People and AI See Things Differently: Implications of Mismatched Perception on HCI for AI Systems

Saleema Amershi
Microsoft Research AI
samershi@microsoft.com

Ece Kamar
eckamar@microsoft.com
Microsoft Research AI

Emre Kıcıman
emrek@microsoft.com
Microsoft Research AI

ABSTRACT

People and AI are increasingly interacting and collaborating in the context of critical application domains (*e.g.*, healthcare, finance, transportation, and legal systems). There is often, however, a fundamental mismatch between how humans and machines perceive and reason about the world. This offers opportunities for bringing together multiple perspectives to reach better outcomes. On the other hand, this mismatch can hurt coordination and result in serious failures (*e.g.*, semi-autonomous vehicle accidents and misdiagnoses by clinical decision support systems). We believe a key solution is to ground communications between humans and machines in their common perceptions while allowing people to inspect and verify the AI and appropriately intervene when necessary. Achieving this requires the HCI and AI communities to address several challenges and to co-design HCI-AI patterns that enable verifiability, control, and consistency.

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

KEYWORDS

Mismatched perceptions, Implications, HCI-AI Design

ACM Reference Format:

Saleema Amershi, Ece Kamar, and Emre Kıcıman. 2018. People and AI See Things Differently: Implications of Mismatched Perception on HCI for AI Systems. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/1122445.1122456

HUMAN AND AI PERCEPTUAL DIFFERENCES

AI is widely employed in a variety of applications ranging from judiciary and health-care decision-making to semi-autonomous driving and support of human decision making. Most AI systems employ statistical learning algorithms to extract complex patterns from large sets of training data to make decisions given inputs, whether those inputs are from cameras, microphones, databases of medical records, or any other source. These statistical learning algorithms enable conversational and other natural interfaces; improved exploration and understanding of complex data-rich situations; and are employed in the real-world to assist people with decision-making situations. As AI capabilities continue to expand, the prominence of AI systems in our daily life and society will increase.

As promising and exciting as such AI-driven applications are, decisions of AI systems are not perfect; they may make serious or dangerous mistakes and in some cases their decisions may be biased against sub-populations of our society. A common strategy to mitigate machine errors is to have humans-in-the-loop, verifying and correcting machine decisions when necessary. When naïvely implemented, however, this simple error mitigation strategy often fails, either because the human user is unable to notice that the AI made a mistake, thinks the AI made a mistake when it did not, or did not have enough time to correct the error. These failures in human-AI collaborative systems go beyond the conventional Automation Irony.[1] *We believe a fundamental cause of such failures in AI-driven applications is the mismatch between how humans perceive and reason about the world, and how computers perceive and reason about the world.*

Machines are equipped with a variety of sensors that may surpass human perception capabilities, such as cameras that sense infrared and ultraviolet, and microphones that hear beyond the range of the human ear. Similarly, the cognitive architecture and models of AI systems are different than humans. Simply put, sometimes, AI-driven computers perceive things that humans do not. For example, recently many Amazon customers found that their Echo/Alexa devices were starting to laugh unprompted. Debugging this behavior, Amazon found that Alexa was hearing the command "Alexa, laugh" even in an apparently quiet room [5]. Relatedly, researchers have been able to command audio chatbots via ultrasound—even though the human ear does not hear the command, the chatbot does, and cannot tell that it is not really coming from a real user [12]. Similar issues arise in computer vision

[1]Automation irony refers to the fact that reducing the need for human intervention in a complex system often exacerbates the likelihood of human error [1]. As simple and frequent tasks are automated, the human is left to perform complex and rare tasks, but without the necessary intuition about system behavior that comes from repeated experience with simpler tasks.

applications, where minor visual changes can cause significant misinterpretations and errors by the computer, *e.g.*, a sticker on a sign causing the vision system in an autonomous car to miss the sign [2].

Of course, humans also perceive things that computers cannot. Machines learning automatically from data may end up with an incomplete representation of the world. For example, a dog classifier may learn to classify Husky dogs based on extraneous environmental factors such as snow on the ground, rather than facial features [7]. Or people may recognize an emergency situation from an audible alarm or sirens, whereas an automated assistant will have no awareness about emergency signals unless it is explicitly trained to have such awareness [10].

MISMATCHED PERCEPTIONS BREAK MENTAL MODELS

Perceptual differences can result in users formulating incorrect mental models of the AI systems they are interacting with, causing people to either expect too much (or too little) from these systems. For example, people riding in semi-autonomous cars may assume the cars can perceive all the information in their own visual fields even if the vehicle may only be perceiving a fraction of what the human notices. For example, many semi-autonomous cars currently optimize for detecting and monitoring moving objects while ignoring stationary ones [8]. This has led to serious crashes when stationary objects are within the path of the car (*e.g.*, a stopped firetruck on the road). In these cases, if users were made aware that stationary objects are often not detected, they may be better able to predict these failures in time to appropriately intervene and take control back from the car.

Incorrect mental models caused by perceptual differences can also lead users to under- or over-estimate how much trust to place in AI applications. Doctors believing AIs perceive more than they do may place too much trust in the recommendations from clinical decision support systems, even when they may disagree with the diagnosis or when the AI is in fact wrong [3]. Conversely, pilots not placing enough trust in AIs, even when those AIs have access to much more sensory information than humans and are able to efficiently communicate and coordinate with AIs in other planes, have resulted in sometimes deadly collisions.

MEDIATING MISMATCHED PERCEPTIONS THROUGH COMMON PERCEPTIONS AND VERIFICATION

While having the ability to perceive the world differently introduces challenges for human-machine coordination, it also offers potential benefits for collaborative machine and human execution. As machines and humans perceive the world differently and use different models for reasoning, it is unlikely that their individual error regions overlap. In these cases, the sum of the human and machine working together may surpass the performance of the AI system or human working alone, not unlike people with diverse perspectives coming together to make better decisions [9].

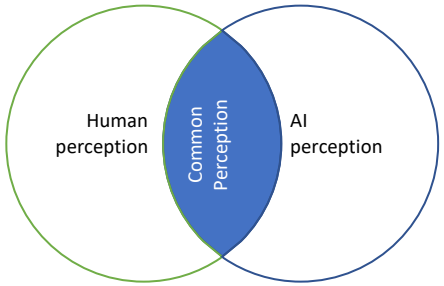

Figure 1: The common intersection of otherwise mismatched human and AI perceptions must serve as the basis for designing human-AI collaborative systems

A recent experiment performed on metastatic breast cancer diagnosis with simulated machine and human partnership provides support for this argument [11]. In this research, Wang et al. demonstrate that human diagnosis error rates can be reduced as much as 85% if the collaboration between the human expert and the AI system can accurately determine whose recommendation to follow at any time. While this example does not provide guidance on how to implement such an effective coordination, in a real-world implementation, the human expert can act as the mediator, deciding when to follow the machine decision and when to override it based on visual evidence provided by the AI system. In another example, some semi-autonomous vehicles now include displays that highlight detected objects to indicate what the vehicle is perceiving. This allows users to form more accurate mental models of the AI and enables earlier verification of AI behavior.

In the two examples above, a key factor in enabling humans to act as mediators is that the AI system's communications with a user are rooted in their common perceptions (Fig. 1). Regardless of what perceptive ability the AI is using to identify a tumor, it may use a heatmap overlaid on a slide image, allowing the human to inspect the slide to validate the AI's findings. Similarly, even if the semi-autonomous vehicle is using a sophisticated LIDAR sensor, it can highlight detected objects in the visual space that both it and its human driver understand. Improving the interpretability and explainability of AI systems will provide important tools to enable such experiences [4, 7].

Beyond grounding of interactions in the common perceptive capabilities of humans and AI systems, there are many additional challenges to verifiability:

- **Evidence:** The human should have sufficient evidence from the AI system or sufficient understanding about it to enable the verification. In some cases, understanding the situation and reviewing machine recommendation may be enough. In others, having techniques for questioning or explaining machine decisions may be necessary.
- **Time:** The human user must have sufficient time to evaluate and verify the AI solution. Reading of a medical image can allow sufficient time to do so, but this approach may not be ideal when a split-second decision must be made (*e.g.*, a human taking over from an autonomous car).
- **Trust:** Placing too much or too little trust in AI systems may interfere with verification. To achieve appropriate levels of trust, users may need to go through a training phase to formulate more accurate mental models about the AI and better understand their capabilities and limitations.
- **(In)consistency of AI models over time:** A challenge to validating and trusting an AI comes from surprising behaviors that might occur when an AI model is retrained [6]. When an AI model is updated, its behavior may change, sometimes significantly. Even when accuracy of the AI model is improved, if an AI that previously worked reliably in a given situation now begins to make errors (*i.e.*, the AI model makes fewer mistakes overall, but now examples correctly classified in the past are now identified incorrectly), the user may fail to correct its behavior

in time, leading to worse end-to-end outcomes. Even if retrained AIs are strictly better than previous models (*i.e.*, they make fewer mistakes and all examples correctly classified in the past are still identified correctly) users may have learned to distrust AIs in certain scenarios. The challenge then may lie in how to notify users of updates so they can anticipate changes and adjust their mental models accordingly.

- **Subjectivity:** In some domains, verification is inherently challenging. This is particularly true when decisions are subjective or correct decisions are not observable (*e.g.*, judging the likelihood of crimes to allocate police resources or predicting criminal recidivism in courts).

CALL TO HCI AND AI COMMUNITIES

Mismatched perceptions between human users and AI is a critical challenge in designing collaborative human-AI systems. At their extreme, such mismatches can result in dangerous and even fatal failures in human-AI collaborative systems. This shared perceptive capability must form the basis of communication and control between humans and AI systems. We propose the following two high-level HCI-AI Design Principles and a concrete research agenda for future progress:

Principle 1: The shared perceptive capabilities of humans and AI must ground the design of interactions between people and AI-driven systems.
Research agenda: Shared perceptive capabilities. We must characterize the effective shared perceptive capabilities of AI and humans, especially across different modalities (audio, image, video, AR/VAR, and text). We must investigate common, practical, and effective methods to map evidence and reasoning into this area of overlap. *E.g.*, when an AI uses extra-human-sensory perception to identify an object, how can that evidence be effectively communicated back to the human?

Principle 2: When differences in human and AI perceptions lead to differing conclusions, the system should be designed for easy human verification of AI outputs.
Research agenda: Human verification of AI outputs through shared perception. We must study trade-offs and strategies for using shared perceptive capabilities and develop robust metaphors for AI reasoning that allow humans to generalize across systems. The way we should communicate evidence and control between humans and AI will vary by application domain, as the need for summarization, deep understanding, quick action, and the cost of error will all vary. *E.g.*, in some cases, we may need to trade off high-fidelity evidence to enable people to make quick decisions. To aid people working across many systems, we should identify reusable metaphors that standardize such communications and trade-offs so that people can be trained to understand AI better.

Beyond this specific agenda, we believe the intersection of human and AI perceptive capabilities should also serve as the basis for reasoning about people's trust in and consistency of AI models over

time. For example, AI models may be retrained, and may act inconsistently over time with respect to their own unique perceptions, but to ensure trust, must appear to be acting consistently with respect to how humans perceive situations. That is, non-audible sounds, radar, etc., may be more or less important in an AI's functioning as it is retrained, as long as the AI appears to be functioning consistently from the user's perspective.

There are many open and fundamental questions around how to best achieve these two principles: Is it imperative that the intersection of human and AI perceptive capabilities grow over time to encompass substantially all of human perceptive capabilities? In critical situations, how can we ensure that human users have the time necessary to verify AI outputs? How can we exploit the common perceptive capabilities of humans and AI to ensure AI results are more explainable, and allow users to explore and validate reasoning and potentially take action to fix or override the AI's behavior? How do we teach users about the limits of specific AI models, especially as the models change over time? How do we build AI models that improve while appearing to behave consistently to human users? When AI models do change inconsistently, how do we ensure users are aware and can adapt?

REFERENCES

[1] Lisanne Bainbridge. 1983. Ironies of automation. In *Analysis, Design and Evaluation of Man–Machine Systems 1982*. Elsevier, 129–135.

[2] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. 2017. Adversarial patch. *arXiv preprint arXiv:1712.09665* (2017).

[3] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *Healthcare Informatics (ICHI), 2015 International Conference on*. IEEE, 160–169.

[4] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[5] Mashable. 2018. Amazon reveals why Alexa is randomly laughing and creeping people out. https://mashable.com/2018/03/07/why-amazon-alexa-laughing/

[6] Gagan Bansal Besmira Nushi, Ece Kamar, Daniel S Weld Walter S Lasecki, and Eric Horvitz. 2019. Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. *AAAI* (2019).

[7] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of ACM KDD 2016*. ACM, 1135–1144.

[8] J. Stewart. 2018. Why Tesla's Autopilot Can't See a Stopped Firetruck. https://www.wired.com/story/tesla-autopilot-why-crash-radar/

[9] James Surowiecki. 2005. *The wisdom of crowds*. Anchor.

[10] Waymo Team. 2017. Recognizing the sights and sounds of emergency vehicles. https://medium.com/waymo/recognizing-the-sights-and-sounds-of-emergency-vehicles-8161e90d137e

[11] Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H Beck. 2016. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718* (2016).

[12] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. 2017. DolphinAttack: Inaudible voice commands. In *Proceedings of ACM SIGSAC 2017*. ACM, 103–117.