Recognizing the Problem of High Trust

SHIVANI KAPANIA, Google Research, India

Our everyday life experiences are increasingly mediated by AI-based systems— which continue to proliferate in high-stakes areas such as loan assessment, hiring, and health. Prior work on human-AI interaction is majorly set in communities that indicate mistrust, albeit with conflicting results. We know much less about contexts where AI is viewed aspirationally. In this research, we investigated the perceptions around AI systems by drawing upon 32 interviews with adult Internet users, who indicate a high trust towards AI, with attitudes such as faith and self-blame. Attitudes of high trust, faith and self-blame for AI outcomes, lead to vulnerability, pointing to higher tolerance for system errors and misfires, introducing potential for real, irreversible harm. We call for recognizing high trust as a distinct problem. High (over-) and low (under-) trust have different effects, exist in different contexts, and often affect different communities. We call for explicitly embracing these differences in HCI research and system design, calibrating high trust, and reconsidering the use of trust as a measure for success.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; • **Social and professional topics** → Cultural characteristics.

Additional Key Words and Phrases: AI, algorithmic decision-making, AI perceptions, India, AI authority, user attitudes, algorithmic perceptions

ACM Reference Format:
Shivani Kapania. 2022. Recognizing the Problem of High Trust. In *Workshop on Trust and Reliance in AI-Human Teams at CHI, New Orleans, LA.* ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

A growing body of research on public attitudes toward algorithmic systems indicates skepticism and moderate acceptance towards these technologies in contexts such as the US, UK and Germany [18], where studies report that individuals express concerns about the fairness and usefulness of these systems [17, 39]. The research on improving human-AI interactions is thus often set in communities where studies indicate user mistrust towards algorithmic systems[1] [7, 21, 25]. However, the acceptance among users may be shaped by specific online trajectories and exposure levels, possibly not generalizing to contexts with newer Internet citizens from under-researched socio-cultural settings.

AI deployments in India are emerging in several niche, high-stakes areas (healthcare [34, 42], finance [38], agriculture [9]). Marda [26] describe how AI is also emerging as a focus for policy development in India. Prior research, however, presents a case for techno-optimism among technology users that envision technology with a socio-economic promise [31] for India. As the world's second largest Internet user population [16], it is important to understand how Indian users perceive AI systems.

We draw from 32 interviews set in India, to report the perceptions of trust towards AI-based outcomes and the user attitudes which accompany these perceptions. Our work has implications for the design of responsible AI systems, by

[1] the results on perceptions of trust and fairness are domain-dependent and mixed, but predominantly negative in contexts such as the US and EU

Manuscript submitted to ACM

highlighting that users from under-studied settings could have different attitudes towards and behaviors with AI due to the contextual or non-technological factors. Our results indicate that there is already a high trust towards adopting AI systems among Indian technology users, which means we must approach design and research differently (*e.g.,* by supporting people in maintaining a healthy distrust and critical awareness of AI). Promoting trust, even as an implicit sub-goal, which is often the case with building transparent and explainable AI systems, can lead to adverse effects for communities which already have high initial trust towards AI-based outcomes. We call for recognizing that levels of trust are heavily influenced by the contextual factors with a potential to generate unwarranted trust in AI.

2 METHODOLOGY

We conducted interviews with 32 adult Internet users based in different regions within India to understand perceptions of AI and their acceptance of AI-based decisions. Participants were located in various tier 1 and tier 2 cities of India, belonging to diverse age groups and occupations. Our sample also consisted of a mix of internet experience, with 16 nascent internet users that first accessed the internet only in the last 2 years, and the remaining 16 more 'experienced' that have been online for more than 2 years. We interviewed 17 male and 15 female participants. We recruited participants through a combination of a UXR database internal to our organization, and a market research company Dowell. We conducted 'screener' conversations with potential participants to ask if they had heard of the words 'Artificial Intelligence'. Each interview session lasted 75-90 minutes each. We conducted interviews in three languages (Hindi, Tamil and English). Each participant received a thank-you gift in the form of a gift card worth 27 USD or 2000 INR.

Interview structure. In line with our goal, each interview had structured sub-sections beginning with the participants' general perceptions and understanding of AI. We asked participants to share "in [their] opinion, what do [they] think Artificial Intelligence means?" and "overall, do [they] think AI is doing more good or more harm for us? Why? In what ways?" In the main part of the interview, we used a scenario-based approach to probe experiences, attitudes and intentions to act. For each scenario, we began by presenting a scenario description accompanied by visuals, and then invited participants to share their initial reactions. We asked questions around their beliefs and intentions (e.g. if AI made a decision on your loan application, would you believe it to be correct?), their preferences for human vs AI decision-maker (e.g. what are the differences between a human making a decision on your loan application and an AI making that decision?), the kinds of information they would like to know about the AI application, and the level of control they believe AI systems should have in the given scenario.

Scenario selection. We draw inspiration from prior research on examining various perceptions of trust, fairness, justice, and explanations in using scenarios. We used four scenarios in the study, 1) loan approval, 2) financial advising, 3) medical diagnosis, 4) hospital finder. The scenarios were a part of a vignette experiment with 2 (decision support vs. decision-making) x 2 (health vs. finance) within subjects design. We randomized the four scenarios across these participants using Latin Square Design to control for ordering effects, if any.

Interview analysis. After transcribing, we translated all the Hindi and Tamil interviews into English, our primary language of analysis. We followed the qualitative data analysis approach by Thomas to conduct multiple rounds of coding at the sentence and paragraph level. We began by open coding instantiations of perceptions, assumptions and expectations of/around AI. Two members of the research team independently read all units multiple times, and identified categories (unit of analysis) together with a description and examples of each category, until a saturation point was reached.

3 FINDINGS

Interview participants indicated a willingness to take high-risk actions based on AI-based outcomes. Several participants noted an acceptance of the medical diagnosis by AI, with an openness to undergo medical treatment, for instance, *"the AI has given me a diagnosis. I will go to the hospital and get the necessary treatment. [...] I have more confidence on an AI instead of a doctor."* (P21) Participants were willing to accept loan assessments, and in fact, preferred that AI evaluated their application over a human loan officer. For the decision-support scenarios, most participants reported an inclination to follow AI recommendations (visit a hospital, budget income, invest in financial schemes), often mediated by their own judgment. As P11 expressed, *"I would definitely invest in businesses suggested by an AI, because I am 100% sure that it would be the right recommendation."* During each interview, we explored situations in which participants might receive contradictory recommendations from a human expert an AI. Several participants reported not only an intention to act on AI outcomes, but also an inclination to follow AI-based recommendations over those by a doctor or financial advisor. As P7 described, *"I would definitely go with AI if there is a contradiction. There is a clear logic behind an AI. "* Overall, participants demonstrated a tendency to accept AI decisions as correct, and believed AI was worthy of this trust unless they had evidence to the contrary.

Participants' beliefs about the AI system's high competence and benevolence underpin their trust towards the system, and different levels of trust may result in diverging intentions or behaviors. AI was considered *reliable*, that it is high-performing and capable of making decisions or recommendations. Participants also perceived AI as *infallible* and emotionless— that systems are based on facts, logic, rules and conditions or parameters, for example, a loan assessment AI looks at whether an individual's salary is above a certain threshold to determine whether they are eligible for a loan. The following attitudes towards AI made participants more likely to accept and rely on AI-based outcomes, even in situations where they receive a wrong or unfavorable outcome.

Faith. A commonly held attitude towards AI was a faith in its capabilities and intent, which persisted even when interviewees received an unfavorable outcome. Participants maintained that AI is between 90% to 99% accurate[2], with a recurrent estimate of 95% accurate. For example, P28 suggested that they *"might face an issue 1 out of 100 times. It is very rare."* for a medical diagnosis application. A normative perception across our interviews was that "computers do not make mistakes like humans do". Many participants had the faith that AI would only provide an outcome if it is confident in its abilities. As P9 suggested, *"I will believe an AI because it will not tell me such a big thing that I have a [blood] clot, unless it is confident in itself. "*

Participants ascribed neutral or good intentions to the AI systems in the scenarios, as opposed to their perception of humans malicious motives to gain out of their circumstances. They believed that AI, as a tool, does not cheat or fraud, simply because it would not receive any benefit from doing that. The only perceived goal of an AI system was to provide an outcome, which was in contrast with participants' prior experiences with human institutions. As P19 suggested, *"AI has nothing to gain, not a fixed salary or a commission or incentive. AI is neutral, it is a machine, it is a robot. So the suggestions from an AI will also be good. ''* Prior work has documented the ways in which 'information' and its use is considered inherently objective, free of personal opinion [3, 33]. Our research confirmed this perspective— AI is seen as fair because it is 'driven by data'. In participants' rationale for accepting AI decisions, the data driven nature of AI was limited to their own data, as opposed to curated training datasets that can be fraught with issues [35]. P2 reported, *"if the system/AI is doing that, it will not take into account that factor of human judgment. In a way, it is good*

[2]Participants did not specifically use technical jargon like accuracy. They expressed the percentage of times they believed an AI system would give the 'right/correct' outcome, which we loosely translate to accuracy.

*that **everyone will be treated impartially or fairly.** The one who deserves the loan and meets the full requirements, gets the loan. The system will take into account every aspect of your ability to pay the loan."*

In addition, AI was seen as a system more capable of fair decision-making than human institutions. Where human processes were riddled with inconsistencies and manipulation, AI was seen as simply a rule-following, clearly specifiable system that took into consideration every parameter that should be a part of the decision making process, and *thus fair*. P12 expressed their intention to use loan assessment AI systems because, *"officers make you go through so many procedures for a single approval. It will be easy if you have connections at the office. Otherwise, they will ask you to visit one counter after another, and make you wait in long queues. It is simply exhausting. "* Colquitt [8] presents four types of justice: distributive, procedural, interactional, informational. Participants consider the procedures, workflows, practices of institutions as a frequent source of unfairness (procedural and interactional injustice). These findings point to a need to reorient our research from an exclusive focus on outcomes towards a perspective which takes into consideration the entire process and the various interactions it encompasses.

Self-blame. Participants believed that AI is infallible, that it makes the right decisions and gives the correct outcomes. Users perceived that AI is built by specifying conditions (like a rule-based system, if-else), and there was not much scope of questioning the AI outcomes. This notion of being based on clear conditions manifested into self-blame. Participants consistently blamed themselves for receiving an adverse outcome, especially in the loan assessment scenario. As P32 described, *"If we give proper documents, it [AI] will give loan, else there must be some problem with our documents. "* Users conjectured that an unfavorable outcome was their error, because an AI rarely ever makes mistakes. For example, participants had a tendency to believe a wrong outcome by AI meant they did not correctly input their medical history into the health application, or entered the wrong location for finding a hospital, or did not upload the necessary documents for financial advising.

Participants viewed AI systems as emotionless, logical entities, therefore, it was perceived that there was little scope for true errors. As P18 mentioned, *"the decision is right if we keep our emotions aside and think logically. If they [institutions/developers] have certain rules, then they will decide based on those rules. If we did not fulfill those requirements, then we did not receive the loan."* Moreover, in some cases, participants believed that they deserved to receive an unfavorable outcome. For instance, users speculated that a loan rejection was *"based on [their] transactions"* (P4), and meant they had ineligible finances or collateral to receive a loan. Overall, participants found various ways to deflect blame away from the AI application onto themselves or others. Self-blame is a dangerous attitude, not only because of its widespread nature, but also because it almost always had no basis. There is an urgent need to combat self-blame, or else users might not recognize when an outcome is biased, unfair. Users might not seek recourse or alternative opportunities if they believe that they received an unfavorable outcome, but one that is deserved.

Factors influencing trust. The factors which contributed to trust, and thus, AI system being perceived as an authority, were often extraneous to the system, and unindicative of the reliability, usefulness, or effectiveness of the given AI system. Trust was heavily influenced by institutional, infrastructural and societal factors that lay outside the boundaries of the AI—users' interactions with ineffectual institutions, polarized narratives and misleading terminologies, users' prior experiences with technology, and the availability of human systems around them. For example, many participants were left exasperated by the corrupt or discriminatory practices in their interactions with financial institutions, which is why they perceived AI as a better alternative to avoid those forms of exploitation. Several interviewees narrated their experiences with bureaucratic decision-making institutions (not just banks), with corruptible officers demanding a bribe as the only way to approve their application. AI was consistently seen as a mechanism to avoid encountering corrupt or prejudiced practices in their interactions with institutions. As P26 expressed,

> *"if a lower middle class individual visits a bank, they ask for so much documentation which you cannot fulfill. The biggest reason for this is that bribes work in many banks for sanctioning loans in India. I have faced this myself. If you go to a bank, the mediators will get a hold of you and ask for 15-20% of the loan amount. Then they will clear all your documents. If there is an AI in this place, then there would not be any issue."*

The narratives about AI that surfaced in our research often originated from media (Sci-Fi movies like the Terminator), news articles, or government perspectives and initiatives. Most descriptions of AI were polarized and extreme: mostly optimistic, and rarely very pessimistic. The realistic, shades-of-grey narratives were often missing in participants' descriptions of AI, its benefits and harms. People carried an optimism about AI owing to the breadth of coverage about AI's potential, applicability, and planned use by the government. As P16 described, *"the government has made everything an online system nowadays. That cannot go wrong at all."* Several participants brought up the promise of various technological deployments (*"Modi is launching a driverless train"* (P28)) as a way to demonstrate their support and interest in adopting, accepting and acting upon AI decisions.

4 DISCUSSION

Greater trust might seem to indicate a well-performing product for users. However, our findings indicate that users' experiences with alternative, human systems could easily confound measurements of trust. Is it possible that a high trust is simply an indication that users rely on AI as opposed to existing human institutions? For instance, an application might be dysfunctional, unsafe, or unfair to certain users, but they would still rely on it because AI is perceived as a better alternative. Overall, high perceived trust might not be an indicator about the system performance. Taking efforts that reduce acceptance of AI outcomes might seem antithetical to business goals: regardless, if trust is well-calibrated over time, then it might mitigate harm, retain users, and increases overall satisfaction with the system, leading to success. Overall, unwarranted trust might seem desirable at first, but it can negatively harm product experience in the long-term, as people continually receive outcomes that do not align with their expectations.

Trust and reliance should be aligned with the actual trustworthiness of AI [14, 41], instead of a trust built through proxy, confounding sources that we document in our findings. The gratitude shared by our participants was not isolated to the applications that they were currently using (social media, navigation, voice assistants), but often extended to all of AI systems, including those in our scenarios. Designers can consider introducing affordances to communicate the actual capabilities of the system while optimizing for understanding [2]. One can set the right expectations from early-use about the system's capabilities. In particular, designers can consider making the limitations of the system explicit before and during early use, by describing the ways in which a system operates and arrives at a decision, offering examples of situations in which the system is likely to provide unreliable results (see the PAIR Guidebook [30], and the Microsoft Human-AI Toolkit [1]). More research is needed to discover the nuances of the ways in which users might adjust these components of trustworthiness with continued interactions with a system. Future work can consider designing alternative metrics for measuring success, beyond user trust in a product or feature, that explicitly takes into account the contextual factors and their prior experiences.

4.1 Build competencies on AI use

Though the research on trust and fairness perceptions offers mixed results (often domain- and task-dependent), several prior studies find that respondents in the EU, US indicate a tendency to trust humans more than algorithmic systems [7, 10, 20], especially for tasks requiring human skills [21] (with some exceptions such as Logg [24] and Lee and Rich

[22]). In our technological context, we find a case of acceptance of AI outcomes with a confidence greater than a system might deserve. Unwarranted, over-trust is consequential: AI systems have potential to cause harm through incorrect or biased outcomes, especially because users might not actively seek out information about the capabilities of the given system. The effects of misplaced trust are exacerbated and far-reaching in high-stakes scenarios (hiring, finance, social benefits). Several participants considered a lack of human involvement in making the decision as desirable. Even the use of general-purpose products in critical situations (virtual assistants for job interview reminders) could lead to adverse outcomes for users.

Mayer, Davis, and Schoorman [27] present an integrative model of trustworthiness with three components: ability, integrity, and benevolence. We use this framework to reflect on our findings and present a path forward. Prior work (based in US or EU) reports that users rate technology companies (and products) with low trust (benevolence) [5, 28], but high ability. Therefore, when a technology system goes wrong, it is seen as a benevolence issue. Our results suggest that both—ability and benevolence—components of trustworthiness are high in India. Users are willing to give rely on AI because they have faith in the competence and benevolence of AI systems. When a system makes a mistake, neither its capabilities (ability) nor its intent (benevolence) were readily questioned. This resulted in self-blame or placing blame on other actors, without a recognition that AI could have made that error. Especially for the loan assessment scneario, many participants in our study believed that either they deserved an unfavorable outcome or could be their own fault. The self-blame was exacerbated when participants indicated a faith in the capabilities of AI. Low or non-contestation of AI outcomes has potential to cause individual harm if people accept decisions which might be wrong. In addition, users reporting decisions or behaviors to the system represents important opportunities for model feedback. This scope for mitigating harm and improving models is lost if users believe the outcomes to be correct, with a possibility of causing harm to other users.

There is a well-established body of work on explainable AI [4, 13, 37], and how explanations impact user trust and reliance [32, 40]. The focus of this research is to explore when to explain (high low stakes, [6]), what to explain (global local [11, 19]), and how to explain [44]. To make the system more transparent, there's increasing interest within the HCI, XAI communities to find human-centered approaches for explainability [23]. There are broadly three levels of building competencies on AI use: (1) outcome/in-the-moment explanation of a certain outcome, (2) an understanding of how a given AI application works, (3) general, accessible education about what AI is, how it works, and what are its strengths and limitations. Most explainability approaches have been context-agnostic, however, there needs to be particular attention to emerging internet users that often have less familiarity with technical jargon, often coupled with lower literacy [29]. Additionally, in-the-moment explanations are not always feasible due to legal or usability constraints. Consider a credit card approval application. Designers might be unable to 'explain' the reason behind a particular decision due to anticipated legal issues. Even without legal constraints, explanations are most useful when actionable [15]. A credit card rejection for an applicant with high trust on AI might mean that they believe that the system made the correct decision. As a result, they might not seek alternative platforms when it could easily be the case that the AI made an error or gave a biased outcome. The onus is on the designers/developers to add safeguards, ensure that users do not share a utopian view of the systems with which they interact [36], and acknowledge the range of system errors that are possible. Another approach to consider is to leverage existing capacities [43] by educating the user about how the system works in general ([12]) could be a valuable starting point wherein users can calibrate their responses based off their general understanding of how an application provided an outcome.

ACKNOWLEDGMENTS

To Robert, for the bagels and explaining CMYK and color spaces.

REFERENCES

[1] Microsoft Aether. 2021. Microsoft HAX Toolkit. https://www.microsoft.com/en-us/haxtoolkit/. (Accessed on 09/08/2021).

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[3] Christopher William Anderson. 2018. *Apostles of certainty: Data journalism and the politics of doubt*. Oxford University Press.

[4] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012* (2019).

[5] Brooke Auxier. 2020. How Americans view U.S. tech companies in 2020 | Pew Research Center. https://www.pewresearch.org/fact-tank/2020/10/27/how-americans-see-u-s-tech-companies-as-government-scrutiny-increases/. (Accessed on 09/06/2021).

[6] Andrea Bunt, Matthew Lount, and Catherine Lauzon. 2012. Are explanations always important? A study of deployed, low-cost intelligent interactive systems. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*. 169–178.

[7] Noah Castelo, Maarten W Bos, and Donald R Lehmann. 2019. Task-dependent algorithm aversion. *Journal of Marketing Research* 56, 5 (2019), 809–825.

[8] Jason A Colquitt. 2001. On the dimensionality of organizational justice: a construct validation of a measure. *Journal of applied psychology* 86, 3 (2001), 386.

[9] Aman Dalmia, Jerome White, Ankit Chaurasia, Vishal Agarwal, Rajesh Jain, Dhruvin Vora, Balasaheb Dhame, Raghu Dharmaraju, and Rahul Panicker. 2020. Pest Management In Cotton Farms: An AI-System Case Study from the Global South. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 3119–3127.

[10] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[11] Shi Feng and Jordan Boyd-Graber. 2019. What can ai do for me? evaluating machine learning interpretations in cooperative play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 229–239.

[12] Google. 2019. How Google Search Works (in 5 minutes) - YouTube. https://www.youtube.com/watch?v=0eKVizvYSUQ. (Accessed on 09/06/2021).

[13] Andreas Holzinger. 2018. Explainable ai (ex-ai). *Informatik-Spektrum* 41, 2 (2018), 138–143.

[14] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 624–635.

[15] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. 2019. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615* (2019).

[16] Sandhya Keelery. 2021. Internet usage in India - statistics & facts | Statista. https://www.statista.com/topics/2157/internet-usage-in-india/. (Accessed on 09/09/2021).

[17] Patrick Gage Kelley, Yongwei Yang, Courtney Heldreth, Christopher Moessner, Aaron Sedley, Andreas Kramm, David Newman, and Allison Woodruff. 2019. "Happy and Assured that life will be easy 10years from now.": Perceptions of Artificial Intelligence in 8 Countries. *arXiv preprint arXiv:2001.00081* (2019).

[18] Anastasia Kozyreva, Philipp Lorenz-Spreen, Ralph Hertwig, Stephan Lewandowsky, and Stefan M Herzog. 2021. Public attitudes towards algorithmic personalization and use of personal data online: evidence from Germany, Great Britain, and the United States. *Humanities and Social Sciences Communications* 8, 1 (2021), 1–11.

[19] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2019. Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 131–138.

[20] Markus Langer, Cornelius J König, and Maria Papathanasiou. 2019. Highly automated job interviews: Acceptance under the influence of stakes. *International Journal of Selection and Assessment* 27, 3 (2019), 217–234.

[21] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 2053951718756684.

[22] Min Kyung Lee and Katherine Rich. 2021. Who Is Included in Human Perceptions of AI?: Trust and Perceived Fairness around Healthcare AI and Cultural Mistrust. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[23] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[24] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103.

[25] Chiara Longoni, Andrea Bonezzi, and Carey K Morewedge. 2019. Resistance to medical artificial intelligence. *Journal of Consumer Research* 46, 4 (2019), 629–650.

[26] Vidushi Marda. 2018. Artificial intelligence policy in India: a framework for engaging the limits of data-driven decision-making. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, 2133 (2018), 20180087.

[27] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review* 20, 3 (1995), 709–734.

[28] Catherine Miller, Hannah Kitcher, Kapila Perera, and Alao Abiola. 2020. People, Power and Technology: The 2020 Digital Attitudes Report. https://doteveryone.org.uk/report/peoplepowertech2020/. (Accessed on 09/06/2021).

[29] Chinasa T Okolo, Srujana Kamath, Nicola Dell, and Aditya Vashistha. 2021. "It cannot do all of my work": Community Health Worker Perceptions of AI-Enabled Mobile Health Applications in Rural India. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–20.

[30] Google Research PAIR Team. 2021. People + AI Research - Guidebok. https://pair.withgoogle.com/guidebook/. (Accessed on 09/08/2021).

[31] Joyojeet Pal. 2012. The machine to aspire to: The computer in rural south India. *First Monday* (2012).

[32] Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert. 2019. How model accuracy and explanation fidelity influence user trust. *arXiv preprint arXiv:1907.12652* (2019).

[33] Sylvain Parasie and Eric Dagiral. 2013. Data-driven journalism and the public good:"Computer-assisted-reporters" and "programmer-journalists" in Chicago. *New media & society* 15, 6 (2013), 853–871.

[34] Munoz Claire Parry and Urvashi Aneja. 2020. 3.AI in Healthcare in India: Applications, Challenges and Risks | Chatham House – International Affairs Think Tank. https://www.chathamhouse.org/2020/07/artificial-intelligence-healthcare-insights-india-0/3-ai-healthcare-india-applications. (Accessed on 12/24/2021).

[35] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns* 2, 11 (2021), 100336.

[36] Edward Santow. 2020. Emerging from AI utopia.

[37] Tjeerd AJ Schoonderwoerd, Wiard Jorritsma, Mark A Neerincx, and Karel van den Bosch. 2021. Human-Centered XAI: Developing Design Patterns for Explanations of Clinical Decision Support Systems. *International Journal of Human-Computer Studies* (2021), 102684.

[38] Anubhutie Singh and Srikara Prasad. 2020. Dvara Research Blog | Artificial Intelligence in Digital Credit in India. https://www.dvara.com/blog/2020/04/13/artificial-intelligence-in-digital-credit-in-india/. (Accessed on 09/06/2021).

[39] Aaron Smith. 2018. Public Attitudes Toward Computer Algorithms | Pew Research Center. https://www.pewresearch.org/internet/2018/11/16/public-attitudes-toward-computer-algorithms/. (Accessed on 08/22/2021).

[40] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S Weld, and Leah Findlater. 2020. No explainability without accountability: An empirical study of explanations and feedback in interactive ml. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[41] Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad Van Moorsel. 2020. The relationship between trust in AI and trustworthy machine learning technologies. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 272–283.

[42] C Vijai and Worakamol Wisetsri. 2021. Rise of Artificial Intelligence in Healthcare Startups in India. *Advances In Management* 14, 1 (2021), 48–52.

[43] Marisol Wong-Villacres, Carl DiSalvo, Neha Kumar, and Betsy DiSalvo. 2020. Culture in Action: Unpacking Capacities to Inform Assets-Based Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[44] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. 2020. How do visual explanations foster end users' appropriate trust in machine learning?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 189–201.