Collaborations Between People and AI Systems (CPAIS)

Human - AI Collaboration

Key Insights from a Multidisciplinary Review of Trust Literature

September, 2019

PARTNERSHIP ON AI

Human - AI Collaboration Key Insights from a Multidisciplinary Review of Trust Literature

Introduction:

Understanding trust between humans and AI systems is integral to promoting the development and deployment of socially beneficial and responsible AI. Successfully doing so warrants multidisciplinary collaboration.

In order to better understand trust between humans and artificially intelligent systems, the Partnership on AI (PAI), supported by members of its Collaborations Between People and AI Systems (CPAIS) Expert Group, conducted an initial survey and analysis of the multidisciplinary literature on AI, humans, and trust. The results include this overview document presenting seven key insights which emerged from the literature review, as well as a bibliography of almost 80 research articles, tagged by theme.

The insights and bibliography can serve as a useful starting place for those working in or around AI, and can help align understandings related to trust between people and AI systems. This work can also inform future research, which should investigate gaps in the research and our bibliography to improve our understanding of how human-AI trust facilitates, or sometimes hinders, the responsible implementation and application of AI technologies.

Key Insights - Discussion:

Several high-level insights emerged when reflecting on the bibliography of submitted articles:

1. There is a presupposition that trust in AI is a good thing, with limited consideration of distrust's value.

The original project proposal emphasized a need to understand the literature on humans, AI, and trust in order to eventually determine appropriate levels of trust and distrust between AI and humans in different contexts. However, the articles included in the bibliography are largely framed with the need and motivation towards trust – not distrust – between AI systems and humans. While certain instances may warrant facilitated trust between humans and AI, others may actually enable more socially beneficial outcomes if they prompt distrust, or caution. Future literature should explore distrust as related, but not necessarily directly opposite, to the concept of trust. For example, an AI system that helps doctors detect cancer cells is only useful if the human doctor and patient trust that information. In contrast, individuals should remain skeptical of AI systems designed to induce trust for malevolent purposes, such as AI-generated malware that may use data to more realistically mimic the conversational style of a target's closest friends.

2. Many of the articles were published before the Internet's ubiquity/ the social implications of AI became a central research focus.

It is important to contextualize recent literature on intelligent systems and humans with literature focused on social and cognitive mechanisms undergirding human to human, or human to organizational, trust. Future work can put many of the foundational, conceptual articles that were written before the 21st century in conversation with those specifically focused on the context of AI systems, and their different use cases. It can also compare foundational, early articles' exploration of trust with how trust is seen specifically in relation to humans interacting with AI.

3. Trust between humans and AI is not monolithic: Context is vital.

Trust is not all or nothing. There often exist varying degrees of trust, and the level of trust sufficient to deploy AI in different contexts is therefore an important question for future exploration. There might also be several layers of trust to secure before someone might trust and perhaps ultimately use an AI tool. For example, one might trust the data upon which an intelligent system was trained, but not the organization using that data, or one might trust a recommender system or algorithm's ability to provide useful information, but not the specific platform upon which it is delivered. The implications of this multifaceted trust between human and AI systems, as well as its implications on adoption and use, should be explored in future research.

4. Promoting trust is often presented simplistically in the literature.

The majority of the literature appears to assert that not only are AI systems inherently deserving of trust, but also that people need guidance in order to trust the systems. The basic formula is that explanation will demonstrate trustworthiness, and once understood to be deserving of trust, people will use AI. Both of these conceptual leaps are contestable. While explaining the internal logic of AI systems does, in some instances, improve confidence for expert users, in general, providing simplified models of the internal workings of AI has not been shown to be helpful or to increase trust.

5. Articles make different assumptions about why trust matters.

Within our corpus, we found a range of implicit assumptions about why fostering and maintaining trust is important and valuable. The dominant stance is that trust is necessary to ensure that people will use AI. The link between trust and adoption is tenuous at best, as people often use technologies without trusting them. What is largely consistent across the corpus – with the exception of some papers concerned about the dangers of overtrust in AI – is the goal of fostering more trust in AI, or stated differently, that more trust is inherently better than less trust. This premise needs challenging. A more reasonable goal would be that people are able to make individual assessments about which AI they ought to trust and which they ought not trust, in the service of their goals for what specifically and in which circumstances. This connects to insight 1: There is a presupposition that trust in AI is a good thing. It is important to think about context, person-level motivations and preferences, as well as instances in which trust might not be a precondition for use or adoption.

6. AI definitions differ between publications.

The lack of consistent definitions for AI within our corpus makes it difficult to compare findings. Most articles do not present a formal definition of AI, as they are concerned with a particular intelligent system applied in a specific domain. The systems in question differ in significant ways, in terms of the types of users who may need to trust the system, the types of outputs that a person may need to trust, and the contexts in which the AI is operating (e.g., high- vs. low-stakes environments). It is likely that these entail different strategies as they relate to trust. There is a need to develop a framework for understanding how these different contributions relate to each other, potentially looking not at trust in AI, but at trust in different facets and applications of AI. For a more detailed analysis of what questions to ask to differentiate particular types of human-AI collaboration, see the PAI CPAIS Human-AI Collaboration Framework.

7. Institutional trust is underrepresented.

Institutional trust might be especially relevant in the context of AI, where there is often a competence or knowledge gap between everyday users and those developing the AI

technologies. Everyday users, lacking high levels of technical capital and knowledge, may find it difficult to make informed judgments of particular AI technologies; in the absence of this knowledge, they may rely on generalized feelings of institutional trust.

About the Project and Process

PAI's first in-person meeting of its CPAIS Expert Group took place in November, 2018 and helped to motivate this project. This expert group consists of about 30 representatives from across technology, academia, and civil society. Within these sectors, group members represent varied disciplinary training and roles (e.g., policy, research, product, psychology, computer science, and sociology).

The process of collecting articles from the Expert Group, via an open call for submissions, sourced content from a multidisciplinary community all aligned around an interest and expertise in human-AI collaboration. Submitted articles were evaluated for inclusion and analyzed by members of a smaller project group from within the PAI Partner community. Four thematic tags were developed, highlighting the ways the article abstracts approached the issue of trust. Specifically:

1. Understanding - lays out a conceptual framework for trust or is primarily a survey of trust-related issues.
2. Promoting - focuses on means for increasing trust
3. Receiving - focuses on the entity (e.g., a robot, a system, a website) that is trusted
4. Impacting - focuses on the nature of changes due to trust being present (e.g., the impact on a group or an organization when it experiences trust)

Two individuals from the smaller project group undertook a thematic tagging exercise to assess inter-rater reliability and the distribution of articles across tags. They tagged themes as primary and secondary (first and second order) for each article from the four thematic options above. There was 69% rater agreement when considering tags included in either first or second order thematic tagging, and 48% rater reliability for the total articles in which both reviewers tagged the same first-order theme.

The CPAIS Trust Literature Bibliography identifies thematic tags for each article, at levels 1 and 2. The "themes" column lists the first order themes and the second order themes, where applicable. The total tags for each article (at both levels) are also provided. "Understanding trust" was the most frequent theme - used with 61 articles (78% of the total). 50 articles (63%) were tagged with "promoting trust," and 29 articles (37%) were tagged with "receiving trust". Finally, 13 articles (16%) were tagged with a focus on impacting trust. .

This bibliography and thematic tags serve as fruitful entry points for those investigating the nuances in the literature on humans, trust, and AI, especially when contextualized with the insights drawn from the corpus presented above.

Acknowledgements

PAI is grateful to all the participants in the CPAIS Expert Group, and, in particular, to the following evaluators for their insights and contributions to this paper: Rachel Bellamy (IBM), Vicki Hanson (ACM), Rhia Jones (BBC), Hiroaki Kitano (SONY), Bran Knowles (Lancaster University, ACM), Frens Kroeger (Coventry University), John Richards (IBM), Jason Stanley (Element AI), and Amber Story (American Psychological Association).

About Partnership on AI

The Partnership on AI (PAI) is a global multistakeholder nonprofit committed to the creation and dissemination of best practices in artificial intelligence through the diversity of its Partners. By gathering the leading companies, organizations, and people differently affected by artificial intelligence, PAI establishes a common ground between entities which otherwise may not have cause to work together – and in so doing – serves as a uniting force for good in the AI ecosystem. Today, PAI convenes more than 90 partner organizations from around the world to realize the promise of artificial intelligence. Find more information about PAI at partnershiponai.org.

PARTNERSHIP ON AI