Towards Use-Based Ethics Audits of LLM-Based Advice-Chatbots

Tobias Christoph
tobias.christoph@tuwien.ac.at
Institute of Logic and Computation
TU Wien
Vienna, Austria

Kees van Berkel
kees.van.berkel@tuwien.ac.at
Institute of Logic and Computation
TU Wien
Vienna, Austria

Katta Spiel
katta.spiel@tuwien.ac.at
Crip Collective || HCI Group
TU Wien
Vienna, Austria

Abstract

This position paper argues for a stakeholder use-based approach to designing ethics-based AI audits of LLM-based advice-chatbots. The proposed Use-Based Ethics Audit (UBEA) methodology facilitates auditing as a continuous process integrated into all phases of an AI's development cycle. Its key characteristics are the involvement of stakeholders in identifying ethical focus areas (EFAs), and in designing a testable environment for the identified EFAs. In particular, the UBEA methodology seeks to bring closer together two key questions preceding any audit practice: what are the relevant values involved and how to translate these values to a testable environment? We identify four desiderata for audits in general and discuss how the UBEA methodology seeks to address these requirements. We reflect on how context-sensitive approaches, such as LLM-based advice chatbots, benefit from a 'stakeholders in the loop' approach and define some key challenges and future work directions.

CCS Concepts

• **Human-centered computing** → **HCI theory, concepts and models**; **HCI design and evaluation methods**; **Natural language interfaces**.

Keywords

Ethics-based audits, Ethical AI, Large Language Models, Human-Centered Design, Human-AI Interaction, Embedded ethics.

1 Introduction

As Large Language Model (LLM)-based chatbots increasingly appear in advice-giving roles in public services, the ethical integrity of such systems becomes ever more critical. While technical and functional aspects of technologies benefit from clearly defined and measurable system requirements, ethical principles remain largely general and abstract [20, 35]. Technology providers thereby benefit from acquiring expert guidance in making ethics more tangible and testable [50]. Additionally, current efforts to integrate ethics frequently focus on post-development, which, although a valuable practice, overlooks the importance of addressing ethical considerations during the design and implementation phases of LLM-based chatbots [5]. When ethics-based audits are treated as an afterthought, the identification and mitigation of ethical shortcomings is inefficient and makes it difficult to assess ethical performance systematically and holistically [30, 37].

Nevertheless, LLM-based advice-chatbots have a shaping impact on society and, thus, need to be ethically aligned – or at least actively contrasted – with the values governing the social environments they are embedded in (cf. [17]). Prior to designing the appropriate auditing methods for such socio-technical systems, proper understanding of the involved values[1] and ethical principles is required. This requirement comes with two fundamental challenges:

- C1 *What are the relevant values involved in the (envisioned) AI application?*
- C2 *How to translate these values to a practical, testable environment?*

This position paper argues for a Use-Based approach to Ethics Audit (UBEA) of LLM-based chatbots to address challenges C1 and C2. In brief, UBEAs involve stakeholders throughout the development phase of an LLM-based chatbot. These stakeholders include, on the one hand, potential end-users – such as clients or individuals who will rely on the chatbots advice – and, on the other hand, a broader range of relevant actors, for instance employees in the company in which the technology will be deployed or representatives of indirect [16] or non-human [40] stakeholders. All are involved both for the identification of Ethical Focus Areas (EFAs) that serve as indicators for relevant values and for interactively shaping a workable testing environment for those EFAs. We argue that UBEAs are worth investigating since they address key desiderata of AI ethics audits and align with effective approaches developed in the field of human-centered interaction.

The UEBA method aims at integrating auditing into the design phase of LLM-based chatbots. By involving stakeholder groups, we seek to explore how ethical considerations can be systematically embedded into technologies from the outset. This approach promotes a proactive and critical-auditing methodology to guide development decisions towards integrating ethical requirements, thus, incorporating auditability into a system's lifecycle.[2] We seek to create a framework that thereby not only assesses ethical performance but makes holistic ethics evaluation approachable to industry partners – thus helping to contribute to broader, systemic change in AI-technologies [18].

We stress that although our method has the potential to generalize to any user-centered AI technology, we focus on advice-giving chatbots asas the act of giving advice is inherently normative and ethically charged [38]. Their interactive nature has a direct impact on users and their environments, requiring robust ethical considerations. Most common technologies in this domain incorporate LLMs and use Retrieval-Augmented Generation (RAG). Such chatbots are highly context-sensitive, operating differently depending on what their use case and area of deployment is. With broad applications

[1] In what follows, we mainly use the term 'values' and leave 'ethical principles' implicit.
[2] Our approach is not concerned with developing a general-purpose LLM. Rather, it focuses on the identification of testable aspects of the domain-specific LLM-based chatbot applications. Ultimately, decisions regarding how to transform or adjust a system in response to these insights remain the responsibility of the developers and maintainers of the chatbot.

HEAL@CHI'25,
2025. ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM

across various sectors, including career counseling [22], customer service [15], and healthcare [4, 32], the advice-giving domain is still sufficiently concrete to allow the development of a structured method. These considerations provide substantive reasons for a use-based approach. Future work needs to be directed to investigating the limits of generalizing our approach.

The remainder of this paper is organized as follows: In Section 2 we provide further background and identify four desiderata for audits of LLM-based advice-chatbots. We outline the key aspects of Use-Based Ethics Audit in Section 3 and provide some methodological considerations. Since this work argues for the investigation of UBEAs we close in Section 4 with a critical reflection of our proposal, highlighting some key challenges.

2 Background

Due to its pressing importance, we see an increasing number of auditing methods being developed for AI systems (including LLMs and LLM-based services). We identify the following central (non-exhaustive) desiderata for ethics audit methodology of AI. The methodology:

- D1 follows a structured method for identifying key values and ethical principles involved in the AI;
- D2 provides a procedure to translate the identified values into testable representations of these values;
- D3 embeds ethics in the entire development phase of the AI;
- D4 adopts a multi-scaled ethics approach, involving representatives of the various stakeholders involved.

In many such contexts, the values to be tested are assumed given or considered straightforward (cf. bias considerations).[3] However, as expressed in C1, it is often a non-trivial task to identify the relevant values/ethical principles involved. A structured approach, which avoids ad hoc considerations of values thus risking ethical shortfalls is key (cf. [20]). Even when such values are known, the central challenge remains how to translate values/ethical principles that are abstract and general to a practical, testable environment. Furthermore, to ensure that ethics audits are employed holistically, they should not only examine the technical parts of a system's workings but also organizational structures and governance [37, 45, 46]. Although regulatory bodies are implicated with enforcing accountability regarding ethical compliance, the actual control of technology creation lies with the developers. Even with extrinsic incentives for industry partners to uphold ethical standards (i.e., via laws and policies), the use of additional resources as well as the absence of methods and existing regulatory gaps make such practices challenging. These considerations are captured by desiderata D1 and D2.

D3 follows, what is called, the 'Embedded Ethics' approach, promoting the involvement of ethicists throughout the AI's entire development process [34, 35]. The reason for doing so is to avoid ad-hoc solutions when ethical concerns are identified in the post-development phase. Furthermore, involving ethicists yields an ongoing discussion throughout the design and implementation processes, documenting ethical decision-making that facilitates comprehensive auditing at later stages. Embedded ethics anticipates rather than responds to social and ethical frictions of AI and enables bridging current regulatory gaps [35].

D4 requires a multi-scale ethics approach to AI [48], which orders and analyzes the effects of technologies by investigating the technology at hand from different perspectives (i.e. individuals, communities, institutions, nations, global, and over time). By ordering the ethical effects of technology by their scale, a comparison of effects across scales is facilitated. The approach, furthermore, ensures that various stakeholders are represented in the analysis of transformative powers such as AI technology.

To address these desiderata, our strategy is twofold. First, it encourages and enables technology providers to integrate ethics from the design stage onward and thereby place ethical considerations within the system and its lifecycle. Second, it advocates for continuous auditing using a stakeholder-centered approach. We build on a growing body of literature that highlights the potential and importance of participatory or user-involved methods to address algorithmic issues. While existing work has explored ways users can help detect problematic behaviors [47] and has proposed participatory mechanisms for democratizing technology governance [25], they are often focused on post-deployment stages and one-time assessments. They demonstrate users' promising role in detecting issues – such as biases or unintended consequences through examples [9] – but do not yet offer testable procedures for continuous auditing practices [11, 24]. In response to that, and since which values and exactly how they are crucial to a system highly depends on its context [42], we propose defining Ethical Focus Areas (EFA) that serve as concrete and tangible highlights of ethical values. To identify these, we suggest directly involving technology stakeholders in the process as their insights are crucial to align the values embedded in an LLM-based chatbot with their needs [18]. Just as User-Centred Design (UCD) leverages users' input during the design and prototyping phases to shape technical requirements, the stakeholders' expert feedback on potential ethical conflicts highlights critical concerns and thereby areas to audit. In Section 3.2, we discuss how UBEAs will address all four desiderata.

3 Use-Based Ethics Audits

Use-Based Ethics Audit (UBEA) is a participatory approach to auditing LLM-based advice-chatbots, grounded in the involvement of skateholders, defined as any individuals who are envisioned to seek advice or guidance from the chatbot in the future and other actors related to the deployed technology and its impacts. Ethics audits fundamentally address two key aspects: *How* ethical concerns are assessed (referring to the methods used during an assessment phase to evaluate a system) and *What* to even evaluate for (relating to the values that should guide this ethical assessment). With UBEA, we build on attempts to make ethical values graspable by defining them at an early development stage [52]. Similarly, existing approaches aim to make the values more tangible and comprehensible to technology providers [23, 49]. UBEA integrates stakeholder involvement in defining both aspects, the *What* (C1) and the *How* (C2), to ensure that the ethical principles are both identified and translated into concrete and testable criteria.

[3] See the work of Laine et al [28] for a study of key ethical principles recurring in the literature on ethics audits.

The core element to this is the introduction of Ethical Focus Areas (EFAs) which serve as key indicators of relevant ethical concerns in a system's development, dependent on its context and use-case. UBEA incorporates EFAs informed by stakeholders' lived experiences and practical concerns, thereby naturally shaping the testing environment of an audit.

Ethics audits can be broadly structured into three phases: an *understanding* phase, which examines both the system and the values of technology providers; an *assessment* phase, that evaluates the system itself along with organizational structures behind its development; and a *recommendation* phase, in which actionable suggestions for improvements are formulated. Withing such a composition, EFAs serve as a bridge between the understanding and the actual assessment, designing the succeeding assessments by guiding the audit process through stakeholder-guided real-world areas of concern.

3.1 Towards a Methodology

Foremost, in ethics audits, a cooperation between a company and the auditing party must be established to facilitate a deeper insight into a system's design, technology, and governance to allow holistic auditing [36, 46]. Rather than being a punitive measure, auditing should be a collaborative effort. Moreover, relying purely on black-box audit limits the identification of root risks and the assessment of the technology [6]. With these considerations in mind, our UBEA framework is structured into two parts, as outlined in Figure 1. It requires a product or its prototype – such as an advice-giving chatbot – to be employed as a preliminary foundation for the ethical assessment and refinement. The process begins by identifying relevant stakeholders who will contribute to the audit. Drawing from User-Centered Design (UCD), ensuring a diverse and representative stakeholder base is vital as it prevents ethical shortcomings and includes multiple perspectives [51]. Best practices include stakeholder mapping to identify directly and indirectly affected parties, recruitment through targeted outreach and snowball sampling [7, 26]. Additionally, effective communication to set expectations and confirm stakeholders understand their role is key in allowing them to establish EFAs [11].

Once a stakeholder group is set, EFAs are identified through established participatory methods such as structured workshops, group techniques and peer discussions, experience sampling, or think-aloud interviews [8, 19, 43]. These activities – corresponding to Part 1 of Figure 1 – guide stakeholders to uncover ethical issues by interacting with a system or its prototype and drawing on their lived experiences. Through critical questioning, scenarios that reveal unintended consequences – such as gaps in ethical design or safeguards – are uncovered. For instance, diary studies can be applied to define concrete observable ethical concerns as they emerge [11]. Prompting stakeholders with (falsification) questions challenges a system's assumptions. For instance, asking "Under what conditions could EFA X lead to unintended harm?" or "Describe scenarios where EFA X was violated. What were the consequences?", compels stakeholders to think beyond expected use cases and actively points out potential ethical harms related to their experiences. These insights are then coded into scenarios, structured similar to ethical user stories [10, 21], which represent the lived examples behind EFAs. The Delphi method, through iterative rounds of feedback, redefines and promotes a consensus on EFAs while guaranteeing that no single perspective dominates the process [39, 53]. Incorporating frameworks such as multi-dimensional STEEPED approach or Failure Modes and Effects Analysis (FMEA) [2, 29] enhances this approach's extensiveness.

The process transitions into a second active phase, outlined in Part 2 of Figure 1. Here, stakeholders assess and refine EFAs based on their real-world related interaction with a system. This phase not only probes limits of ethical safeguards but its findings also enable iterative improvements in both a system's design and the auditing framework of its ethics. Contextualizing stakeholder experiences and guaranteeing scenario-based validation is crucial for the specific auditing methods of EFAs. Assessment criteria and practices are intrinsically revealed, corresponding to the *How* part (C2). As different EFAs, and the contexts in which they arise, demand different auditing methods, the wide range of research-based available methods for auditing LLMs can be utilized [1, 3, 27, 31, 33, 54]. For instance, uncovering algorithmic bias may call for techniques distinct from those used to evaluate accessibility, explainability or the development's underlying governance processes [37]. In short, the *What* hereby directly informs the *How*. Mapping appropriate methods to the specific EFAs results in a well-defined testing environment. Through the actual usage context, EFAs have the strong potential to confront edge cases and potential failures for a more robust and adaptive auditing.

Establishing a well-structured EFA is achieved by systematically defining its different components. Besides the ethical focus area of concern and an assessment methodology (e.g. scenario-based prompting, benchmarks, document reviews), EFAs are associated with the broader ethical principles (e.g. fairness, transparency, autonomy) [28]. Furthermore, an EFA requires an assigned typology, e.g. behavior-linked for observable system actions; data-driven, addressing issues related to data sourcing and quality; governance and process oriented, examining organisational policies and oversights; stakeholder-interaction based, considering the UX design. Classifying EFAs with these categorizations enables more structured audits, permitting similar concerns to be grouped together and relevantly related audit techniques to be applied efficiently. It further allows an overview of which ethical values are *covered* and which need more a targeted investigation.

Similarly to grouping, the impact of EFAs can be ranked based on risk assessment methodologies such as weighted scoring, Pareto analysis, or risk matrices [13, 14, 44]. This allows to prioritize EFAs, as some may present greater risks or require urgent attention based on their impact severity, allowing technology providers to focus their development efficiently. Ultimately, the audit can proceed into the assessment phase where the EFAs ethical values be systematically tested with their corresponding methods.

Consistent with principles of human-centered design [12], the process is inherently iterative – as represented by arrows looping back to the two phases. EFAs are dynamic constructs that evolve in response to changes in the system, stakeholder perspectives, and emerging ethical challenges. As the technology progresses through its lifecylce, previously identified EFAs may require adaption and new EFAs may be identified. A continuous engagement with ethical concerns through EFAs is essential in embedding ethics.

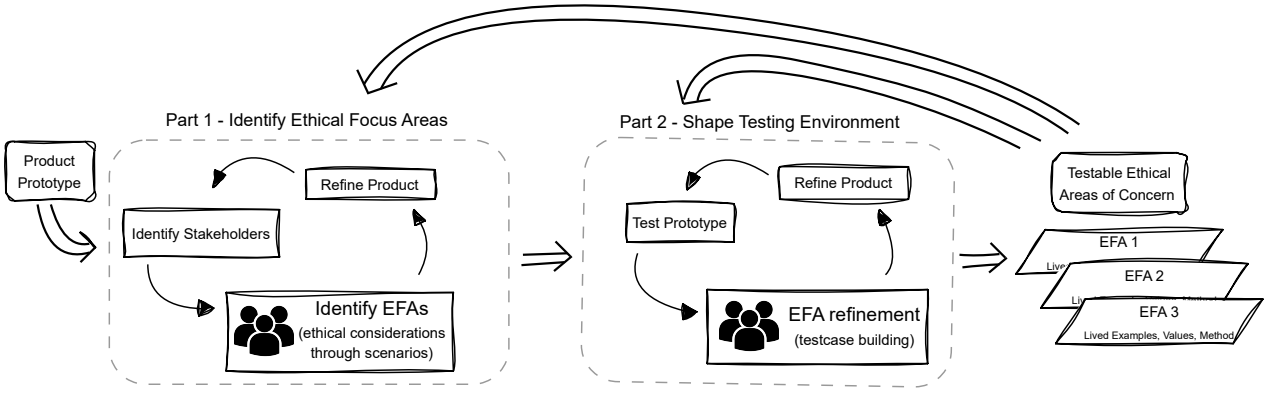

Figure 1: The UBEA methodology, consisting of two parts: (1) Identifying ethical focus areas and stakeholders and (2) interactively shaping the testing environment. Both serve to yield a continues feedback process in which ethical considerations are identified, integrated, and tested. $\Longrightarrow$ denote transitions between distinct phases and illustrate how testable EFAs feed back into Part 1 and Part 2 to support iterative refinement and product testing. $\longrightarrow$ represents internal reflective cycles withing a phase.

3.2 How UBEA Aims to Address the Desiderata

Desiderata D1 and D2 are simultaneously addressed by the UBEA methodology: by taking stakeholders and use-cases as a departure point, our methodology takes a pragmatic stance in identifying key values and ethical considerations. By focusing on how stakeholders experience the ethical dimensions of an AI technology, the identification of values is closer connected to testability (e.g. by means of falsification procedures), thus bringing together the *What* question (of C1 and D1) and the *How* question (of C2 and D2).

D3 is directly accommodated by the UBEA methodology since it constitutes an interactive audit involved at various stages of the development process. We stress here that, although UBEA is stakeholder-focused, to avoid ethical pitfalls a close collaboration is required between developers, ethicists, and the involved stakeholders. In particular, these parties aim at jointly shaping the testing environment.

D4 is addressed by involving stakeholders in identifying EFAs, i.e. ethical dimensions, and relevant stakeholders. In this way, developers, ethicists, and stakeholders can jointly identify and recruit the relevant involved parties across various scales that need to be dynamically included in the iterative UBEA process (cf. Fig. 1).

4 Critical Reflection and Challenges

While the UBEA methodology offers a structured approach, several challenges must be addressed.

First, even though EFAs can help map values into concrete considerations and provide a more testable framework, measuring ethical performance remains inherently difficult. In particular, whether testable environments adequately approximate abstract values/ethical principles requires ongoing critical evaluation.

Second, a key concern of UBEA is its resource intensity, since workshops, iterative feedback, and stakeholder engagement in general requires significant effort. Consequently, effectively cooperating in UBEA could be challenging for smaller companies. A potential way to streamline or automate certain processes could make UBEA more feasible while remaining the upsides of its participatory approach. Embedding ethics early can also reduce costly post-development risks, required audits, and related adjustments. Furthermore, audits are often considered regulatory burdens and fear of potential PR backlashes when ethical shortcomings get exposed could exacerbate industry resistance. Yet, an adaption of UBEA can be framed as a competitive advantage, underscoring proactive and responsible technology providers. Structured ethics auditing could be positioned similarly to Formal Verification Testing (FVT) which is already a standard for ensuring system reliability and accountability.

Third, while ensuring stakeholder diversity is critical in a use-based approach, a focus on stakeholder groups may create ethical shortcomings in the development phase. For this reason, we stress the need to enhance UBEA with expert validation of the EFAs and a transparent selection process. Such measures can additionally help mitigate bias and promote trustworthiness efficiently.

Lastly, translating the approach of UBEA and EFAs to other AI technologies requires additional investigations, but their design is technically adaptable. By conducting further research, case studies, and industry collaboration, we aim to find answers to these obstacles. Represent a promising step toward making ethical consideration a natural and testable part of system development and a suitable baseline for addressing these challenges.

5 Conclusion

In this position paper, we argue for a stakeholder-based approach to designing ethics-based AI audits. The resulting Use-Based Ethics Audit (UBEA) methodology facilitates auditing as a continuous process integrated into all phases of an AI's development cycle. We argue that context-sensitive approaches, such as LLM-based advice chatbots, benefit from a 'stakeholders in the loop' approach since it brings together questions on *What* to audit and *How* to audit, seeking to bridge the gap between abstract values/ethical principles and practicable criteria that can guide ethical assessment. For this purpose, we introduce Ethical Focus Areas (EFAs), which

provide for a testable environment of relevant values. Nevertheless, we stress that the design of UBEAs relies on ongoing interaction between stakeholders, developers, and ethicists to mitigate potential ethical shortfalls.

By involving stakeholders directly in the design process of ethics auditing, our approach not only enhances ethical oversight, but also makes it inherently practical for companies and technology providers to adopt. By integrating UBEA during the design phase, organizations can embed ethical considerations into their development workflow from the outset. This establishes *ethics-by-design*, rather than treating ethics and audits as external compliance hurdles. While maintaining auditability, UBEA allows industry stakeholders to systematically anticipate and address ethical concerns from the start, guaranteeing testability and proactive ethical governance (which is particularly promising in light of regulatory gaps). UBEA can therefore serve as a self-auditing tool for technology providers without solely relying on external third-party evaluations (cf. [41]). We conjecture that such an approach makes it easier and less resource-intensive to conduct a full audit on a deployed AI product.

Possible adaptions and the feasibility of UBEAs in other AI based technologies can be further explored. Developing and investigating this approach allows us to pinpoint these strengths, but additional research is needed to explore its practical implementation. Case studies and direct engagement with industry partners – consulting them on whether our approach effectively makes ethical values more accessible through EFAs – are needed to refine our method.

References

[1] Maryam Amirizaniani, Elias Martin, Tanya Roosta, Aman Chadha, and Chirag Shah. 2024. Auditllm: a tool for auditing large language models using multiprobe approach. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 5174–5179.

[2] Valentina Amuso and Lieve Van Woensel. 2025. Reflections on applying systems thinking to stakeholder mapping: the stoa unit at the european parliament. *Global Policy*.

[3] Leif Azzopardi and Yashar Moshfeghi. 2024. Prism: a methodology for auditing biases in large language models. *arXiv preprint arXiv:2410.18906*.

[4] Iqra Basharat and Subhan Shahid. 2024. Ai-enabled chatbots healthcare systems: an ethical perspective on trust and reliability. *Journal of Health Organization and Management*.

[5] Kathrin Bednar and Sarah Spiekermann. 2024. The power of ethics: uncovering technology risks and positive value potentials in it innovation planning. *Business & Information Systems Engineering*, 66, 2, 181–201.

[6] Stephen Casper et al. 2024. Black-box access is insufficient for rigorous ai audits. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2254–2272.

[7] Elizabeth Chen, Cristina Leos, Sarah D Kowitt, and Kathryn E Moracco. 2020. Enhancing community-based participatory research through human-centered design strategies. *Health promotion practice*, 21, 1, 37–48.

[8] Tamlin Conner Christensen, Lisa Feldman Barrett, Eliza Bliss-Moreau, Kirsten Lebo, and Cynthia Kaschub. 2003. A practical guide to experience-sampling procedures. *Journal of Happiness Studies*, 4, 1, 53–78.

[9] Wesley Hanwen Deng, Wang Claire, Howard Ziyu Han, Jason I Hong, Kenneth Holstein, and Motahhare Eslami. 2025. Weaudit: scaffolding user auditors and ai practitioners in auditing generative ai. *arXiv preprint arXiv:2501.01397*.

[10] Christian Detweiler and Maaike Harbers. 2014. Value stories: putting human values into requirements engineering. In *REFSQ Workshops*. Vol. 1138, 2–11.

[11] Alicia DeVos, Aditi Dhabalia, Hong Shen, Kenneth Holstein, and Motahhare Eslami. 2022. Toward user-driven algorithm auditing: investigating users' strategies for uncovering harmful algorithmic behavior. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, 1–19.

[12] ISO DIS. 2009. 9241-210: 2010. ergonomics of human system interaction-part 210: human-centred design for interactive systems. *International Standardization Organization (ISO). Switzerland*, 2.

[13] Christian Enyoghasi and Fazleena Badurdeen. 2023. Sustainable product design decision-making through integrated risk likelihood and impact analyses. In *International Manufacturing Science and Engineering Conference*. Vol. 87233. American Society of Mechanical Engineers, V001T04A003.

[14] Seena Fazel and Achim Wolf. 2018. Selecting a risk assessment tool to use in practice: a 10-point guide. *BMJ Ment Health*, 21, 2, 41–43.

[15] Asbjørn Følstad, Cecilie Bertinussen Nordheim, and Cato Alexander Bjørkli. 2018. What makes users trust a chatbot for customer service? an exploratory interview study. In *Internet Science: 5th International Conference, INSCI 2018, St. Petersburg, Russia, October 24–26, 2018, Proceedings 5*. Springer, 194–208.

[16] Batya Friedman, David G Hendry, Alan Borning, et al. 2017. A survey of value sensitive design methods. *Foundations and Trends® in Human–Computer Interaction*, 11, 2, 63–125.

[17] Iason Gabriel and Vafa Ghazavi. 2022. The challenge of value alignment. In *The Oxford handbook of digital ethics*. Oxford University Press Oxford.

[18] Karine Gentelet and Sarit K Mizrahi. 2023. A human-centered approach to ai governance: operationalizing human rights through citizen participation. In *Human-Centered AI*. Chapman and Hall/CRC, 215–230.

[19] Karen Gravett. 2019. Story completion: storying as a method of meaning-making and discursive discovery. *International Journal of Qualitative Methods*, 18, 1609406919893155.

[20] Thilo Hagendorff. 2022. Blind spots in ai ethics. *AI and Ethics*, 2, 4, 851–867.

[21] Erika Halme, Marianna Jantunen, Ville Vakkuri, Kai-Kristian Kemell, and Pekka Abrahamsson. 2024. Making ethics practical: user stories as a way of implementing ethical consideration in software engineering. *Information and Software Technology*, 167, 107379.

[22] Deirdre Hughes et al. 2024. An international evidence review.

[23] 2021. Ieee standard model process for addressing ethical concerns during system design. *IEEE Std 7000-2021*, 1–82. doi:10.1109/IEEESTD.2021.9536679.

[24] Eunkyung Jo, Young-Ho Kim, Yuin Jeong, SoHyun Park, and Daniel A Epstein. [n. d.] Incorporating multi-stakeholder perspectives in evaluating and auditing of health chatbots driven by large language models. ().

[25] Leah R Kaplan, Mahmud Farooque, Daniel Sarewitz, and David Tomblin. 2021. Designing participatory technology assessments: a reflexive method for advancing the public role in science policy decision-making. *Technological Forecasting and Social Change*, 171, 120974.

[26] Shawal Khalid and Chris Brown. 2024. Exploring stakeholder challenges in recruitment for human-centric computing research. In *2024 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, 432–438.

[27] Hannah Rose Kirk, Yennie Jun, Filippo Volpin, Haider Iqbal, Elias Benussi, Frederic Dreyer, Aleksandar Shtedritski, and Yuki Asano. 2021. Bias out-of-the-box: an empirical analysis of intersectional occupational biases in popular generative language models. *Advances in neural information processing systems*, 34, 2611–2624.

[28] Joakim Laine, Matti Minkkinen, and Matti Mäntymäki. 2024. Ethics-based ai auditing: a systematic literature review on conceptualizations of ethical principles and knowledge contributions to stakeholders. *Information & Management*, 103969.

[29] Jamy Li and Mark Chignell. 2022. Fmea-ai: ai fairness impact assessment using failure mode and effects analysis. *AI and Ethics*, 2, 4, 837–850.

[30] Yueqi Li and Sanjay Goel. 2024. Making it possible for the auditing of ai: a systematic review of ai audits and ai auditability. *Information Systems Frontiers*, 1–31.

[31] Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

[32] Bingjie Liu and S Shyam Sundar. 2018. Should machines express sympathy and empathy? experiments with a health advice chatbot. *Cyberpsychology, Behavior, and Social Networking*, 21, 10, 625–636.

[33] Louise McCormack and Malika Bendechache. 2024. Ethical ai governance: methods for evaluating trustworthy ai. *arXiv preprint arXiv:2409.07473*.

[34] Stuart McLennan, Amelia Fiske, Leo Anthony Celi, Ruth Müller, Jan Harder, Konstantin Ritt, Sami Haddadin, and Alena Buyx. 2020. An embedded ethics approach for ai development. *Nature Machine Intelligence*, 2, 9, 488–490.

[35] Stuart McLennan, Amelia Fiske, Daniel Tigard, Ruth Müller, Sami Haddadin, and Alena Buyx. 2022. Embedded ethics: a proposal for integrating ethics into the development of medical ai. *BMC Medical Ethics*, 23, 1, 6.

[36] Jakob Mökander. 2023. Auditing of ai: legal, ethical and technical approaches. *Digital Society*, 2, 3, 49.

[37] Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. 2024. Auditing large language models: a three-layered approach. *AI and Ethics*, 4, 4, 1085–1115.

[38] Jenny Ng, Emma Haller, and Angus Murray. 2022. The ethical chatbot: a viable solution to socio-legal issues. *Alternative Law Journal*, 47, 4, 308–313.

[39] Chitu Okoli and Suzanne D Pawlowski. 2004. The delphi method as a research tool: an example, design considerations and applications. *Information & management*, 42, 1, 15–29.

[40] Athanasia Pouloudi, Reshma Gandecha, Christopher Atkinson, and Anastasia Papazafeiropoulou. 2004. How stakeholder analysis can be mobilized with actor-network theory to identify actors. *Information systems research: Relevant theory and informed practice*, 705–711.

[41] Emanuele Ratti and Mark Graves. 2025. A capability approach to ai ethics. *American Philosophical Quarterly*, 62, 1, 1–16.
[42] Maribeth Rauh et al. 2022. Characteristics of harmful text: towards rigorous benchmarking of language models. *Advances in Neural Information Processing Systems*, 35, 24720–24739.
[43] John Rieman. 1993. The diary study: a workplace-oriented research tool to guide laboratory efforts. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, 321–326.
[44] Dejan Ristić. 2013. A tool for risk assessment. *safety Engineering*, 3, 3, 121–127.
[45] Daniel S Schiff, Stephanie Kelley, and Javier Camacho Ibáñez. 2024. The emergence of artificial intelligence ethics auditing. *Big Data & Society*, 11, 4.
[46] Noah Schöppl, Mariarosaria Taddeo, and Luciano Floridi. 2022. Ethics auditing: lessons from business ethics for ethics auditing of ai. *The 2021 Yearbook of the Digital Ethics Lab*, 209–227.
[47] Hong Shen, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. 2021. Everyday algorithm auditing: understanding the power of everyday users in surfacing harmful algorithmic behaviors. *Proceedings of the ACM on Human-Computer Interaction*, 5, CSCW2, 1–29.
[48] Melanie Smallman. 2022. Multi scale ethics—why we need to consider the ethics of ai in healthcare at different scales. *Science and Engineering Ethics*, 28, 6, 63.
[49] S Spiekermann and T Winkler. 2020. Value-based engineering for ethics by design. *See: https://ssrn. com/abstract*, 3598911.
[50] Sarah Spiekermann et al. 2022. Values and ethics in information systems: a state-of-the-art analysis and avenues for future research. *Business & Information Systems Engineering*, 64, 2, 247–264.
[51] Brian Still and Kate Crane. 2017. *Fundamentals of user-centered design: A practical approach*. CRC press.
[52] Steven Umbrello and Ibo Van de Poel. 2021. Mapping value sensitive design onto ai for social good principles. *AI and Ethics*, 1, 3, 283–296.
[53] Marianne Wilson, David Brazier, Dimitra Gkatzia, and Peter Robertson. 2024. Participatory design with domain experts: a delphi study for a career support chatbot. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, 1–12.
[54] Xiang Zheng, Longxiang Wang, Yi Liu, Xingjun Ma, Chao Shen, and Cong Wang. 2025. Calm: curiosity-driven auditing for large language models. *arXiv preprint arXiv:2501.02997*.