

Challenges and Future Directions for Integration of Large Language Models into Socio-technical Systems

Torkamaan, H.; Steinert, S.; Pera, M.S.; Kudina, O.; Kernan Freire, S.; Verma, H.; Kelly, Sage ; Sekwenz, M.T.; Yang, J.; van Nunen, K.L.L.

DOI

[10.1080/0144929X.2024.2431068](https://doi.org/10.1080/0144929X.2024.2431068)

Publication date

2024

Document Version

Final published version

Published in

Behaviour and Information Technology

Citation (APA)

Torkamaan, H., Steinert, S., Pera, M. S., Kudina, O., Kernan Freire, S., Verma, H., Kelly, S., Sekwenz, M. T., Yang, J., van Nunen, K. L. L., Warnier, M., Brazier, F. M., & Oviedo-Trespalacios, O. (2024). Challenges and Future Directions for Integration of Large Language Models into Socio-technical Systems. *Behaviour and Information Technology*. <https://doi.org/10.1080/0144929X.2024.2431068>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.



Challenges and future directions for integration of large language models into socio-technical systems

Helma Torkamaan, Steffen Steinert, Maria Soledad Pera, Olya Kudina, Samuel Kernan Freire, Himanshu Verma, Sage Kelly, Marie-Therese Sekwenz, Jie Yang, Karolien van Nunen, Martijn Warnier, Frances Brazier & Oscar Oviedo-Trespalacios

To cite this article: Helma Torkamaan, Steffen Steinert, Maria Soledad Pera, Olya Kudina, Samuel Kernan Freire, Himanshu Verma, Sage Kelly, Marie-Therese Sekwenz, Jie Yang, Karolien van Nunen, Martijn Warnier, Frances Brazier & Oscar Oviedo-Trespalacios (10 Dec 2024): Challenges and future directions for integration of large language models into socio-technical systems, Behaviour & Information Technology, DOI: [10.1080/0144929X.2024.2431068](https://doi.org/10.1080/0144929X.2024.2431068)

To link to this article: <https://doi.org/10.1080/0144929X.2024.2431068>



© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



[View supplementary material](#)



Published online: 10 Dec 2024.



[Submit your article to this journal](#)



Article views: 242
















[View related articles](#)



[View Crossmark data](#)

Challenges and future directions for integration of large language models into socio-technical systems

Helma Torkamaan ^a, Steffen Steinert ^b, Maria Soledad Pera ^c, Olya Kudina ^b, Samuel Kernan Freire ^d, Himanshu Verma ^d, Sage Kelly ^{b,e}, Marie-Therese Sekwenz ^a, Jie Yang ^c, Karolien van Nunen ^b, Martijn Warnier ^a, Frances Brazier ^a and Oscar Oviedo-Trespalacios ^b

^aMulti-Actor Systems Department, Faculty of Technology, Policy, and Management, Delft University of Technology, Delft, the Netherlands; ^bDepartment of Values, Technology and Innovation, Faculty of Technology, Policy, and Management, Delft University of Technology, Delft, the Netherlands; ^cKnowledge and Intelligence Design, Faculty of Industrial Design Engineering, Delft University of Technology, Delft, the Netherlands; ^dWeb Information Systems – Department of Software Technology, Faculty of Electrical Engineering Mathematics, and Computer Science, Delft University of Technology, Delft, the Netherlands; ^eSchool of Psychology and Counseling, Queensland University of Technology, Kelvin Grove, Australia

ABSTRACT

Large Language Models (LLMs) are expected to significantly impact various socio-technical systems, offering transformative possibilities for improved interaction between humans and technology. However, their integration poses complex challenges due to the intricate interplay between societal structures, human behaviour, and technological innovation. This research explores these multifaceted challenges, emphasising the need for a human-centered approach in integrating LLMs to ensure that technological advancements are aligned with ethical standards and societal needs. Utilizing a structured methodology comprising a workshop, literature analysis, and expert collaborations, the study uses a multi-dimensional human-centered AI framework to guide the responsible integration of LLMs. Key insights include the importance of inclusive data, considering unintended consequences, maintaining privacy, and respecting intellectual property rights. The paper identifies and advocates for principles like human-in-the-loop, continuous longitudinal studies, proactive awareness campaigns, and regular audits to develop LLMs that are ethically sound, adaptable, and effectively integrated into various socio-technical systems, thus addressing user needs and broader societal impacts. The paper also underlines the importance of collaboration among academia, industry, and policymakers to develop LLMs that are ethically aligned, socially beneficial, and adaptable to future societal needs. The findings offer valuable insights into the strategic integration of LLMs, advocating for a broader research perspective beyond industrial motivations to fully understand and leverage LLMs in socio-technical landscapes.

ARTICLE HISTORY

Received 29 February 2024
Accepted 10 November 2024

KEYWORDS



AI; systems thinking; co-design; emerging technology; socio-technical systems

1. Introduction

1.1. Background

Generative Artificial Intelligence (AI) is an emerging technology enabling computers to process and create text, images, audio, and video. Large Language Models (LLMs) are a specific form of generative models built from vast data sources. They process human language as a series of symbols, likely to appear next to each other, and predict language patterns, extrapolating from their training database. LLMs have shown great promise at challenging text generation tasks, such as translation, summarisation, paraphrasing, and dialogue generation, and also perform well in text classification,

including sentiment analysis, among others (Dong et al. 2022; Hanqing Zhang et al. 2023; Li et al. 2021; Min et al. 2023). Currently, there are several LLMs in the market, such as GPT-4o (OpenAI 2024) and GPT-4 (OpenAI et al. 2024) (OpenAI models used for ChatGPT¹, a popular publicly available LLM-powered chatbot), Gemini models² (Team, Anil et al. 2024; Team, Georgiev, et al. 2024), Claude 3.5 Sonnet³, Meta Llama⁴ (Touvron, Lavril, et al. 2023; Touvron, Martin, et al. 2023), BLOOM (BigScience Workshop et al. 2023), Gemma 2⁵, OLMo⁶, Mistral Large⁷, Med-PaLM⁸ (Singhal et al. 2023), among others. LLMs are transforming our interaction with systems and technology, particularly within various socio-technical systems.

CONTACT Helma Torkamaan  h.torkamaan@acm.org  Jafflaan 5, 2628DX, Delft, the Netherlands

 Supplemental data for this article can be accessed online at <http://dx.doi.org/10.1080/0144929X.2024.2431068>.

© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

The role of LLMs in socio-technical systems can be compared to the role of language in society. Language is foundational in society, influencing every aspect of human interaction and life. It serves as the basis for communication, information sharing, coordination, and expressing needs, thoughts, feelings, and observations (Downes 1998; Trudgill 1991). Language also provides mental models to view the world and interact with others: cultural, social, cognitive, and conceptual frameworks, as well as political frameworks (Lakoff and Johnson 2008; Thomas and Wareing 1999). The ability of LLMs to generate content and respond to human commands creates new pathways for interactions between humans, socio-technical systems, and the technologies they represent, necessitating a thorough examination of their challenges and impacts.

LLMs can help streamline processes, enhance decision making, and personalise experiences within socio-technical systems through their language processing capabilities. In this context, socio-technical systems refer to the intricate interplay between societal infrastructure, human behaviour, and technological innovation (Kroes et al. 2006). For example, in healthcare, LLMs can aid in diagnostics, improve patient communication, and manage medical records more efficiently. In education, they can help create adaptive learning environments and provide personalised experiences to students (Murgia, Pera, et al. 2023). As evidenced by the literature (Haghani 2023), the growing trend towards adopting and integrating LLMs across diverse socio-technical systems suggests an inevitable trajectory towards their widespread utilisation.

This integration, however, is a double-edged sword. While it provides opportunities to refine and address existing issues within socio-technical systems, if not managed carefully, incorporating LLMs could exacerbate existing issues (Bender et al. 2021) or even introduce new, unforeseen challenges. The literature highlights several areas of concern, including but not limited to the potential for increasing various forms of biases in algorithms (Blodgett et al. 2020; Lopez 2021; Prabhakaran, Hutchinson, and Mitchell 2019), which may lead to unfair or discriminatory outcomes (Drage and Mackereth 2022; European Union Agency for Fundamental Rights 2022; Navigli, Conia, and Ross 2023), and the risk of data privacy breaches when handling sensitive information (Bender et al. 2021; Schappert 2023). Additionally, overreliance on automation without human oversight could undermine critical decision-making processes and reduce accountability (Cummings 2006; Goddard, Roudsari, and Wyatt 2012; Skitka, Mosier, and Burdick 2000; Sterz et al. 2024). Furthermore, the digital divide could widen

(Xiao et al. 2024) if access to LLM-powered systems is not made equitable, leaving specific populations disadvantaged in benefiting from these technological advancements. This underscores the importance of developing strategies for the effective and responsible deployment and integration of LLMs across diverse domains.

The transformative potential of LLMs, characterised by their unprecedented scale, human-like interactions, and adaptability, sets them apart from conventional AI systems. Effective integration of LLMs into socio-technical systems is complex and presents unique challenges. Unlike traditional AI applications, which are often constrained by task-specific datasets and contexts, LLMs possess a cross-system reach that necessitates a comprehensive examination of their broader societal impacts. There needs to be a greater understanding of the societal concerns, challenges, and risks involved across domains. The present study seeks to address this lack of understanding, which is not only of academic importance but essential to safeguard society against potential adverse consequences. To bridge this knowledge gap, this paper emphasises the need for research to go beyond specific use cases of the technology and delve into the broader domain of understanding how LLMs should ultimately function within socio-technical systems. Specifically, this paper:

- Offers a set of strategies to improve LLM-related research ideas for socio-technical systems, identifying four key strategies (Section 4.1)
- Identifies common challenges in integrating LLMs across diverse socio-technical systems (Section 4.2)
- Presents action plans and key research themes, setting a research agenda (Section 4.3).

The present research takes a proactive approach, which would help navigate unforeseen difficulties and pitfalls due to insufficient preparation and understanding of the technology integration. To guide this exploration and inquiry, this paper has adopted a human-centered AI framework (Torkamaan et al. 2024), which we describe in Section 1.2.

1.2. Foundations of the human-centered AI framework

The interdisciplinary nature of integrating LLMs into socio-technical systems necessitates a nuanced approach. The human-centered AI framework, first introduced in Torkamaan et al. (2024), provides a comprehensive lens through which the multifaceted interactions between technology, users, and societal structures can

be examined and understood. This framework (Torkamaan et al. 2024) provides a multi-dimensional approach to understanding, designing, and implementing AI systems and comprises four main components: (1) paradigms, (2) actors, (3) values, and (4) the level of realisation and evaluation. The framework illustrates AI research's dynamic and evolving nature and has a paradigmatic approach to AI systems. It introduces four paradigms, each presenting a complementary viewpoint through which AI systems can be designed, conceptualised, and refined: technology-centric, user-centric, human-centric, and future-centric perspectives (Figure 1). In a nutshell, this framework underscores the evolution of AI research from a primarily technology-focussed approach, emphasising algorithmic accuracy and performance, towards more inclusive and comprehensive paradigms that consider the broader societal impacts, ethical considerations, and long-term implications of AI technologies.

The transition from a technology-centric paradigm to a more user-centric perspective results from acknowledging the significance of user experience and interaction with AI systems. This transition marks a shift from solely algorithmic efficiency to a broader understanding of AI's role in human lives, emphasising the importance of usability, user preferences, and experience.

Subsequently, the human-centric paradigm broadens the lens beyond individual users to include various stakeholders, such as non-users affected by AI decisions, communities influenced by AI deployment, developers, policymakers, and others impacted by AI systems. This

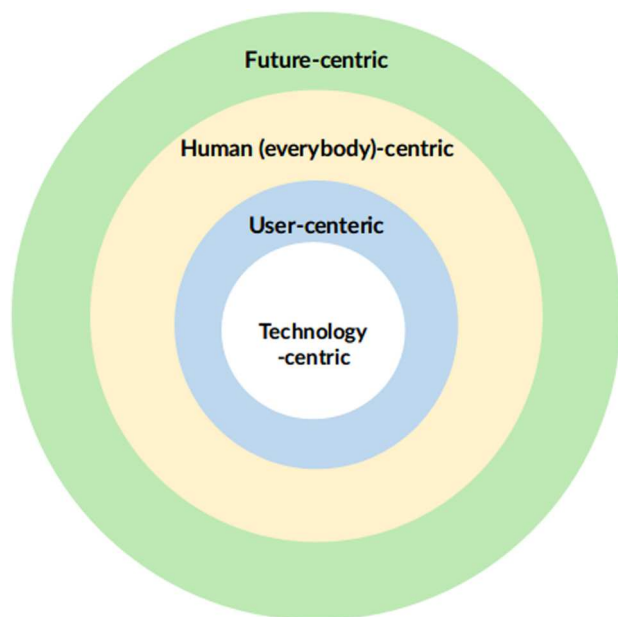


Figure 1. Paradigms of human-centered AI framework, adapted from Torkamaan et al. (2024).

approach delves into crucial issues and values like fairness, privacy, and ethical dimensions of AI. It acknowledges the varied impacts of AI systems, illustrating how they might affect different societal groups. For instance, it considers how data privacy concerns differ between end-users and individuals whose data are collected indirectly. Similarly, it addresses fairness, not just in terms of user experience but in how AI systems can perpetuate or mitigate societal inequalities (Costanza-Chock 2020; Eubanks 2018). This perspective emphasises the importance of equitable outcomes, ensuring that AI technologies are developed and implemented in a just and beneficial way for all members of society, not just a select few. The human-centric paradigm focuses on analysing current practices and their real-life implications. It concerns itself with existing and pressing societal issues and emphasises alignment with current societal norms, legal regulations, and ethical guidelines.

The *future-centric paradigm* represents the most advanced stage in this evolution, where the focus extends to designing AI systems that anticipate long-term societal consequences (Torkamaan et al. 2024). With a proactive and forward-looking approach, AI system owners and developers are encouraged to acknowledge their accountability for the long-term consequences of their creation. The future-centric paradigm fosters open discussion regarding the realisation and future of ethical principles and involves long-term studies, ensuring AI aligns with evolving societal values and needs.

Commonly used and relevant scope, methods, success criteria, and limitations are specified within each paradigm. The human-centered AI framework (Torkamaan et al. 2024) presents a transition from a narrow, technical focus to an inclusive and forward-looking approach that considers the complex interplay between technology, individuals, society, and ethical considerations. Accordingly, it provides a foundation for AI research and development that is both comprehensive and adaptable to the changing landscape of AI technology and its societal implications. This is the rationale behind its adoption in the study presented in this paper. By employing this framework as a guiding structure, the workshop (described in the next section) explored the issue of LLMs' integration into socio-technical systems, highlighting the need for AI systems that are ethically aligned, human-centric, socially impactful, and future-resilient. The framework thus served as an instrumental tool in exploring, shaping, and assessing a research agenda that aligns with the ethical, user engagement, societal influence, and anticipatory governance demands in the rapidly advancing domain of artificial intelligence.

2. Related work

The integration of LLMs into various research domains and socio-technical systems has gained significant attention since 2022. Recent studies (e.g. Gordon et al. 2024; Hacker, Engel, and Mauer 2023; Haghani 2023; Lo 2023; Murgia, Abbasiantaeb, et al. 2023; Novelli et al. 2024; Oviedo-Trespalacios et al. 2023; Thirunavukarasu et al. 2023) have extensively examined the applications and impacts of LLMs, including ChatGPT, across diverse fields, underscoring their interdisciplinary utility. This research spans disciplines such as Medicine and Computer Science, with growing interest in the Social Sciences, Arts, and Humanities. For a literature trend analysis of existing research on LLMs across these fields, see Appendix B, which offers a more comprehensive overview. Notably, these studies highlight the significant roles of LLMs in education and healthcare, with frequent intersections in areas like decision-making, ethics, and data privacy. The widespread and rapid adoption of LLM technology underscores its complex and interdisciplinary nature. These insights emphasise the need for a holistic approach to conducting technically sound, ethically responsible, and practically applicable socio-technical research. Additionally, they underscore the importance of developing research strategies for effectively integrating LLMs into socio-technical systems, guided by principles from human-centered AI.

The concept of human-centered AI plays a crucial role in designing and developing AI systems that align with human values and societal needs. Research and novel approaches in this field, such as the framework proposed by Sousa et al. (2024), place user trust at the core of socio-technical designs. This framework advocates for AI development incorporating social and community insights, human-computer trust interactions, and user trust characteristics. Similar work has highlighted that human-centered explanations can significantly increase user reliance on AI systems (Scharowski et al. 2023). Shneiderman's Human-Centered AI (HCAI) framework (Shneiderman 2022, 2020a, 2020b) is particularly relevant as it promotes balancing human control and AI automation, ensuring trustworthiness, safety, and reliability. This framework also incorporates governance structures with audit and ethical considerations. Additionally, frameworks such as the US National Institute of Standards and Technology (NIST) AI Risk Management Framework (AI 2023) and various ethics-based auditing models (Mökander and Floridi 2021; Raji et al. 2020) and standards (IEEE 2020; IEEE Standards Committee and Joanna Isabelle Olszewska 2020; ISO 2021) support the advancement of trustworthy AI.

Several initiatives within the European AI and robotics landscape aim to create trustworthy AI systems that align with European values, such as respect for human dignity, freedom, democracy, and the rule of law. The High-Level Expert Group on Artificial Intelligence⁹ has highlighted seven characteristics that AI systems should exhibit to be considered trustworthy: human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity and non-discrimination, societal and environmental well-being, and accountability. The European Union has established six AI Networks of Excellence to address these characteristics: AI4Media¹⁰, ELISE¹¹, ELSA, euROBIN, HUMANE-AI¹², and TAILOR¹³. For instance, AI4Media focuses on media-related AI technologies, ELISE emphasises machine learning innovations to create trustworthy AI systems, and HUMANE-AI and TAILOR focus on building the scientific foundations of trustworthy AI.

The Joint Strategic Research Agenda of the EU¹⁴ highlights legal questions surrounding training data and methods to keep LLMs factual and updated as key research directions. This paper complements these efforts to analyse and build trustworthy AI by focussing on the challenges of integrating LLMs across socio-technical systems and domains. Furthermore, the EU addresses AI through its Digital Rights Principles (European Commission 2023), emphasising the importance of freedom of choice. The EU also has provided the first regulatory framework for AI that specifically sets out rules for general purpose models like LLMs (recital 99, 105 Art 3 (66), Art 75 AI Act (EUR-Lex 2021)). In the context of elections, the EU addresses generative AI content within the Digital Services Act, which regulates content moderation and platform governance. This includes guidelines for online platforms that have reached a specific user threshold, emphasising the mitigation of risks associated with generative AI, misuse in political advertising, media literacy initiatives, and the labelling of AI-generated content (EUR-Lex 2024).

Other human-centered AI frameworks emphasise human engagement and participation in AI development, leveraging data donors as domain experts (Yoo et al. 2024) and addressing the under-represented communities (Freeman 2020; Mack et al. 2024) for creating supportive environments that place these groups' lived experiences at the forefront of design. For instance, studies have shown that existing text-to-image generative AI models often produce biased outputs that misrepresent people with disabilities (Mack et al. 2024). By focussing on specific user groups, developers can enhance the accuracy of LLMs by combining human and machine intelligence, resulting in better

technologies and outcomes for users (Monarch 2021). However, the consequences of AI (related) decisions extend beyond active users and can affect non-users and unintended audiences. The design justice approach, as articulated by Costanza-Chock (2020), advocates for inclusive design processes led by marginalised communities. This approach ensures that AI systems do not perpetuate societal biases but instead foster equitable outcomes for all stakeholders. Addressing issues such as algorithmic bias is essential to ensure that AI systems are carefully designed and evaluated to promote justice and accountability. Further work is needed to guarantee equitable outcomes for all AI stakeholders in using and deploying human-centric systems and properly integrating LLMs into socio-technical systems.

To effectively integrate our research with existing frameworks and initiatives in human-centered AI, we have employed a comprehensive, multi-dimensional framework that centres on human values and societal needs (Torkamaan et al. 2024) (Section 1.2). Building on the foundation established by existing research on human-centered AI and trustworthy AI initiatives and agendas, this study enhances these efforts by adopting a flexible and generally applicable framework (Section 1.2) to address the complex challenges of integrating LLMs into diverse socio-technical systems. This approach aligns this study with notable frameworks, such as those discussed above. It promotes interdisciplinary collaboration, proactive ethical considerations, and value-driven design, addressing critical issues such as bias, accountability, trustworthiness, and inclusivity. This ensures that our findings are robust and relevant, providing actionable insights for the responsible and effective deployment of LLMs in diverse socio-technical contexts.

3. Methodology

This study employed a structured, multi-part approach to explore the integration and impact of LLMs within socio-technical systems. The methodology comprised three distinct parts, the third part building upon the insights and outputs of the previous ones to develop a comprehensive understanding, reflective of the interdisciplinary nature of integrating LLMs into these systems.

Part 1 – Collecting Research Ideas. We launched a workshop call focussing on how LLMs are shaping socio-technical systems, addressing their integration and impacts. We aimed to bring together experts from various disciplines, including computer science, safety and security, psychology, ethics, design, and sociology, to share their knowledge, ensuring various experiences and perspectives in the discussion. Interested

participants were asked to prepare an abstract of 200–500 words, detailing either a hypothetical future research project or their viewpoints, interests, ideas, or concerns regarding LLMs. A total of 13 abstracts were submitted. These abstracts served as the starting point for analysis, further dialogue, and ideation during the workshop (see Appendix A for the abstracts). A thematic analysis approach was used to categorise and interpret the abstracts, identifying key patterns, themes, and research trajectories proposed by the participants.

Part 2 – Literature Trend Analysis. We conducted a trend analysis of ChatGPT and LLMs research to have an updated and broader scope of inquiry within the domain of LLMs using the Scopus database (Scopus n.d.). We extracted search results for the term ‘ChatGPT’ and the more general ‘large language model*’ as textual BibTeX entries. This query was conducted in August 2023 before the workshop and was repeated and updated in November 2023 for this paper. No significant difference in trends between the two analyses was observed. In November 2023, the query for ‘ChatGPT’ resulted in 2,841 titles, and the query for ‘large language model*’ returned 2,543 items. Upon gathering a substantial number of abstracts, we relied on the insights generated via Scopus and conducted additional automated literature trend analysis using basic topic extraction algorithms¹⁵. This computational analysis aimed to identify prevailing themes, gaps, and emergent patterns within the existing body of work. Details on the results of this trend analysis are listed in Appendix B.

Part 3 – The Workshop. Following the identification of key trends and practices within the literature, the research progressed to Part 3—the workshop component (Figure 2). The choice of a workshop over other methods, such as interviews or purely literature reviews, was intentional to foster real-time, dynamic exchanges and collective problem-solving, and to facilitate the integration of diverse viewpoints and the synthesis of collective intelligence, both essential for addressing complex, multifaceted issues. Twelve scholars confirmed their participation in the workshop following the invitation. A total of ten individuals attended the in-person session. Another session was scheduled for the other two individuals, while logistical challenges led to separate discussions with them. These discussions ensured their contributions were included in the workshop tasks and outcomes. Additionally, to maintain objectivity, one experienced scientist with expertise in participatory systems, who did not submit an abstract, acted as an unbiased observer and provided an external perspective, helping reduce potential bias and structure the workshop results. All 13 contributors (listed in Table 1)¹⁶

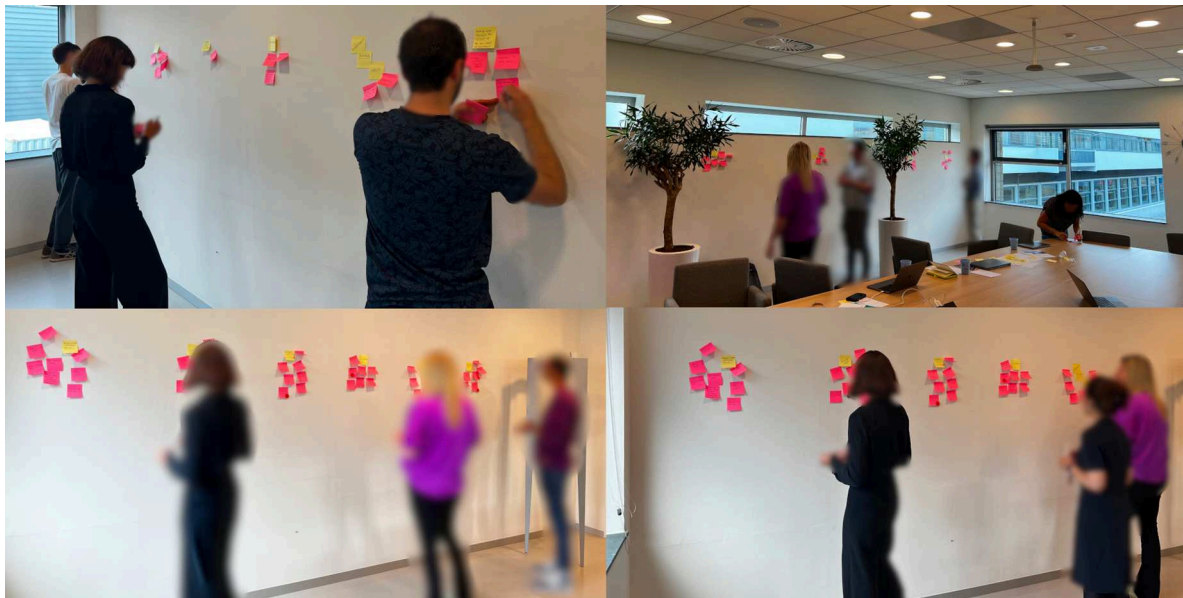


Figure 2. The workshop as an interactive colloquium bringing distinct expertise together.

Table 1. Contributors' backgrounds and areas of expertise.

| Identified gender (Years of research experience) | Role | PhD degree in | Expertise |
|--|---|-----------------------------------|--|
| Female (8 years) | Assistant professor in AI for health systems | Computer science | Digital Health, Recommender Systems, Human-Centered AI, User Modeling and personalisation, Ubiquitous computing |
| Male (10 years) | Assistant professor in Ethics & Philosophy | Philosophy, History and Sociology | Philosophy of technology, ethics of technology, emotions, values |
| Female (15 years) | Associate professor in Web Information Systems | Computer science | Information Retrieval (IR), Children Information Retrieval, Information Access, Recommender Systems, IR for Non-traditional Users, Education |
| Male (4 years) | PhD candidate in Knowledge and Intelligence Design | Industrial design | Human-centered AI, Knowledge Sharing, AI and Design, Conversational agents |
| Female (2 years) | PhD candidate in Platform Governance and Digital Regulation | Technology Policy and Management | Content moderation, platform governance and regulation, AI and socio-technical and legal system design |
| Female (5 years) | PhD candidate in AI & Society | Human factors engineering | Psychology, Technology Acceptance, AI & Society, Data Stewardship |
| Male (14 years) | Assistant professor in Knowledge and Intelligence Design | Computer science | HCI, Social Cognition, Human-Centered AI, Empathy-Centric Design, CSCW |
| Male (11 years) | Assistant professor in Web information systems | Computer science | Human-Centered AI, Crowd Computing, Human Language Technologies, Recommender Systems |
| Female (14 years) | Assistant professor in Safety and Security Science | Safety science | Safety Culture, Human Behaviour, Safety Education and Training |
| Male (22 years) | Full Professor in Complex Systems Design | Computer Science | Complex Systems, Distributed Systems, Artificial Intelligence, Self-Management, Security & Privacy, Energy |
| Male (13 years) | Assistant professor in Responsible Risk Management | Human factors engineering | Safety Science, Misuse of Technology, Socio-technical systems, Engineering Psychology, Transport, Industrial Management |
| Female (9 years) | Assistant professor in Ethics/Philosophy of Technology | Ethics of technology | Ethics of technology, Postphenomenology, AI, Mental healthcare, Voice assistants |
| Female (37 years) | Full Professor in Systems Engineering | Cognitive Psychology | Participatory, Systems, Distributed Systems, Multi-agent Systems, Design, Autonomic Computing |

had varied areas of expertise and experience in their respective areas ranging from two to 37 years and are confirmed as co-authors of the final paper, having actively engaged in the process, had access to the data, and contributed to the collaborative writing of the paper. Furthermore, throughout the entire process, we maintained structured critical reflection, detailed

documentation, and open discussion to ensure objectivity and minimise bias in our findings and reporting.

The interdisciplinary nature of LLM applications and their cross-disciplinary impact on various socio-technical systems necessitates a broad spectrum of expertise to address the associated challenges effectively. Expertise in AI, human-computer interaction, ethics, sociology,

and system design is crucial to navigating the complexities of integrating LLMs into socio-technical systems. By fostering a multidisciplinary environment, the workshop aimed to leverage the diverse expertise of participants to develop comprehensive strategies for the responsible and effective integration of LLMs.

The workshop was structured as a sequence of six distinct phases depicted in Figure 3. The welcome and group introduction were conducted in Phase 0, 'Introduction'. Then, Phase 1, 'Exploration,' proceeded through presentations and discussions based on preliminary literature trend analysis insights. Additionally, participants individually reflected on how their research is connected to such literature trends, relying on the submitted abstracts and group exploration to identify key issues.

In Phase 2, 'Elevation,' participants worked towards elevating their research ideas to a more future-centered perspective guided by the Human-Centered AI Framework (See Section 1.2), which was presented to the participants. The framework (Torkamaan et al. 2024) describes different paradigms in detail and provides an

overview of each paradigm's common methods, success criteria, scope, and limitations. Each participant read their abstracts and identified the scope of their research within the paradigms of the framework, specifying the focus, methods, and scope of their research. Subsequently, participants redefined their research ideas through group discussions, modifying them to align with a higher-level perspective. This process involved critically examining their initial concepts and adjusting their approaches to fit within the broader, future-centric paradigms outlined in the framework.

Phase 3, 'Expansion,' focussed on opportunities for expansion across other socio-technical systems. During this phase, participants reflected on the scope of their research and discovered opportunities to align their work with broader socio-technical perspectives. They identified common challenges and concerns for designing and integrating socio-technical systems across expert domains and other socio-technical contexts. Phase 4, 'Ideation and Prioritization,' was brainstorming and ideating solutions to solve the identified integration and design challenges. Phase 4 transitioned

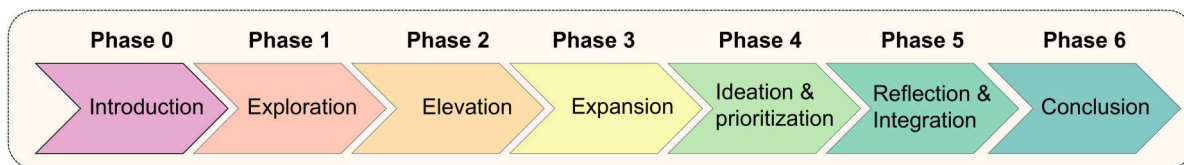


Figure 3. Sequential phases of the conducted workshop with experts.

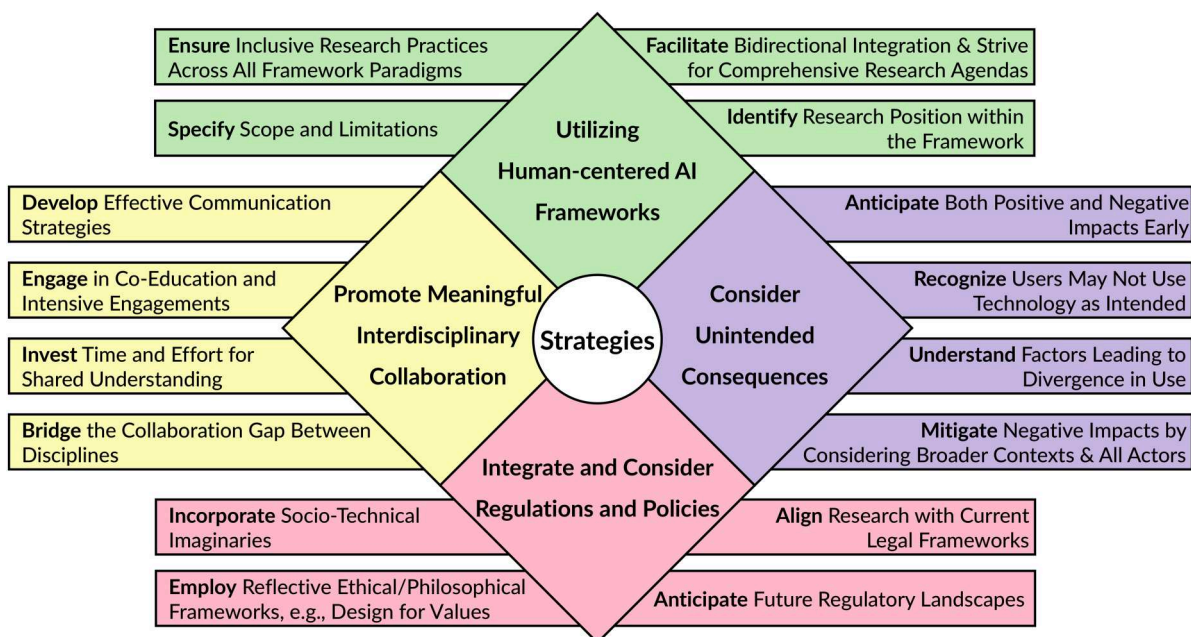


Figure 4. An overview of strategies to improve LLM-related research ideas for socio-technical systems.

into Phase 5, ‘Reflection and Integration,’ a reflective session where participants revisited their initial ideas and discussed the practical implications of LLMs in addressing societal challenges. This iterative and participatory approach ensured a comprehensive field exploration, blending individual expertise with collective intelligence to identify and prioritise research directions. Phase 6 concluded the workshop with further group reflection and discussion. The workshop lasted for four hours.

Part 4 – Analysis and Collaborative Write-up. Following the workshop, a thematic analysis was conducted to organise and extract the insights gained from all previous sections of the study. This analysis helped to identify emerging themes and refine the rich input from the workshop into actionable research directions. Each contributor reviewed the analysis and interpretation of the results independently. Subsequently, the lead and last authors structured the first draft of the manuscript. All workshop participants, who are also co-authors of this paper, then engaged in collaborative writing with researchers jointly contributing to this paper by adding and editing content¹⁷. This collaborative effort ensured that our final outputs were not only multidisciplinary but also benefited from workshop participants’ collective expertise and critical insights.

4. Results and discussion

4.1. Strategies to improve LLM-related research ideas for socio-technical systems

To identify strategies to improve LLM-related research ideas for socio-technical systems, we started with an exploration of the abstracts¹⁸ of hypothetical LLMs research ideas submitted by the workshop participants through a group reflection (Phase 1 of the workshop). The abstracts covered a wide range of application domains, namely, education ($n = 2$), health ($n = 4$), governance ($n = 2$), and risk and safety management ($n = 5$), with some overlaps, each offering unique insights into the integration and implications of LLMs. As can be seen by the diversity of topics covered, the presence of LLMs in academic discourse is evident, indicating a collective recognition of their potential to reshape various disciplines. This diversity spans areas, such as education integration, health collaboration, ethical considerations, social implications, and technological challenges. It reflects a growing awareness of the multifaceted nature of LLMs and their profound influence on societal dynamics. Additionally, it was noted that the collection of abstracts used in this workshop maintains similarity to the current trend in the LLMs literature, as discussed

during the Exploration Phase and in the literature trend analysis (Appendix B).

As part of the exploration phase, participants assessed their research ideas in light of the human-centered AI framework (Torkamaan et al. 2024) introduced earlier, aiming to bridge technical aspects with social considerations. This framework, delineated into four paradigms, transitions from technology-centric to socio-centric concepts. Participants utilised these paradigms to analyse their submitted abstracts in the workshop’s initial phase, identifying the framework’s specific paradigms relevant to their research concepts. Participants were able to assign their research to one of the paradigms. The majority of the attendees placed their submitted abstract in either technology-centric or user-centric paradigms, as specified in Figure 1. For example, Abstract 2 (A2) focussed on designing for human-AI collaboration in healthcare. Particularly, it described the abstract’s author’s engagement with physicians in an attempt to understand local practices at the cancer treatment facility of a major European hospital and propose pathways for meaningful incorporation of AI that are better suited to support physicians in their clinical and research practices. The abstract’s author positioned it in the user-centric paradigm since it goes far beyond the scope and methods of the technology-centric paradigm. Yet, it does not consider other stakeholders involved or affected. The success criteria are also limited to efficiency metrics in daily practice and do not consider collective or long-term outcomes or impacts.

Upon determining the positioning of their research within the human-centered AI paradigms, the Elevation Phase of the workshop followed. In this phase, participants engaged in reflective discussions about the breadth, depth, and limitations of their research, exploring avenues for refining and advancing their research ideas. Participants then identified strategies to enhance the human-centric or future-centric nature of their research using the framework (Torkamaan et al. 2024). By finding the gaps and overlooked areas, the framework guided participants in determining the key actors, values, and the extent to which these values should be realised, all of which facilitate the development of a future-centric socio-technical research agenda. Overall, the framework helped participants recognise areas where their research was lacking in terms of inclusivity and long-term impact. This recognition prompted discussions on how to address these gaps effectively.

Several strategies (Figure 4) were discussed to make research more inclusive of other paradigms, including *integrating considerations of regulations and policies, interdisciplinary collaboration, consideration for unintended consequences, and utilising proper human-*

centered AI frameworks for reflection from the start. A strategy discussed involves integrating considerations of regulations and policies into research. Such an approach ensures that research not only aligns with current legal frameworks but also anticipates future regulatory landscapes. Reflective frameworks, especially those rooted in philosophical and ethical considerations, like Value-sensitive Design (Friedman and Hendry 2019) and Design for Values (van den Hoven, Vermaas, and van de Poel 2015), guided by the human-centered AI framework (Torkamaan et al. 2024) utilised in this paper or other suitable frameworks, e.g. Shneiderman (2022, 2020a), were identified as crucial strategies in this process. They provide a means to address and mitigate the potential negative impacts of AI technologies. For example, employing philosophical or ethical frameworks could not only guide the rectification of adverse AI aspects but also help to anticipate them by ensuring the inclusion of crucial stakeholder perspectives and values. In addition, socio-technical imaginaries (Mlynar et al. 2022; Verma et al. 2023) emerged as a key component in strategies for designing and integrating AI within socio-technical systems. These shared visions and perspectives significantly shape the deployment and acceptance of AI in various social settings. Recognizing and incorporating these socio-technical imaginaries is essential for creating inclusive AI solutions that resonate with the diverse needs and expectations of different stakeholder groups. This approach ensures that AI technologies are not only technologically sound but also socially relevant and accepted.

Interdisciplinary collaboration stands out as a cornerstone in advancing this human-centric research agenda. There is a notable collaboration gap that, if bridged, would result in more comprehensive and impactful socio-technical research outcomes. Currently, a divide persists between technical experts and professionals from fields such as social science, humanities, and policy, among others. This division often results in teams focussing exclusively on either technical or non-technical aspects, thereby overlooking the opportunity to understand the complex interplay between these two elements. Meaningful interdisciplinary collaboration (Mariano 1989) requires co-education and intensive engagements to define, execute, and assess the research and its future. By fostering an environment where both sides can learn from each other and work together closely, we can create AI solutions that are not only innovative but also resonate deeply with the diverse needs and contexts of society. This comprehensive approach is the key to unlocking the full potential of AI in serving humanity.

From the workshop discussions, it became clear that such interdisciplinary collaboration often faces

challenges in terms of communication and shared understanding. For example, the general assumptions about the technology, fuelled by misunderstanding, bias, and ignorance, often do not align with its actual abilities and limitations. Sometimes, it can take several months for members of an interdisciplinary team to align their language and still more time to reach a common understanding of the problems they are addressing. This process requires more than just dialogue; it necessitates a deep investment in time and effort.

Another strategy that can help in having more future-(or human-)centric research is to consider both positive and negative impacts, particularly unintended consequences, beyond the scope of the specific solution developed from the start. Over time, the application and utilisation of technology have consistently revealed a fascinating trend: users often do not use technology in the ways its creators initially intended, e.g. Dourish (2003). This divergence between intended use and actual use can be attributed to several factors, including the evolving needs of users, the creative repurposing of technology, and the unforeseen contexts in which technology finds application. The negative aspects of the unforeseen impacts and unintended uses of the technology can be mitigated by considering that the specific solution that is being developed may be used by everyone and in a broader context. The human-centered AI framework (Torkamaan et al. 2024) encourages consideration of all stakeholders (including non-users), which aids in resolving this issue.

Systematic considerations and utilising proper human-centered AI frameworks for reflection and position of the research from the start give the scientists the tools to identify the position of their research and strive for more comprehensive research agendas by using proper methodologies and specifying the scope and limitations of their studies. This strategy works both for solutions that are focussed only on the inner layers of the human-centered AI framework (Figure 1), e.g. technology-centric, and those solutions that exclude inner layers. Some workshop attendees noticed that their abstracts are focussed merely on the outer layer and are exclusive of the inner layers, i.e. user- or technology-centric. For instance, the A9 abstract was about the potential negative consequences the use of LLMs could have on the upcoming European Elections 2024 in light of the European Regulation (EUR-Lex 2022) for online platforms and content moderation, not taking the crucial inner layers that encompass user and technology aspects into account. This observation underscores the importance of inclusive research practices encompassing all framework paradigms. Researchers can create more balanced and impactful solutions by

ensuring that each layer of research incorporates the essential elements of the inner layers (Figure 1). The transition between the human-centered AI framework functions both ways. While technology-centric scientists think of how to bring the solution outwards, society-centric scientists think of how to bring the solution inwards from future-centric toward more technology-centric considerations. It then opens the conversation of not setting one regulation but how regulating AI use, and particularly LLMs, would impact or should impact other regulations.

4.2. Common challenges in integrating LLMs across diverse socio-technical systems

A key question in socio-technical systems is whether different systems face similar challenges despite their varied disciplines and contexts. This includes exploring if problems with LLMs in one system mirror those in others and understanding their mutual impacts. Recognizing these common challenges in incorporating LLMs across various socio-technical settings can guide the development of future research agendas.

In this workshop, we explored the common challenges in integrating LLMs across diverse socio-technical systems by prompting participants to reflect on expanding their research ideas—which had been elevated to align with higher-level socio-technical paradigms—across other socio-technical systems. It was commonly understood that the common challenges and concerns are rooted in the complexities and ethical considerations involved in designing and integrating LLMs into socio-technical systems. The following challenges and issues were identified cross-systems during the brainstorming phase (Phase 4 of the workshop):

Changes in the context of use. Traditionally, systems are designed to satisfy specific user needs. In user-centered design, it is important to clearly understand this context to design technology that effectively meets such needs. However, dynamics in the appropriation and adoption of technology can lead to alterations in the context of its use. Consequently, a particular application of technology might shift from one context to another. Such transitions can lead to serious disruptions, as the technology may not have been designed to accommodate these contextual changes, thereby increasing vulnerabilities. A significant limitation of LLMs is their lack of contextual awareness, which poses a safety risk, e.g. see Oviedo-Trespalcacios et al. (2023). This oversight in design means that, while powerful, LLMs may not effectively adapt to changing user environments or requirements, leading to potential misalignments between the technology's function and the evolving needs of its users.

Responsibility, Liability, and Accountability. These values, which correspond with Assurance and Accountability values in the human-centered AI framework (Torkamaan et al. 2024), are multifaceted. For instance, responsibility can mean different things to different people. Furthermore, philosophers have identified different kinds of responsibility concerning technology (Doorn 2012). There is also an ongoing debate over who is responsible for the outcome related to LLMs, which currently involves creators, users, data generators, and annotators, among other actors and stakeholders. The challenge of attribution of responsibility is sometimes described as the problem of many hands because so many actors are involved, including artificial ones. This situation may create a so-called responsibility gap, where it is not clear who is responsible. To ensure accountability and responsibility, it is crucial to address these issues. Another important question is, 'Who is liable when something goes wrong?' The contextual dynamics of different socio-technical systems likely yield varied responses to this question. For example, the discussion of the liability of entities like banks and AI companies in financial systems might result in different results than discussions about the liability of doctors in healthcare systems or content providers in digital entertainment systems. In addition, 'Who can one sue?' and 'Where can I go when something goes wrong?' and 'To what extent can an AI agent be considered autonomous?' are collective challenges, and responses may vary for different systems and social domains.

Trustworthiness and User Perception. The concept of trustworthiness in relation to LLMs encompasses two distinct aspects. Firstly, it involves assessing the inherent trustworthiness of the system, which includes evaluating the reliability of content and responses generated by LLMs within a socio-technical system. Secondly, it concerns the user or human perception of trustworthiness, which acts as a primary safeguard against potential harm arising from interaction with the system or external entity. Gaining user trust is essential for successfully integrating, adapting, and accepting new AI solutions (Kelly, Kaye, and Oviedo-Trespalcacios 2023; Wu et al. 2011). Ideally, a system's actual trustworthiness would align with how users perceive it. However, in reality, there often exists a discrepancy between these two aspects. Another challenge with LLMs is their capacity to produce human-like responses. This capability can obscure indicators of unreliable information, potentially leading to a misplaced sense of trustworthiness. Such misalignment can foster confidence tricks in interactions, resulting in a false sense of security and potentially harmful consequences.

Impact on Critical Thinking and Bias. The increasing reliance on LLMs raises concerns about the deterioration of human creativity and critical thinking abilities (Jakesch et al. 2023). This issue is part of a broader discourse on how technology dependence affects human skills, health, and behaviour. For instance, habitual GPS use impacts navigational capabilities and can negatively affect a person's spatial memory abilities (Dahmani and Bohbot 2020). The critical issue with LLMs is their widespread application in tasks involving human writing and critical thinking, which are pivotal for the future of societies. Such overreliance may lead to greater acceptance of biased or incorrect information, heightening societal vulnerability to manipulation. This phenomenon underscores the need to cultivate and maintain intellectual virtues, like critical thinking skills, in the digital age, ensuring that reflection and unbiased information processing remain integral components of societal decision-making.

Inclusivity in Data and Stakeholders. The development of LLMs currently faces significant challenges related to inclusivity, both in the data used and in stakeholder representation. Historically, the lack of diverse representation has led to outcome biases (Noble 2018). Therefore, it is crucial to critically reflect on the data used (versus data that should be used) to train/power/develop these models in the context of potential pitfalls. Currently, well-known models powering popular chatbots and other technologies, both academic and commercial, rely heavily on specific data sources. For instance, GPT-based models leverage Wikipedia and CommonCrawl data sources (Minaee et al. 2024), which fail to capture the full spectrum of human experience and perspectives adequately. Despite this limitation, its reach is propagated as GPT models have been fine-tuned to yield OpenAI's ChatGPT, raising questions about its applicability to serve a wide range of users on inquiries on diverse topics. To foster the long-term design of LLMs that can effectively represent and cater to a broader audience, we argue that it is important to incorporate representative data and engage a diverse range of stakeholders right from the initial stages of LLM design. This approach will ensure that LLMs are developed and integrated into socio-technical systems in a truly inclusive manner that reflects the diverse dimensions of human society.

Unintended Consequences and Future Harm Prevention. The deployment of LLMs can lead to unintended consequences, necessitating a proactive approach to harm prevention. This encompasses not only the risks associated with intended malicious use and misuse, like manipulation and deception, but also those arising from unintended yet creative applications beyond the tool's tested functionalities and scope. An example

might be the creation of mis- or disinformation in the context of elections and the dissemination of such content on online platforms targeting political candidates. Developing robust strategies to prevent harm is essential in the evolving landscape of LLM applications.

False Information and Hallucinations. From a technological standpoint, LLMs may unintentionally generate false and potentially unethical content, often called 'hallucinations'. Addressing this issue involves not only technical solutions to improve the accuracy and ethical compliance of LLMs but also a comprehensive study of the broader implications these hallucinations have on socio-technical systems. This dual approach presents a significant challenge, as it requires an intricate understanding of both the technical mechanisms of LLMs and the complex interactions within socio-technical ecosystems.

Intellectual Property. One of the challenges of LLMs and generally generative AI is intellectual property (IP) issues, particularly in relation to attribution and the use of various content types in training LLMs. The challenges related to IP include understanding and navigating the complexities of crediting sources for datasets, algorithms, human feedback, and multimedia materials utilised in model development while adhering to legal and ethical guidelines. However, current legal frameworks do not cover the complexities of IP issues related to LLMs (Hacker, Engel, and Mauer 2023). Several ongoing lawsuits against developers of LLMs, such as The New York Times case against OpenAI¹⁹ and Github's users case against Microsoft, OpenAI, and Copilot, are likely to be instrumental in setting precedents²⁰. This issue is intertwined with other challenges discussed, such as harm prevention, false information, inclusivity in data, and particularly responsibility, liability, and accountability.

Privacy and Human Rights Considerations. An ongoing challenge for the integration of LLMs and the design of future socio-technical systems is balancing the trade-off between ensuring data privacy (Winograd 2022), obtaining informed consent (e.g. when an individual might be unable to provide it or considering the implications of IP and personal autonomy), and safeguarding other fundamental rights, such as the right to equal treatment, non-discrimination, the right to a fair trial, free speech or freedom to seek information and maintaining data reliability (EUR-Lex 2012). This trade-off extends to broader considerations, particularly in relation to human rights, where the ethical deployment and application of LLMs necessitate careful navigation to protect individual privacy and data integrity without compromising the effectiveness and accuracy of these models or preventing effective logging, tracing, and mandatory auditing in the EU AI Acts fundamental rights impact assessment for high-risk AI systems.

Individual Differences and Personalization. The challenge in designing inclusive socio-technical systems lies not only in accommodating diverse stakeholder data but also in reconciling the varying, sometimes conflicting, needs and rights of these actors. This complexity is amplified at the individual level, particularly in sensitive domains like healthcare and education, where a one-size-fits-all approach falls short (Kelly, Kaye, and Oviedo-Trespalacios 2022). Effective system personalisation, therefore, becomes both a practical necessity and an ethical imperative. It ensures that AI systems, especially those using LLMs, are not just effective but also equitable, genuinely meeting the diverse needs of all users.

The workshop discussions resulted in ten identified common challenges inherent to the integration of LLMs across diverse socio-technical systems. While these challenges share similarities with those encountered in responsible technological innovation broadly, the distinctive capacities and extensive reach of LLMs amplify and extend these concerns significantly. Unlike smartphones, blockchain, or other technologies (Appendix B), LLMs have the potential to fundamentally alter information processing, decision-making processes, and even the nature of human-machine interaction across various socio-technical landscapes.

Integrating LLMs into socio-technical systems presents unique challenges that are distinct from those encountered in broader AI research. Unlike traditional AI systems, which are typically designed for specific applications and trained on task-specific datasets, LLMs are generalists trained on vast and varied datasets. This generality complicates the development of context-aware responses, and adds complexities of intellectual property (Novelli et al. 2024), and accountability (Cheong et al. 2024) for errors, particularly in high-stakes scenarios, e.g. Cheong et al. (2024) and Ullah et al. (2024). Moreover, the human-like responses of LLMs can weaken innate human safeguards against machine-generated content (Oviedo-Trespalacios et al. 2023). Consequently, they can mislead users into overestimating their reliability, exacerbating trustworthiness concerns (Choudhury and Shamszare 2023). Additionally, LLMs' versatility makes them susceptible to unintended uses, including malicious applications like the spread of misinformation and manipulation (Ai et al. 2024; De Angelis et al. 2023; Yizhou Zhang et al. 2024).

Traditional AI systems face bias issues primarily related to the data they are trained on; the biases are often bound to specific applications that they are developed for, allowing for more targeted mitigation strategies. LLMs also, due to their training on broad, massive, and diverse datasets, are prone to inheriting and amplifying a wide range of biases (Bender et al.

2021; Gallegos et al. 2024). The sheer volume, variety, scale, and generality of processed data and the process of building LLMs make it harder to identify and mitigate these biases comprehensively (Kruspe 2024; Zack et al. 2023) and increase the risk of unintentional memorisation (Chiyuan Zhang et al. 2023) or exposure to sensitive information. The pervasive nature of LLMs means that their integration can have far-reaching implications, from amplifying existing societal inequities to introducing complex ethical challenges concerning autonomy, agency, privacy (Winograd 2022), and many more. Moreover, the rapid evolution and deployment of LLMs necessitate an agile and anticipatory approach to research, governance, and policy-making, ensuring that socio-technical systems remain adaptable and resilient in these transformative technologies. It is essential for researchers, practitioners, and policy-makers to collaboratively work towards solutions that are ethically grounded, socially beneficial, and technologically feasible, taking into account the multifaceted nature of socio-technical systems and the diverse needs and values of stakeholders involved.

4.3. A research agenda to integrate LLMs into socio-technical systems

This section outlines action plans and key research themes, forming the research agenda for integrating LLMs into socio-technical systems. This agenda is guided by insights from the workshop, in which we examined specific solutions for individual research cases and ideation using the human-centered AI framework (Torkamaan et al. 2024) (Section 4.1) and then broadened our focus to identifying common challenges across diverse socio-technical systems (Section 4.2).

4.3.1. Plans of action

Several ideas emerged (listed in Appendix C) during the workshop brainstorming session to address the identified challenges described in the previous section. These ideas were discussed as plans of action, listed in Table 2. While current studies often focus solely on technological or user-centric paradigms, our proposed agenda emphasises the need to shift towards more future-centric and inclusive approaches. This shift involves embracing methodologies prioritising human-centric and future-oriented research, as outlined in the human-centered AI framework (Torkamaan et al. 2024).

4.3.2. Research directions

In the workshop, we explored how to integrate LLMs into various socio-technical systems, identifying key

Table 2. Research directions for effective integration of LLMs into socio-technical systems.

| Challenges | Plans of Action |
|---|---|
| Context Recognition and Adaptation in LLMs | <ul style="list-style-type: none"> • Developing methods for LLMs to recognise and adapt to the context of use • Identifying limitations, refining LLMs, and particularly incorporating organisational and cultural nuances into LLMs responses |
| Legal and Ethical Frameworks for LLMs | <ul style="list-style-type: none"> • Exploring dynamic adaptation techniques for LLMs in varying socio-technical environments • Investigating legal and ethical frameworks for assigning responsibility in LLM-related incidents • Studying liability in different sectors (e.g. finance, healthcare, entertainment, etc.) • Developing guidelines for dealing with liability in different sectors and varying contexts • Standardizing and defining different stages and life cycles of LLMs, from creation to deployment, for accountability |
| Trustworthiness and Reliability of LLMs | <ul style="list-style-type: none"> • Enhancing trustworthiness and reliability and studying user perception of LLMs • Studying factors that influence user trust in LLMs • Developing proper metrics to assess and align the actual and perceived trustworthiness of LLMs • Researching the impact of human-like responses on user trust and information reliability, and potential associated risks |
| Impact of LLMs on Cognitive Skills and Bias Mitigation | <ul style="list-style-type: none"> • Increasing research investments and focussing on analysing the impact of LLMs on human cognitive skills and critical thinking abilities • Developing practical and effective methodologies for bias detection and mitigation in LLMs outputs • Dedicating fundamental investments to exploring educational strategies that can help maintain critical thinking skills in both children and adults in the presence of LLMs in the future |
| Inclusive Data Collection in LLM Training | <ul style="list-style-type: none"> • Making a conscious and deliberate effort to encompass a diverse range of perspectives and voices for inclusive data collection and representation in LLM training • Developing frameworks for diverse stakeholder participation in LLM design and policy-making, as well as for assessing the impacts of different data sources on LLM behaviour and outputs |
| Anticipating and Mitigating Unintended Uses of LLMs | <ul style="list-style-type: none"> • Developing strategies for anticipating and mitigating unintended uses of LLMs and LLMs-empowered socio-technical systems • Researching robust harm prevention mechanisms in various application domains • Exploring the ethical implications of creative applications of LLMs as critical steps toward responsible and safe design, development, and deployment |
| Reducing Misinformation and Managing Hallucinations | <ul style="list-style-type: none"> • Investigating solutions for reducing the generation of false or misleading content by LLMs • Studying the societal implications of LLM-generated misinformation • Developing guidelines for managing, mitigating, and correcting LLM hallucinations |
| Intellectual Property and Content Use in LLMs | <ul style="list-style-type: none"> • Researching legal and ethical issues related to the use of content and intellectual property in LLM training • Developing frameworks for attribution and source crediting in LLMs • Exploring the intersection of intellectual property with other LLM challenges |
| Privacy Protection and Human Rights in LLM Applications | <ul style="list-style-type: none"> • Investigating methods to protect privacy and uphold human rights in LLM applications with a dedicated research initiative that follows previous privacy-aware technology development, and considers particular challenges inherent in LLMs • Finding ways around anonymization and revealing an individual identity using LLMs • Exploring trade-offs between data privacy and model effectiveness • Developing technologies for secure data handling and model auditing to maintain the delicate balance between technological advancement and human rights |
| Adaptive and Personalized Approaches in LLMs | <ul style="list-style-type: none"> • Researching adaptive and personalised approaches and accommodating individual differences for LLMs in diverse socio-technical domains • Studying the ethical implications of personalisation in sensitive areas like healthcare and education • Developing inclusive design practices that cater to a wide range of user needs and rights, which can then translate into technical solutions with future-centric paradigm considerations |

research directions. Our discussions emphasised the importance of these ideas across different fields, their broad appeal, and the risks of ignoring these strategies, which could leave critical challenges unresolved. Based on a detailed analysis of ideas from Section 4.3.1 and Appendix C, we identified four main research themes that are crucial for effectively tackling these challenges. Focusing on these themes and building on the action plans outlined previously, we have developed research directions. These research directions aim to direct future research toward a deeper understanding and innovative approaches for smoothly incorporating LLMs into socio-technical systems.

Key Research Theme 1: Human-in-the-loop. The core of our research agenda considers the principle of human-in-the-loop to guide the development,

integration, and usage of LLMs. Human-in-the-loop emphasises the active involvement of humans in the design, development (including data creation, interpretation of model behaviour), integration, and application (e.g. presence of human operator and oversight in AI decisions) of the solutions from the start and continuing their involvement by continuously capturing their feedback to improve the solutions in an iterative process. By adopting co-design and participatory methodologies, this key research theme advocates for a collaborative framework where feedback from a diverse array of stakeholders—including end-users, experts, decision-makers, and those indirectly affected—is systematically incorporated, and their imaginaries from different contexts are adequately collected and considered. The group discussion revealed that a human-centered design

and human-in-the-loop can have different meanings for different disciplines. This key research theme necessitates a nuanced understanding of the term ‘human’ within various disciplinary and contextual boundaries, prompting a critical examination of stakeholder roles, actors involved, specifying their qualities or the expertise they need to possess, and the outcomes they influence, in integration of LLMs into socio-technical systems.

This approach is pivotal not only for addressing the challenges of *taking into account the context of use* but also for addressing broader challenges, such as *responsibility, liability, and accountability, inclusivity in data and stakeholders*, and long-term performance and data quality by continuously improving LLMs based on human feedback, expert in the loop feedback, development of human-centric and context-sensitive KPIs and stakeholder involvement. Integrating human values and knowledge was also mentioned during the workshop as a possible solution to the challenge of *false information and hallucinations*. However, human-in-the-loop approaches may require more effort and resources for technology development, impacting the speed of automation, despite having the potential to improve its quality. Human-in-the-loop can serve both as a method and as a foundation for deep research within the context of integrating LLMs into socio-technical systems. For example, as a method, it emphasises the active involvement of humans, ensuring alignment with values, needs, and ethical needs. Beyond its application as a method, it presents a rich venue for deep research, exploring various dimensions of interaction, system design, ethics, and broader societal implications.

Key Research Theme 2: Multi-stakeholder longitudinal studies. This theme underscores the necessity for comprehensive, long-term studies to assess the real-world impacts of LLMs within socio-technical systems. This theme was first mentioned in the workshop to address the challenge of *inclusivity in data and stakeholders*. It was also discussed as a solution to *trustworthiness and user perception, unintended consequences and future harm prevention* challenges. The core idea here is that offline evaluation, a one-time field evaluation, and traditional short-term studies are insufficient to uncover the problems that may lie within the developed solution or latent issues. Long-term and longitudinal studies should be increased and given higher priority to investigate and determine the real-world impacts or the extent of the impact resulting from integrating LLMs in socio-technical systems. Technology’s impact on human well-being and behaviour, as well as social impacts and the effect of decisions made using such technology, require a long-term perspective and a more future-centric view.

This research theme also encompasses the continuous improvement of LLMs through ongoing human feedback. Long-term observation and detailed study of LLM usage are integral to addressing challenges related to long-term performance and data quality. This involves a sustained and systematic approach to collecting human feedback, coupled with regular benchmarking and performance analysis of LLMs. Evaluating these systems based on extensive, real-world evidence ensures that they not only meet current needs but also adapt and evolve in response to emerging trends, feedback, and societal values over time.

Key Research Theme 3: Education and awareness. This research theme has a fundamental role in the challenges of *responsibility, liability, and accountability* and *impact on critical thinking and bias*. It also plays an important role in effective interdisciplinary collaboration and human-in-the-loop processes. This research theme emphasises the importance of enhancing LLM literacy among all stakeholders involved in the design, development, and deployment of these models. The potential (societal) impacts of such models (i.e. awareness) should also be of interest to those stakeholders tasked with or advocating for deploying these models, as doing so is never without consequence. The education and awareness research theme advocates for increasing awareness, managing expectations, promoting critical thinking, and increasing LLM literacy. By enhancing collective understanding of LLM strengths, limitations, and societal implications, this theme aims to equip individuals with the knowledge needed to engage with LLM technologies critically and constructively. Education and awareness (even in informal settings) efforts should extend beyond professional circles and interdisciplinary teams to reach the general public, empowering individuals with the knowledge to navigate the complexities of LLM technologies effectively, fostering a realistic understanding of the capabilities and constraints of these models, as well as a comprehensive grasp of their operational mechanisms and inherent limitations.

Key Research Theme 4: Audits. Adopting robust auditing mechanisms is a vital research theme, addressing challenges related to *responsibility, liability, and accountability*. Audits are envisioned as comprehensive and multidimensional evaluation frameworks that assess the ethical, technical, and social dimensions of LLMs, ensuring their alignment with established norms and values. Facilitating ethical training for engineers and technical experts, emphasising explainable AI and transparency, in addition to developing and implementing frequent and resilient auditing mechanisms, are approaches that can mitigate these challenges. This research theme can provide a solution to *false*

information and hallucination challenges, improve the trustworthiness of interaction with the intelligent systems, and help assess and increase the reliability of these systems' outputs when integrated with feedback and adjustment of the systems. Lastly, audits for LLM impact assessment can help find a solution to *privacy and human rights considerations*. That is to ensure that the impact assessments conducted are comprehensive and accurately reflective of the diverse societal implications. This ensures a balanced approach to AI deployment, safeguarding individual rights while maximising technological benefits.

The four key research themes discussed in this section, derived from collaborative workshop discussions, form the cornerstone of the research directions for the successful, effective, and responsible integration of LLMs into socio-technical systems. These themes not only highlight critical areas for immediate investigation but also suggest a roadmap for long-term and in-depth research endeavours concerning various identified challenges and related action plans. By emphasising Human-in-the-Loop research and methodologies, Multi-Stakeholder Longitudinal Studies, Education and Awareness initiatives, and comprehensive Audits, these themes collectively address the multifaceted nature of socio-technical system integration. This approach ensures that LLMs are developed and deployed in a manner that is not only technologically robust but also ethically sound, socially inclusive, and adaptable to the evolving landscape of societal needs and values. Consequently, pursuing these research themes promises to advance our understanding and implementation of LLMs in a way that harmonises innovation with human-centric values, fostering socio-technical systems that are resilient, equitable, and aligned with the broader objectives of societal well-being and progress.

Limitations. This paper described the challenges and outlined future directions for integrating LLMs into socio-technical systems employing a methodology comprising a workshop and expert collaborations. However, it is plausible that a number of limitations could have influenced the results and insights obtained. Namely, the reliance on workshop discussions, although valuable, may not fully encompass the diversity of perspectives and experiences from various stakeholders involved in LLM development and deployment. A significant constraint is the self-selection bias of participants in the workshop. This bias stems from the recruitment method, which likely attracted individuals already interested or invested in the subject matter, potentially skewing the diversity of viewpoints. Additionally, despite the wide array of expertise areas

covered, as listed in [Table 1](#), the participants' shared institutional background (all experts involved in the workshop but one were affiliated with the same university from diverse faculties) may limit the generalizability of the findings. Another limitation is that having participants of the workshop as co-authors of the paper could have introduced potential bias, as their involvement may have influenced the interpretation of the results. However, the writing process was intentionally kept open and transparent to ensure mutual accountability. This approach, combined with the use of an external observer role, structured reflection and documentation, and separation of roles, helped minimise bias while still leveraging participants' valuable insights and expertise, which were essential for advancing the study. Furthermore, the rapid pace of advancements in the field of LLMs necessitates continuous updating of the research agenda to stay aligned with emerging technologies and societal impacts. Recognizing these limitations is crucial for contextualising the results and guiding future research efforts to refine and expand upon the insights provided in this study.

5. Conclusion

This paper emphasises the significant potential of LLMs to transform socio-technical systems and the necessity for a meticulous, human-centric approach to their integration. It highlights the importance of collaboration between academia, industry, and policymakers in creating LLMs that are technologically sophisticated, ethically grounded, and socially advantageous. Addressing both the challenges and possibilities of LLMs, the paper advocates for a strategic integration that is mindful of techno-optimism and proactive in mitigating risks. The research underscores the need for transdisciplinary efforts and early stakeholder involvement, paving the way for holistic, value-driven solutions that transcend mere technological or user-centric focuses.

The urgency of preparing for LLM adoption is apparent, given its widespread current usage and the mix of outcomes it brings. A multifaceted strategy, involving interdisciplinary research, educational initiatives, tools for critical analysis, and a boost in independent research, is vital to manage its impact effectively. The paper also points out the limitations of relying solely on large technology companies for LLM advancements and studies, suggesting regulatory measures like channeling a part of their revenue into independent academic research. Furthermore, the paper stresses the need for a broader perspective in research, beyond financially or industrially motivated studies, to evaluate the integration of LLMs into socio-technical systems

objectively. Unlike previous technologies, such as social media, the impact of LLMs is more far-reaching, affecting multiple socio-technical systems and society at large. Therefore, a comprehensive and inclusive approach in research and implementation is essential, considering ethical, societal, and cross-systemic impacts. This calls for a robust and accountable framework for the development and deployment of LLMs, ensuring that their impact is beneficial and aligned with societal values.

Notes

1. <https://openai.com/chatgpt>
2. <https://deepmind.google/technologies/gemini>
3. <https://www.anthropic.com/news/claude-3-5-sonnet>
4. <https://llama.meta.com/>
5. <https://blog.google/technology/developers/google-gemma-2/>
6. <https://allenai.org/olmo>
7. <https://mistral.ai/technology>
8. <https://sites.research.google/med-palm/>
9. <https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai>
10. <https://www.ai4media.eu/>
11. <https://www.elise-ai.eu>
12. <https://www.humane-ai.eu>
13. <https://tailor-network.eu>
14. <https://www.vision4ai.eu/sra/>
15. We used Sci-bert (Beltagy, Lo, and Cohan 2019) and hierarchical clustering algorithm, and independently, LDA, and clustering algorithms, namely, k-means, and gpt-4.
16. Note: participants possessed multiple areas of expertise.
17. ChatGPT (version 4o) was employed to improve language and enhance the overall readability of this paper.
18. The abstracts (Appendix A) covered a wide range of application domains, demonstrating the extensive potential reach and versatility of LLMs in society. These domains ranged from education ($n = 2$), health ($n = 4$), and governance ($n = 2$) to risk and safety management ($n = 5$), with some overlaps, each offering unique insights into the integration and implications of LLMs. For instance, in the education domain, the abstracts highlight LLMs' capacity to democratise access to information (A1), adapt learning paths to varying cognitive skills (A1), and reshape or impact the educational paradigm by influencing intellectual virtues (A5), curiosity (A5), open-mindedness (A5), and creativity (A10). In the health domain, themes revolved around mental well-being (A11, A12), workforce sustainability (A2, A6, A12), the risks of medical disinformation (A8), and the potential for human-AI collaboration in clinical settings (A2). In the domain of governance, compliance with digital regulations (A9) and value-based LLMs for democratic societies (A4) are discussed. In risk and safety, the abstracts discuss quality assurance as crowdsourcing labour (A3), potential misuse by malicious users (A11), adversarial attacks (A13), LLMs use as a source for safety managers (A6),

meaning and context integration (A7), and inclusion of confidence as a way to increase trustworthiness (A10).

19. https://nytco-assets.nytimes.com/2023/12/NYT_Complaint_Dec2023.pdf
20. <https://githubcopilotlitigation.com/>

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Helma Torkamaan  <http://orcid.org/0000-0003-1094-4059>
 Steffen Steinert  <http://orcid.org/0000-0001-8784-7607>
 Maria Soledad Pera  <http://orcid.org/0000-0002-2008-9204>
 Olya Kudina  <http://orcid.org/0000-0001-5374-1687>
 Samuel Kernan Freire  <http://orcid.org/0000-0001-8684-0585>
 Himanshu Verma  <http://orcid.org/0000-0002-2494-1556>
 Sage Kelly  <http://orcid.org/0000-0002-1022-0467>
 Marie-Therese Sekwenz  <http://orcid.org/0000-0002-3686-6100>
 Jie Yang  <http://orcid.org/0000-0002-0350-0313>
 Karolien van Nunen  <http://orcid.org/0000-0001-6057-4557>
 Martijn Warnier  <http://orcid.org/0000-0002-4682-6882>
 Frances Brazier  <http://orcid.org/0000-0002-7827-2351>
 Oscar Oviedo-Trespalacios  <http://orcid.org/0000-0001-5916-3996>

References

- OpenAI. Achiam, Josh, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, et al. March 4, 2024. "GPT-4 Technical Report." arXiv:2303.08774. <https://doi.org/10.48550/arXiv.2303.08774>.
- Ai, Lin, Tharindu Kumarage, Amrita Bhattacharjee, Zizhou Liu, Zheng Hui, Michael Davinroy, James Cook, et al. October 11, 2024. "Defending Against Social Engineering Attacks in the Age of LLMs." arXiv:2406.12263. <https://doi.org/10.48550/arXiv.2406.12263>.
- AI, NIST. January 26, 2023. "Artificial Intelligence Risk Management Framework (AI RMF 1.0)." NIST AI 100-1. Gaithersburg, MD: National Institute of Standards and Technology (U.S.), NIST AI 100-1. <http://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>. <https://doi.org/10.6028/NIST.AI.100-1>.
- Beltagy, Iz, Kyle Lo, and Arman Cohan. November 2019. "SciBERT: A Pretrained Language Model for Scientific Text." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, edited by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, 3615–3620. Hong Kong, China: Association for Computational Linguistics. <https://aclanthology.org/D19-1371>.
- Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. March 1, 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*,

- 610–623. New York, NY, USA: Association for Computing Machinery. ISBN: 978-1-4503-8309-7. <https://doi.org/10.1145/3442188.3445922>.
- Blodgett, Su Lin, Solon Barocas, Hal Daumé III, and Hanna Wallach. July 2020. “Language (Technology) is Power: A Critical Survey of ‘Bias’ in NLP.” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL 2020. edited by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, 5454–5476. Online: Association for Computational Linguistics. <https://aclanthology.org/2020.acl-main.485>.
- Cheong, Inyoung, King Xia, K. J. Kevin Feng, Quan Ze Chen, and Amy X. Zhang. June 5, 2024. “(A)I Am Not a Lawyer, But...: Engaging Legal Experts Towards Responsible LLM Policies for Legal Advice.” In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24*, 2454–2469. New York, NY, USA: Association for Computing Machinery. ISBN:9798400704505. <https://doi.org/10.1145/3630106.3659048>.
- Choudhury, Avishek, and Hamid Shamszare. 2023. “Investigating the Impact of User Trust on the Adoption and Use of ChatGPT: Survey Analysis.” *Journal of Medical Internet Research* 25 (1): e47184. <https://doi.org/10.2196/47184>.
- Costanza-Chock, Sasha. 2020. *Design Justice: Community-Led Practices to Build the Worlds We Need*. The MIT Press. ISBN: 978-0-262-04345-8. <https://library.oapen.org/handle/20.500.12657/43542>.
- Cummings, M. L. 2006. “Automation and Accountability in Decision Support System Interface Design.” Accepted: 2014-09-24T18:48:54Z Publisher: Journal of Technology Studies, ISSN: <http://hdl.handle.net/1721.1/90321>. <https://dspace.mit.edu/handle/1721.1/90321?show=full>.
- Dahmani, Louisa, and Véronique D. Bohbot. April 14, 2020. “Habitual Use of GPS Negatively Impacts Spatial Memory during Self-Guided Navigation.” *Scientific Reports* 10 (1): 6310. ISSN: 2045-2322. <https://www.nature.com/articles/s41598-020-62877-0>. <https://doi.org/10.1038/s41598-020-62877-0>.
- De Angelis, Luigi, Francesco Baglivo, Guglielmo Arzilli, Gaetano Pierpaolo Privitera, Paolo Ferragina, Alberto Eugenio Tozzi, and Caterina Rizzo. 2023. “ChatGPT and the Rise of Large Language Models: The New AI-driven Infodemic Threat in Public Health.” *Frontiers in Public Health* 11 (8): 173:1–173:38. ISSN: 2296-2565. [10.3389/fpubh.2023.1166120/full](https://doi.org/10.3389/fpubh.2023.1166120/full).
- Dong, Chenhe, Yinghui Li, Haifan Gong, Miaoxin Chen, Junxin Li, Ying Shen, and Min Yang. 2022. “A Survey of Natural Language Generation.” *ACM Computing Surveys* 55 (8): 173:1–173:38. ISSN: 0360-0300. <https://doi.org/10.1145/3554727>.
- Doorn, Neelke. March 1, 2012. “Responsibility Ascriptions in Technology Development and Engineering: Three Perspectives.” *Science and Engineering Ethics* 18 (1): 69–90. ISSN: 1471-5546. <https://doi.org/10.1007/s11948-009-9189-3>.
- Dourish, Paul. December 1, 2003. “The Appropriation of Interactive Technologies: Some Lessons from Placeless Documents.” *Computer Supported Cooperative Work (CSCW)* 12 (4): 465–490. ISSN: 1573-7551. <https://doi.org/10.1023/A:1026149119426>.
- Downes, William. September 24, 1998. *Language and Society*, 532 pp. Cambridge University Press. ISBN: 978-0-521-45663-0.
- Drage, Eleanor, and Kerry Mackereth. October 10, 2022. “Does AI Debias Recruitment? Race, Gender, and AI’s ‘Eradication of Difference’.” *Philosophy & Technology* 35 (4): 89. ISSN: 2210-5441. <https://doi.org/10.1007/s13347-022-00543-1>.
- Eubanks, Virginia. January 23, 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*, 273 pp. St. Martin’s Publishing Group. ISBN: 978-1-4668-8596-7.
- EUR-Lex. 2012. “Charter of Fundamental Rights of the European Union EUR-Lex - 12012P/TXT - EN - EUR-Lex.” Official Journal of the European Union C 326/391. Doc ID: 12012P/TXT Doc Sector: 1 Doc Title: Charter of Fundamental Rights of the European Union Doc Type: P Usr_lan: en. 2012. https://eur-lex.europa.eu/eli/treaty/char_2012/oj.
- EUR-Lex. 2021. “Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts.” <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206>.
- EUR-Lex. 2022. “Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) (Text with EEA relevance).” Doc ID: 32022R2065 Doc Sector: 3 Doc Title: Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) (Text with EEA relevance) Doc Type: R Usr_lan: en. <http://data.europa.eu/eli/reg/2022/2065/oj/eng>.
- EUR-Lex. 2024. “Communication from the Commission – Commission Guidelines for Providers of Very Large Online Platforms and Very Large Online Search Engines on the Mitigation of Systemic Risks for Electoral Processes Pursuant to Article 35(3) of Regulation (EU) 2022/2065.” 52024XC03014. Doc ID: 52024XC03014 Doc Sector: 5 Doc Title: Communication from the Commission – Commission Guidelines for Providers of Very Large Online Platforms and Very Large Online Search Engines on the Mitigation of Systemic Risks for Electoral Processes Pursuant to Article 35(3) of Regulation (EU) 2022/2065 Doc Type: XC Usr_lan: en. <https://eur-lex.europa.eu/eli/C/2024/3014/oj/eng>.
- European Commission. 2023. “Europe’s Digital Decade: 2030 Targets | European Commission.” https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/europes-digital-decade-digital-targets-2030_en.
- European Union Agency for Fundamental Rights. November 29, 2022. “Bias in algorithms – Artificial Intelligence and Discrimination.” Vienna. <https://fra.europa.eu/en/publication/2022/bias-algorithm>.
- Freeman, Jonathan B. 2020. “Measuring and Resolving LGBTQ Disparities in STEM.” *Policy Insights from the Behavioral and Brain Sciences* 7 (2): 141–148. ISSN: 2372-7322. <https://doi.org/10.1177/2372732220943232>.
- Friedman, Batya, and David G. Hendry. May 21, 2019. *Value Sensitive Design: Shaping Technology with Moral Imagination*. MIT Press. ISBN: 978-0-262-03953-6.

- Gallegos, Isabel O., Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. **September 1, 2024**. “Bias and Fairness in Large Language Models: A Survey.” *Computational Linguistics* 50 (3): 1097–1179. ISSN: 0891-2017. https://doi.org/10.1162/coli_a_00524.
- Goddard, Kate, Abdul Roudsari, and Jeremy C. Wyatt. **January 1, 2012**. “Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators.” *Journal of the American Medical Informatics Association* 19 (1): 121–127. ISSN: 1067-5027. <https://doi.org/10.1136/amiajnl-2011-000089>.
- Gordon, Emile B., Alexander J. Towbin, Peter Wingrove, Umber Shafique, Brian Haas, Andrea B. Kitts, Jill Feldman, and Alessandro Furlan. **2024**. “Enhancing Patient Communication with Chat-GPT in Radiology: Evaluating the Efficacy and Readability of Answers to Common Imaging-Related Questions.” *Journal of the American College of Radiology* 21 (2): 353–359. ISSN: 1067-5027. <https://doi.org/10.1016/j.jacr.2023.09.011>.
- Hacker, Philipp, Andreas Engel, and Marco Mauer. **June 12, 2023**. “Regulating ChatGPT and Other Large Generative AI Models.” In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23*, 1112–1123. New York, NY, USA: Association for Computing Machinery. ISBN: 9798400701924. <https://doi.org/10.1145/3593013.3594067>.
- Haghani, Milad. **August 28, 2023**. “Riding the Wave of ChatGPT Research: An Analysis of Early-Stage Scholarly Output and the Associated Authorship Anomalies.” Rochester, NY. <https://papers.ssrn.com/abstract=4553479>. <https://doi.org/10.2139/ssrn.4553479>.
- Hu, Krystal. **February 2, 2023**. “ChatGPT Sets Record for Fastest-Growing User Base – Analyst Note.” Reuters. <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/>.
- IEEE. **2020**. “Recommended Practice for Organizational Governance of Artificial Intelligence | StandICT.eu 2026.” P2863. <https://standict.eu/standards-repository/recommended-practice-organizational-governance-artificial-intelligence>.
- IEEE Standards Committee and Joanna Isabelle Olszewska. **May 1, 2020**. “IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being: IEEE Standard 7010-2020.” Piscataway, NJ. <https://doi.org/10.1109/IEEESTD.2020.9084219>.
- ISO. **2021**. “ISO/IEC TR 24027:2021 Information technology – Artificial Intelligence (AI) – Bias in AI Systems and AI Aided Decision Making.” <https://www.iso.org/standard/77607.html>.
- Jakesch, Maurice, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. **April 19, 2023**. “Co-Writing with Opinionated Language Models Affects Users’ Views.” In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23*, 1–15. New York, NY, USA: Association for Computing Machinery. ISBN: 978-1-4503-9421-5. <https://doi.org/10.1145/3544548.3581196>.
- Kelly, Sage, Sherrie-Anne Kaye, and Oscar Oviedo-Trespalacios. **2022**. “A Multi-Industry Analysis of the Future Use of AI Chatbots.” *Human Behavior and Emerging Technologies* 2022 (1): 2552099. ISSN: 2578-1863. <https://doi.org/10.1155/2022/2552099>.
- Kelly, Sage, Sherrie-Anne Kaye, and Oscar Oviedo-Trespalacios. **February 2023**. “What Factors Contribute to the Acceptance of Artificial Intelligence? A Systematic Review.” *Telematics and Informatics* 77:101925. ISSN: 07365853. <https://doi.org/10.1016/j.tele.2022.101925>.
- Kroes, Peter, Maarten Franssen, Ibo van de Poel, and Maarten Ottens. **2006**. “Treating Socio-Technical Systems as Engineering Systems: Some Conceptual Problems.” *Systems Research and Behavioral Science* 23 (6): 803–814. ISSN: 1099-1743. <https://doi.org/10.1002/sres.703>.
- Kruspe, Anna. **April 3, 2024**. “Towards Detecting Unanticipated Bias in Large Language Models.” arXiv:2404.02650. <https://doi.org/10.48550/arXiv.2404.02650>.
- Lakoff, George, and Mark Johnson. **December 19, 2008**. *Metaphors We Live by*, 292 pp. University of Chicago Press. ISBN: 978-0-226-47099-3.
- Li, Junyi, Tianyi Tang, Wayne Xin Zhao, and Ji-Rong Wen. **August 9, 2021**. “Pretrained Language Model for Text Generation: A Survey.” In *Twenty-Ninth International Joint Conference on Artificial Intelligence*, Montreal, Vol. 5, 4492–4499. ISSN: 1045-0823. <https://www.ijcai.org/proceedings/2021/612>. <https://doi.org/10.24963/ijcai.2021/612>.
- Lo, Chung Kwan. **April 2023**. “What is the Impact of ChatGPT on Education? A Rapid Review of the Literature.” *Education Sciences* 13 (4): 410. ISSN: 2227-7102. <https://doi.org/10.3390/educsci13040410>.
- Lopez, Paola. **December 7, 2021**. “Bias Does Not Equal Bias: A Socio-Technical Typology of Bias in Data-Based Algorithmic Systems.” *Internet Policy Review* 10 (4): 29. <https://policyreview.info/articles/analysis/bias-does-not-equal-bias-socio-technical-typology-bias-data-based-algorithmic>. <https://doi.org/10.14763/2021.4.1598>.
- Mack, Kelly Avery, Rida Qadri, Remi Denton, Shaun K. Kane, and Cynthia L. Bennett. **May 11, 2024**. “‘They Only Care to Show Us the Wheelchair’: Disability Representation in Text-To-Image AI Models.” In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, CHI '24*, 1–23. New York, NY, USA: Association for Computing Machinery. ISBN: 9798400703300. <https://doi.org/10.1145/3613904.3642166>.
- Mariano, C. **November 1, 1989**. “The Case for Interdisciplinary Collaboration.” *Nursing Outlook* 37 (6): 285–288. ISSN: 1528-3968.
- Min, Bonan, Hayley Ross, Elinor Sulem, Amir Pournan Ben Veysseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. **September 14, 2023**. “Recent Advances in Natural Language Processing Via Large Pre-Trained Language Models: A Survey.” *ACM Computing Surveys* 56 (2): 30:1–30:40. ISSN: 0360-0300. <https://doi.org/10.1145/3605943>.
- Minaee, Shervin, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. **2024**. “Large language models: A survey.” arXiv preprint arXiv:2402.06196. <https://arxiv.org/abs/2402.06196>.
- Mlynar, Jakub, Farzaneh Bahrami, André Ourednik, Nico Mutzner, Himanshu Verma, and Hamed Alavi. **April 29, 2022**. “AI Beyond Deus ex Machina – Reimagining Intelligence in Future Cities with Urban Experts.” In

- Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI '22*, 1–13. New York, NY, USA: Association for Computing Machinery. ISBN: 978-1-4503-9157-3. <https://doi.org/10.1145/3491102.3517502>.
- Mökander, Jakob, and Luciano Floridi. June 1, 2021. “Ethics-Based Auditing to Develop Trustworthy AI.” *Minds and Machines* 31 (2): 323–327. ISSN: 1572-8641. <https://doi.org/10.1007/s11023-021-09557-8>.
- Monarch (Munro), Robert. August 17, 2021. *Human-In-the-Loop Machine Learning: Active Learning and Annotation for Human-Centered AI*, 422 pp. Simon and Schuster. ISBN: 978-1-63835-103-0.
- Murgia, Emiliana, Zahra Abbasiantaeb, Mohammad Aliannejadi, Theo Huibers, Monica Landoni, and Maria Soledad Pera. June 16, 2023. “ChatGPT in the Classroom: A Preliminary Exploration on the Feasibility of Adapting ChatGPT to Support Children’s Information Discovery.” In *Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization, UMAP '23 Adjunct*, 22–27. New York, NY, USA: Association for Computing Machinery. ISBN: 978-1-4503-9891-6. <https://doi.org/10.1145/3563359.3597399>.
- Murgia, Emiliana, Maria Soledad Pera, Monica Landoni, and Theo Huibers. June 16, 2023. “Children on ChatGPT Readability in an Educational Context: Myth or Opportunity?” In *Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization, UMAP '23 Adjunct*, 311–316. New York, NY, USA: Association for Computing Machinery. ISBN: 978-1-4503-9891-6. <https://doi.org/10.1145/3563359.3596996>.
- Navigli, Roberto, Simone Conia, and Björn Ross. June 22, 2023. “Biases in Large Language Models: Origins, Inventory, and Discussion.” *Journal Data and Information Quality* 15 (2): 10:1–10:21. ISSN: 1936-1955. <https://doi.org/10.1145/3597307>.
- Nikkei Asia. n.d.. “China Tells Big Tech Companies Not to Offer ChatGPT Services.” Nikkei Asia. <https://asia.nikkei.com/Business/China-tech/China-tells-big-tech-companies-not-to-offer-ChatGPT-services>.
- Noble, Safiya Umoja. February 20, 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press. ISBN: 978-1-4798-3364-1. <https://www.degruyter.com/document/doi/10.18574/nyu/9781479833641.001.0001/html>.
- Novelli, Claudio, Federico Casolari, Philipp Hacker, Giorgio Spedicato, and Luciano Floridi. March 15, 2024. “Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity.” arXiv:2401.07348. <https://doi.org/10.48550/arXiv.2401.07348>.
- OpenAI. 2024. “OpenAI Platform.” Retrieved June 14, 2024. <https://platform.openai.com>.
- Oviedo-Trespalacios, Oscar, Amy E. Peden, Thomas Cole-Hunter, Arianna Costantini, Milad Haghani, J. E. Rod, Sage Kelly. November 1, 2023. “The Risks of Using ChatGPT to Obtain Common Safety-Related Information and Advice.” *Safety Science* 167:106244. ISSN: 0925-7535. <https://doi.org/10.1016/j.ssci.2023>.
- Prabhakaran, Vinodkumar, Ben Hutchinson, and Margaret Mitchell. November 2019, 2019. “Perturbation Sensitivity Analysis to Detect Unintended Model Biases.” In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), EMNLP-IJCNLP 2019*, edited by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, 5740–5745. Hong Kong, China: Association for Computational Linguistics. <https://aclanthology.org/D19-1578>. <https://doi.org/10.18653/v1/D19-1578>.
- Raji, Inioluwa Deborah, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. January 27, 2020. “Closing the AI Accountability Gap: Defining an End-To-End Framework for Internal Algorithmic Auditing.” In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20*, 33–44. New York, NY, USA: Association for Computing Machinery. ISBN: 978-1-4503-6936-7. <https://doi.org/10.1145/3351095.3372873>.
- BigScience Workshop, Scao, Teven Le, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, et al. June 27, 2023. “BLOOM: A 176B-Parameter Open-Access Multilingual Language Model.” arXiv:2211.05100. <https://doi.org/10.48550/arXiv.2211.05100>.
- Schappert, Stefania. March 24, 2023. “ChatGPT Leaks User Credit Card Details.” <https://cybernews.com/news/payment-info-leaked-openai-chatgpt-ouage/>.
- Scharowski, Nicolas, Sebastian A. C. Perrig, Melanie Svab, Klaus Opwis, and Florian Brühlmann. July 17, 2023. “Exploring the Effects of Human-Centered AI Explanations on Trust and Reliance.” *Frontiers in Computer Science* 5:15. ISSN: 2624-9898. <https://doi.org/10.3389/fcomp.2023.1151150>.
- Scopus. n.d.. “Scopus Preview – Scopus – Welcome to Scopus.” Scopus. <https://www.scopus.com/home.uri>.
- Shneiderman, Ben. January 13, 2022. *Human-Centered AI*, 390 pp. Oxford University Press. ISBN: 978-0-19-266000-8.
- Shneiderman, Ben. October 16, 2020a. “Bridging the Gap between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-Centered AI Systems.” *ACM Transactions on Interactive Intelligent Systems* 10 (4): 26:1–26:31. ISSN: 2160-6455. <https://doi.org/10.1145/3419764>.
- Shneiderman, Ben. September 30, 2020b. “Human-Centered Artificial Intelligence: Three Fresh Ideas.” *AIS Transactions on Human-Computer Interaction* 12 (3): 109–124. ISSN: 1944-3900. <https://doi.org/10.17705/1thci.00131>.
- Singhal, Karan, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales. 2023. “Large Language Models Encode Clinical Knowledge.” *Nature* 620 (7972): 172–00. <https://doi.org/10.1038/s41586-023-06291-2>.
- Skitka, Linda J., Kathleen Mosier, and Mark D. Burdick. April 1, 2000. “Accountability and Automation Bias.” *International Journal of Human-Computer Studies* 52 (4): 701–717. ISSN: 1071-5819. <https://doi.org/10.1006/ijhc.1999.0349>.
- Sousa, Sonia, David Lamas, José Cravino, and Paulo Martins. March 2024. “Human-Centered Trustworthy Framework: A Human-Computer Interaction Perspective.” *Computer* 57 (3): 46–58. ISSN: 1558-0814. <https://doi.org/10.1109/MC.2023.3287563>.
- Stertz, Sarah, Kevin Baum, Sebastian Biewer, Holger Hermanns, Anne Lauber-Rönsberg, Philip Meinel, and Markus Langer.

2024. "On the Quest for Effectiveness in Human Oversight: Interdisciplinary Perspectives." In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24*, 2495–2507. New York, NY, USA: Association for Computing Machinery. ISBN: 9798400704505. <https://doi.org/10.1145/3630106.3659051>.
- Team, Gemini, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, et al. **June 17, 2024**. "Gemini: A Family of Highly Capable Multimodal Models." arXiv:2312.11805. <https://doi.org/10.48550/arXiv.2312.11805>.
- Team, Gemini, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, et al. **August 8, 2024**. "Gemini 1.5: Unlocking Multimodal Understanding Across Millions of Tokens of Context." arXiv:2403.05530. <https://doi.org/10.48550/arXiv.2403.05530>.
- Thirunavukarasu, Arun James, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. **August 2023**. "Large Language Models in Medicine." *Nature Medicine* 29 (8): 1930–1940. ISSN: 1546-170X. <https://doi.org/10.1038/s41591-023-02448-8>.
- Thomas, Linda, and Shân Wareing. **May 27, 1999**. *Language, Society and Power: An Introduction*, 240 pp. London: Routledge. ISBN: 978-0-203-42696-8. <https://doi.org/10.4324/9780203426968>.
- Torkamaan, Helma, Mohammad Tahaei, Stefan Buijsman, Ziang Xiao, Daricia Wilkinson, and Bart P. Knijnenburg. **2024**. "The Role of Human-Centered AI in User Modeling, Adaptation, and Personalization-Models, Frameworks, and Paradigms." In *A Human-Centered Perspective of Intelligent Personalized Environments and Systems*, edited by Bruce Ferwerda, Mark Graus, Panagiotis Germanakos, Marko Tkalcic, 43–84. Cham: Springer Nature Switzerland. ISBN: 978-3-031-55109-3. https://doi.org/10.1007/978-3-031-55109-3_2.
- Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, et al. **February 27, 2023**. "LLaMA: Open and Efficient Foundation Language Models." arXiv:2302.13971. <https://doi.org/10.48550/arXiv.2302.13971>.
- Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, et al. **July 19, 2023**. "Llama 2: Open Foundation and Fine-Tuned Chat Models." arXiv:2307.09288. <https://doi.org/10.48550/arXiv.2307.09288>.
- Trudgill, Peter. **January 16, 1991**. *Dialects in Contact*, 174 pp. 1st ed. Oxford: Wiley-Blackwell. ISBN: 978-0-631-12733-8.
- Ullah, Ehsan, Anil Parwani, Mirza Mansoor Baig, and Rajendra Singh. **2024**. "Challenges and Barriers of Using Large Language Models (LLM) Such as ChatGPT for Diagnostic Medicine with a Focus on Digital Pathology – a Recent Scoping Review." *Diagnostic Pathology* 19 (1): 43. ISSN: 1746-1596. <https://doi.org/10.1186/s13000-024-01464-7>.
- van den Hoven, Jeroen, Pieter E. Vermaas, and Ibo van de Poel. **2015**. "Design for Values: An Introduction." In *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*, edited by Jeroen van den Hoven, Pieter E. Vermaas, and Ibo van de Poel, 1–7. Dordrecht: Springer Netherlands. ISBN: 978-94-007-6970-0. https://doi.org/10.1007/978-94-007-6970-0_40.
- Verma, Himanshu, Jakub Mlynar, Roger Schaer, Julien Reichenbach, Mario Jreige, John Prior, Florian Evéquoz, and Adrien Depeursinge. **April 19, 2023**. "Rethinking the Role of AI with Physicians in Oncology: Revealing Perspectives from Clinical and Research Workflows." In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23*, 1–19. New York, NY, USA: Association for Computing Machinery. ISBN: 978-1-4503-9421-5. <https://doi.org/10.1145/3544548.3581506>.
- Winograd, Amy. **2022**. "Loose-Lipped Large Language Models Spill Your Secrets: The Privacy Implications of Large Language Models." *Harvard Journal of Law & Technology (Harvard JOLT)* 36:615. <https://heinonline.org/HOL/Page?handle=hein.journals/hjlt36&id=622> &div=&collection=.
- Wu, Kewen, Yuxiang Zhao, Qinghua Zhu, Xiaojie Tan, and Hua Zheng. **December 1, 2011**. "A Meta-Analysis of the Impact of Trust on Technology Acceptance Model: Investigation of Moderating Influence of Subject and Context Type." *International Journal of Information Management* 31 (6): 572–581. ISSN: 0268-4012. <https://doi.org/10.1016/j.ijinfomgt.2011.03.004>.
- Xiao, Anran, Zeshui Xu, Marinko Skare, Yong Qin, and Xinxin Wang. **June 21, 2024**. "Bridging the Digital Divide: The Impact of Technological Innovation on Income Inequality and Human Interactions." *Humanities and Social Sciences Communications* 11 (1): 1–18. ISSN: 2662-9992. <https://doi.org/10.1057/s41599-024-03307-8>.
- Yoo, Dong Whi, Hayoung Woo, Sachin R. Pendse, Nathaniel Young Lu, Michael L. Birnbaum, Gregory D. Abowd, and Munmun De Choudhury. **April 26, 2024**. "Missed Opportunities for Human-Centered AI Research: Understanding Stakeholder Collaboration in Mental Health AI Research." *Proceedings of the ACM on Human-Computer Interaction* 8 (CSCW1): 95:1–95:24. <https://doi.org/10.1145/3637372>.
- Zack, Travis, Eric Lehman, Mirac Suzgun, Jorge A. Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, et al. **July 17, 2023**. "Coding Inequity: Assessing GPT-4's Potential for Perpetuating Racial and Gender Biases in Healthcare." <https://doi.org/10.1101/2023.07.13.23292577>.
- Zhang, Chiyuan, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. **December 15, 2023**. "Counterfactual Memorization in Neural Language Models." *Advances in Neural Information Processing Systems* 36:39321–39362. https://proceedings.neurips.cc/paper_files/paper/2023/hash/7bc4f74e35bcfe8cf e43b0a860786d6a-Abstract-Conference.html.
- Zhang, Yizhou, Karishma Sharma, Lun Du, and Yan Liu. **May 13, 2024**. "Toward Mitigating Misinformation and Social Media Manipulation in LLM Era." In *Companion Proceedings of the ACM Web Conference 2024, WWW '24*, 1302–1305. New York, NY, USA: Association for Computing Machinery. ISBN: 9798400701726. <https://doi.org/10.1145/3589335.3641256>.
- Zhang, Hanqing, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. **October 6, 2023**. "A Survey of Controllable Text Generation Using Transformer-Based Pre-Trained Language Models." *ACM Computing Surveys* 56 (3): 64:1–64:37. ISSN: 0360-0300. <https://doi.org/10.1145/3617680>.