Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems

RISHABH DEVGON, Indraprastha Institute of Information Technology, Delhi, India

CCS Concepts: • **Human-centered computing → HCI theory, concepts and models**; • **Computing methodologies → Philosophical/theoretical foundations of artificial intelligence**.

Additional Key Words and Phrases: Human-centered AI, Human-Computer Interaction, Artificial Intelligence, reliable, safe, trustworthy, software engineering practices, management strategies, independent oversight, design

ACM Reference Format:
Rishabh Devgon. 2020. Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. In *ATHCC '21: ATHCC '21: Critical Review, January 26, 2021, New Delhi, IN*. ACM, New York, NY, USA, 2 pages.

1 CRITICAL REVIEW

Ben Shneiderman highlights the importance of shifting to Human Centered Artificial Intelligence from traditional approaches to avoid negative consequences[1]. HCAI differs from AI because it focuses on amplifying, augmenting and enhancing human performance rather than emulating human behaviour. HCAI uses the participatory design approach and makes human users central to the process. There is mention of a three-layered governance structure for HCAI systems to bridge the gap between principles and practise. The design and recommendations provided in the paper are although not novel, but they are integral to imbibe qualities such as reliability, safety, trustworthiness and addressing concerns such as privacy, security, environmental protection, social justice, and human rights.

The first layer is about reliable systems based on sound software engineering practices putting in use firstly, audit trails and analysis tools which can act as preventative maintenance measures and for practical retrospective forensic analysis. Secondly, software engineering workflows must include user experience design in order to gauge the repercussions of made decisions. Thirdly, verification and validation testing with large, robust test sets are vital to ensure that an HCAI system performs as per the user's wishes and comfort without any unintended consequences. Fourthly, bias testing to enhance fairness. And fifth, creating explainable user interfaces helps users understand what is actually happening inside the black box of an AI system. The second layer is based on safety culture through business management strategies exploring themes such as leadership commitment to safety through explicit mentions of values, beliefs, policies and norms, hiring and training oriented to safety, extensive reporting of failures and near misses, internal review boards for problems and future plans and finally, alignment with industry-standard practices. The third layer talks about trustworthy certification by independent oversight and explores legal liability, professional accountability, moral responsibility, and ethical bias. It touches upon government interventions and regulations, accounting forms conducting external audits, insurance companies compensating for AI failures and professional organizations and research institutes.

The research methods used are extensive literature reviews and case studies. The methods used are justified, given that the paper collates various insights from work all around HCAI and provides its overview. I believe that the paper

2020. Manuscript submitted to ACM

lacked the human centered aspect which could have been included to also add to experiential understanding of the domain as a whole. The author mentions during the introduction that here are some practitioners who resist moving towards a more human centered approach for AI, however, in the paper I did not find any arguments that shed light on this branch of thinking which makes me believe that perhaps the study is a little biased in nature.

A point to reflect on during the testing is to make the study more participatory and pluralistic by having a diverse set of external critics and participants which will contribute to making the HCAI system more robust and reduce the consequent biases. I felt that even though the author has managed to take into account a lot of stakeholders, there was still a lack of focus on the user. The author has focused on participation, he has not really explored the context of use and the user groups' ecology and what effect would the system have on a meta-level. Of course, the argument against making AI systems more explainable is finding the line where the explanation does not reveal too much for other institutions to plagiarise and makes it difficult for people to follow. Through business management strategies, the safety culture could benefit from collaboration and data sharing within the industry to train AI models better to make them safer. Another potential addition that I can see is the dissemination of safe practices to better train individuals to align themselves with the goals of HCAI. The oversight could massively benefit from a more holistic and global approach because of the migratory nature of companies to move to underdeveloped spaces with lax policies, which is witnessed frequently. Taking a more critical approach about each of the themes explored by the author could shed light on deeper explorations of ethics and moral responsibility and the misuse of AI.

REFERENCES

[1] Ben Shneiderman. 2020. Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-Centered AI Systems. *ACM Trans. Interact. Intell. Syst.* 10, 4, Article 26 (Oct. 2020), 31 pages. https://doi.org/10.1145/3419764