A healthy debate: Exploring the views of medical doctors on the ethics of artificial intelligence. *Artificial Intelligence in Medicine, 121*, 102190. https://doi.org/10.1016/J.ARTMED.2021.102190

Martins, T., de Almeida, A. M., Cardoso, E., et al. (2024). Explainable artificial intelligence (XAI): A systematic literature review on taxonomies and applications in finance. *IEEE Access, 12*, 618–629. https://doi.org/10.1109/ACCESS.2023.3347028

Maruyama, Y. (2021). Categorical artificial intelligence: The integration of symbolic and statistical AI for verifiable, ethical, and trustworthy AI. In Goertzel, B., Iklé, M., & Potapov, A. (eds) *Artificial General Intelligence - 14th International Conference, AGI 2021, Palo Alto, CA, USA, October 15-18, 2021, Proceedings, Lecture Notes in Computer Science*, vol 13154. Springer, pp 127–138, https://doi.org/10.1007/978-3-030-93758-4_14

McDonald, N., Schoenebeck, S., & Forte, A. (2019). Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice. *Proc ACM Hum-Comput Interact, 3*(CSCW).https://doi.org/10.1145/3359174

McDowell, J. (1979). Virtue and reason. *The monist, 62*(3), 331–350.

Meo, R., Nai, R., & Sulis, E. (2022). Explainable, interpretable, trustworthy, responsible, ethical, fair, verifiable AI... what's next? In Chiusano, S., Cerquitelli, T., & Wrembel, R. (eds) *Advances in Databases and Information Systems - 26th European Conference*, ADBIS 2022, Turin, Italy, September 5-8, 2022, Proceedings, Lecture Notes in Computer Science, vol 13389. Springer, pp 25–34, https://doi.org/10.1007/978-3-031-15740-0_3

Metcalf, J., & moss, e., & boyd, d. (2019). Owning ethics: Corporate logics, silicon valley, and the institutionalization of ethics. *Social Research, 86*, 449–476. https://doi.org/10.1353/sor.2019.0022

Mill, J. S. (1979). *Utilitarianism*. Hackett Publishing.

Miller, T. (2023). Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New York, NY, USA*, FAccT '23, p 333-342, https://doi.org/10.1145/3593013.3594001

Mittelstadt, B. D. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence, 1*(11), 501–507. https://doi.org/10.1038/S42256-019-0114-4

Mittelstadt, B.D., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. In danah boyd, Morgenstern, J.H. (eds) *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA*, January 29-31, 2019. ACM, pp 279–288, https://doi.org/10.1145/3287560.3287574

Mökander, J., & Axente, M. (2023). Ethics-based auditing of automated decision-making systems: intervention points and policy implications. *AI society, 38*(1), 153–171. https://doi.org/10.1007/S00146-021-01286-X

Morley, J., Elhalal, A., Garcia, F., et al. (2021). Ethics as a service: A pragmatic operationalisation of AI ethics. *Minds Mach, 31*(2), 239–256. https://doi.org/10.1007/S11023-021-09563-W

Morley, J., Kinsey, L., Elhalal, A., et al. (2023). Operationalising AI ethics: barriers, enablers and next steps. *AI Soc, 38*(1), 411–423. https://doi.org/10.1007/S00146-021-01308-8

Morris, M., Song, E., Rajesh, A., et al. (2023). Ethical, legal, and financial considerations of artificial intelligence in surgery. *Am Surg, 89*(1), 55–60. https://doi.org/10.1177/00031348221117042. arXiv:2022 Aug 17

Muralidharan, A., Savulescu, J., & Schaefer, G.O. (2024). Ai and the need for justification (to the patient). *Ethics Inf Technol*, *26*(1) ,16. https://doi.org/10.1007/s10676-024-09754-w, epub 2024 Mar 4. PMID: 38450175; PMCID: PMC10912120

Nannini, L., Balayn, A., & Smith, A.L. (2023). Explainability in AI policies: A critical review of communications, reports, regulations, and standards in the eu, us, and UK. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2023, Chicago, IL, USA*, June 12-15, 2023. ACM, pp 1198–1212, https://doi.org/10.1145/3593013.3594074

Narayanan, D., & Tan, Z. M. (2023). Attitudinal tensions in the joint pursuit of explainable and trusted AI. *Minds Mach, 33*(1), 55–82. https://doi.org/10.1007/s11023-023-09628-y

Nicodeme, C. (2020). Build confidence and acceptance of ai-based decision support systems - explainable and liable AI. In *13th International Conference on Human System Interaction, HSI 2020, Tokyo, Japan*, June 6-8, 2020. IEEE, pp 20–23, https://doi.org/10.1109/HSI49210.2020.9142668

Nussbaum, M. (1988). Non-relative virtues: an aristotelian approach. *Midwest Studies in Philosophy, 13*(1), 32–53.

Nyrup, R., & Robinson, D. (2022). Explanatory pragmatism: a context-sensitive framework for explainable medical AI. *Ethics and Information Technology, 24*(1), 13. https://doi.org/10.1007/S10676-022-09632-3

Oakley, J. (1996). Varieties of virtue ethics. *Ratio, 9*(2), 128–152.

Okolo, C.T. (2023). Towards a praxis for intercultural ethics in explainable AI. CoRR arXiv:2304.11861. https://doi.org/10.48550/ARXIV.2304.11861

Okolo, C.T., Dell, N., & Vashistha, A. (2022). Making ai explainable in the global south: A systematic review. In *Proceedings of the 5th ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies. Association for Computing Machinery, New York, NY, USA*, COMPASS '22, p 439-452, https://doi.org/10.1145/3530190.3534802

O'Neill, O. (1975). *Acting on principle: An essay on Kantian ethics*. Columbia University Press.

van Otterlo, M., & Atzmueller, M. (2020). A conceptual view on the design and properties of explainable AI systems for legal settings. In Rodríguez-Doncel, V., Palmirani, M., Araszkiewicz, M., et al. (eds) *AI Approaches to the Complexity of Legal Systems XI-XII - AICOL International Workshops 2018 and 2020:*

AICOL-XI@JURIX 2018, AICOL-XII@JURIX 2020, XAILA@JURIX 2020, Revised Selected Papers, Lecture Notes in Computer Science, vol 13048. Springer, Luxembourg, pp 143–153, https://doi.org/10.1007/978-3-030-89811-3_10

Parfit, D. (1984). *Reasons and persons*. Oxford University Press.

Pastaltzidis, I., Dimitriou, N., Quezada-Tavarez, K., et al. (2022). Data augmentation for fairness-aware machine learning: Preventing algorithmic bias in law enforcement systems. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea*, June 21 - 24, 2022. ACM, Seoul, Republic of Korea, June 21 - 24, 2022, pp 2302–2314,https://doi.org/10.1145/3531146.3534644

Phillips, P.J., Hahn, C., Fontana, P., et al. (2021). Four principles of explainable artificial intelligence. Tech. rep., NIST, https://doi.org/10.6028/NIST.IR.8312. https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=933399

do Prado, J. C. S., & Cappelli, C. (2010). Software transparency. *Bus Inf Syst Eng, 2*(3), 127–139. https://doi.org/10.1007/s12599-010-0102-z

Railton, P. (1984), Alienation, consequentialism, and the demands of morality. *Philosophy & Public Affairs* pp 134–171.

Robbins, S. (2019). A misdirected principle with a catch: Explicability for AI. *Minds Mach, 29*(4), 495–514. https://doi.org/10.1007/S11023-019-09509-3

Ross, W. D. (1930). *The right and the good*. Clarendon Press.

Saeed, W., & Omlin, C. W. (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-based systems, 263*, 110273. https://doi.org/10.1016/J.KNOSYS.2023.110273

Samek, W., Wiegand, T., & Müller, K. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. CoRR arXiv:1708.08296

Scheffler, S. (1982). *The rejection of consequentialism* (7th ed.). Oxford University Press.

Schmid, U., & Wrede, B. (2022). What is missing in XAI so far? *Künstliche Intell, 36*(3), 303–315. https://doi.org/10.1007/S13218-022-00786-2

Sen, A. (1979). Utilitarianism and welfarism. *The Journal of Philosophy, 76*(9), 463–489.

Shafer-Landau, R. (2012). *Ethical theory: an anthology*. John Wiley & Sons.

Shamsabadi, A.S.., Yaghini, M., Dullerud, N., et al. (2022). Washing the unwashable : On the (im)possibility of fairwashing detection. In Koyejo, S., Mohamed, S., Agarwal, A., et al. (eds) *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022*, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, http://papers.nips.cc/paper_files/paper/2022/hash/5b84864ff8474fd742c66f219b2eaac1-Abstract-Conference.html

Sherman, N. (1989). *The fabric of character: Aristotle's theory of virtue*. Oxford University Press.

Sibai, F.N. (2020). AI crimes: A classification. In *2020 International Conference on Cyber Security and Protection of Digital Services, Cyber Security 2020, Dublin, Ireland*, June 15-19, 2020. IEEE, pp 1–8, https://doi.org/10.1109/CYBERSECURITY49315.2020.9138891

Sidgwick, H. (1907). *The methods of ethics* (7th ed.). Hackett Publishing.

Slack, D., Hilgard, S., Jia, E., et al. (2020). Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. In [116], pp 180–186, https://doi.org/10.1145/3375627.3375830

Slote, M. (1992). *From morality to virtue*. Oxford University Press.

Smart, J. J. C., & Williams, B. (1973). *Utilitarianism: For and against*. Cambridge University Press.

Sokol, K., & Flach, P. (2020). Explainability fact sheets: A framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New York, NY, USA*, FAT* '20, p 56-67, https://doi.org/10.1145/3351095.3372870. https://doi-org.ezbusc.usc.gal/10.1145/3351095.3372870

Solomon, M. Z. (2005). Realizing bioethics' goals in practice: ten ways "is'' can help "ought''. *Hastings Center Report, 35*(4), 40–47.

Standard for XAI - eXplainable AI Working Group IEEE Computational Intelligence Society/ Standards Committee (IEEE CIS/SC/XAI WG) (2024) Ieee cis/sc/xai wg p2976 - standard for xai - explainable artificial intelligence - for achieving clarity and interoperability of ai systems design. https://standards.ieee.org/ieee/2976/10522/

Stepin, I., Alonso, J. M., Catalá, A., et al. (2021). A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access, 9*, 11974–12001. https://doi.org/10.1109/ACCESS.2021.3051315

Sullivan, E., & Verreault-Julien, P. (2022). From explanation to recommendation: Ethical standards for algorithmic recourse. In Conitzer, V., Tasioulas, J., Scheutz, M., et al. (eds) *AIES '22: AAAI/ACM Conference on AI, Ethics, and Society, Oxford, United Kingdom*, May 19 - 21, 2021. ACM, pp 712–722, https://doi.org/10.1145/3514094.3534185

Theodorou, A., Wortham, R. H., & Bryson, J. J. (2017). Designing and implementing transparency for real time inspection of autonomous robots. *Connect Sci, 29*(3), 230–241. https://doi.org/10.1080/09540091.2017.1310182

Theunissen, M., & Browning, J. (2022). Putting explainable AI in context: institutional explanations for medical AI. *Ethics and Information Technology, 24*(2), 23. https://doi.org/10.1007/S10676-022-09649-8

Tsamados, A., Aggarwal, N., Cowls, J., et al. (2022). The ethics of algorithms: key problems and solutions. *AI Soc, 37*(1), 215–230. https://doi.org/10.1007/S00146-021-01154-8

Vainio-Pekka, H., Agbese, M. O. O., Jantunen, M., et al. (2023). The role of explainable ai in the research field of ai ethics. *ACM Trans Interact Intell Syst, 13*(4). https://doi.org/10.1145/3599974

van der Waa, J., Verdult, S., van den Bosch, K., et al. (2021). Moral decision making in human-agent teams: Human control and the role of explanations. *Frontiers Robotics AI, 8*, 640647. https://doi.org/10.3389/FROBT.2021.640647

Waefler, T., & Schmid, U. (2021). Explainability is not enough: Requirements for human-ai-partnership in complex socio-technical systems. In *Proceedings of the 2nd European Conference on the Impact of Artificial Intelligence and Robotics (ECIAIR 2020)* / ed. by Florinda Matos. Lissabon: ACPIL, 2020, S. 185-194. - ISBN 9781912764747. Otto-Friedrich-Universität, Bamberg, pp 185–194, https://doi.org/10.20378/irb-49775, jahr der Erstpublikation: 2020

Wagner, B. (2018a), Ethics As An Escape From Regulation. From "Ethics-Washing" To Ethics-Shopping?, Amsterdam University Press, Amsterdam, pp 84–89. https://doi.org/10.1515/9789048550180-016

Wagner, B. (2018b). Ethics As An Escape From Regulation. From "Ethics-Washing" To Ethics-Shopping?, Amsterdam University Press, Amsterdam, pp 84–89. https://doi.org/10.1515/9789048550180-016

Weller, A. (2019). Transparency: Motivations and challenges. In Samek, W., Montavon, G., Vedaldi, A., et al. (eds) *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Lecture Notes in Computer Science*, vol 11700. Springer, p 23–40, https://doi.org/10.1007/978-3-030-28954-6_2

Wu, S.T., Demetriou, D., & Husain, R.A. (2023). Honor ethics: The challenge of globalizing value alignment in AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and*

Transparency, FAccT 2023, Chicago, IL, USA, June 12-15, 2023. ACM, pp 593–602,https://doi.org/10.1145/3593013.3594026

Zerilli, J., Knott, A., Maclaurin, J., et al. (2019). Transparency in algorithmic and human decision-making: is there a double standard? *Philosophy & Technology, 32*(4), 661–683. https://doi.org/10.1007/s13347-018-0330-6

Zhang, J., & Yu, H. (2022). A methodological framework for facilitating explainable AI design. In: Meiselwitz, G. (ed) *Social Computing and Social Media: Design, User Experience and Impact - 14th International Conference, SCSM 2022, Held as Part of the 24th HCI International Conference, HCII 2022, Virtual Event, June 26 - July 1, 2022, Proceedings, Part I, Lecture Notes in Computer Science*, vol 13315. Springer, Online, pp 437–446, https://doi.org/10.1007/978-3-031-05061-9_31

Zhang, Y., Liao, Q.V., & Bellamy, R.K.E. (2020). Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In [81], pp 295–305.https://doi.org/10.1145/3351095.3372852

Zhou, T., Sheng, H., & Howley, I. (2020). Assessing post-hoc explainability of the BKT algorithm. In: [116], pp 407–413, https://doi.org/10.1145/3375627.3375856

Zicari, R. V., Brodersen, J., Brusseau, J., et al. (2021). Z-inspection®: A process to assess trustworthy ai. *IEEE Transactions on Technology and Society, 2*(2), 83–97. https://doi.org/10.1109/TTS.2021.3066209

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.