Anja Thieme, Microsoft Research
Ed Cutrell, Microsoft Research
Cecily Morrison, Microsoft Research
Alex Taylor, City, University of London
Abigail Sellen, Microsoft Research

Interpretability as a Dynamic of Human-AI Interaction

Insights

We should design AI systems as a resource for people to extend their own capabilities. A few key guidelines:

- Consider interpretability as evolving through dynamic, situated system interactions.
- Provide users tools to work with the system to improve its perception accuracy.
- Create easy access to differently derived AI system outputs to enable users to determine an appropriate level of trust in the system.

Rapid development in AI technologies such as computer vision have enabled the robust perception of many aspects of our world, inviting tantalizing new experience designs. Yet these technology advances also raise many fundamental design considerations, as they become embedded in real-world applications. Designers must think carefully upon the dynamic between the person and the AI system they aim to create, and how that is realized through the design of an interpretable system. In this article, we reflect upon interpretability as a dynamic of human-AI interaction through sharing our design journey of developing an AI-enabled experience that provides ongoing information about people in the immediate vicinity for people who are blind or have low vision [1].

Through the narrative describing our design journey, we capture key insights on two themes. We begin with a discussion, central to any conversation about human-AI interaction, of how we positioned the AI system in relationship to the person using it. Here we illustrate how AI can become a resource for enabling people to extend their capabilities, standing in contrast to the automation of experiences, or systems that attempt to emulate human ability. With the human-AI partnership

articulated, we then consider how it is supported through designing for interpretability. We focus on how this can be achieved by taking advantage of continuous interactions with a dynamic AI system. We then pull these insights together into a set of guidelines for designing future AI-enabled experiences.

POSITION THE AI AS A RESOURCE FOR ENABLING HUMAN CAPABILITIES

Much AI application development has focused on enabling standalone systems like self-driving cars or analytics tools that aid company decision making, (e.g., choosing who qualifies for a loan). In these examples, AI systems are positioned as either capable of *emulating humans* (e.g., driving) or *superior to humans*, potentially outperforming them through improved data insights or productivity. We propose a different orientation to AI systems. Moving beyond the emulation and replacement of human activities, we suggest that AI systems can serve as a useful resource *for* humans, helping them expand their agency to develop new or extend existing skills. This requires consideration of the dynamic of the partnership between person and AI system.

We explored the nuance of human-AI partnerships through a series of design research [2] and ethnographic fieldwork [3] activities with people who have vision impairments. This work highlighted that social relationships and interactions are critical for how our participants made sense of their surroundings, connected with others, or sought help. Based on this insight, we began to imagine how perception technology could provide functionality that would offer people with vision impairments *dynamic, in situ access to information about others nearby*. Such information could make it easier, for example, for a blind person to proactively approach someone to socialize rather than waiting for someone to reach out. It might also mitigate the embarrassing situation of starting a conversation with someone who had quietly left the room. The design aim was to build a system that would enable people with vision impairments to be more confident in how they approached social situations, rather than trying to design a system to replace their vision.

The AI system and enabled user experiences. In conjunction with a user team of eight blind or low-vision people, we developed an AI system (Figure 1) that runs on a head-worn HoloLens device modified to remove the lenses. The device captures a near 180-degree field of view surrounding the person wearing the device, tracks their head position, and provides high-quality spatialized audio from non-occluding speakers above the ears. Multiple state-of-the art computer vision algorithms process the captured images to continuously identify other people nearby, including their identity, location, activity, and gaze direction. The outputs of the underlying perception models are further integrated into a real-time tracking model of all people detected.

Users can filter and receive information about people in the vicinity acoustically via spatialized audio using various input controls on a wrist-worn controller (Figure 2). We created three core experience modes for users to interact with. When users access the "overview" mode, the system instantly reads out the total number of people that it presently detects (e.g., "three people"). In this mode, through twist-by-twist interactions, users can receive additional details for each detected person: their name (or the system states "unknown"), approximate location, and time passed since the person was last detected (e.g., "John, near-front, 10 seconds ago"). The user can either dwell to receive all those details, or quickly skip over these through further twists. This information access can make it easier for the user to build up an understanding of who and where certain people might be in a room, and facilitate their approaching. For example, "John" might be a friend the user would like to talk to. Once in conversation with "John," the user may activate the "person in front" mode. In this mode, whenever the user looks directly at another person (indicated through the device orientation), they would hear the name of the person, or, if not identified, a spatialized sound to indicate the presence of a person. This functionality can help confirm who a conversation partner may be and enable the user to better adjust their body orientation toward that person. Finally, the "ambient" mode provides a sound for each person nearby at regular intervals (e.g., every 30 seconds) without any names or details. These interval-based updates allow users to have a more continuous sense of people's presence through peripheral, low-frequency audio that they can easily tune into or ignore if irrelevant.

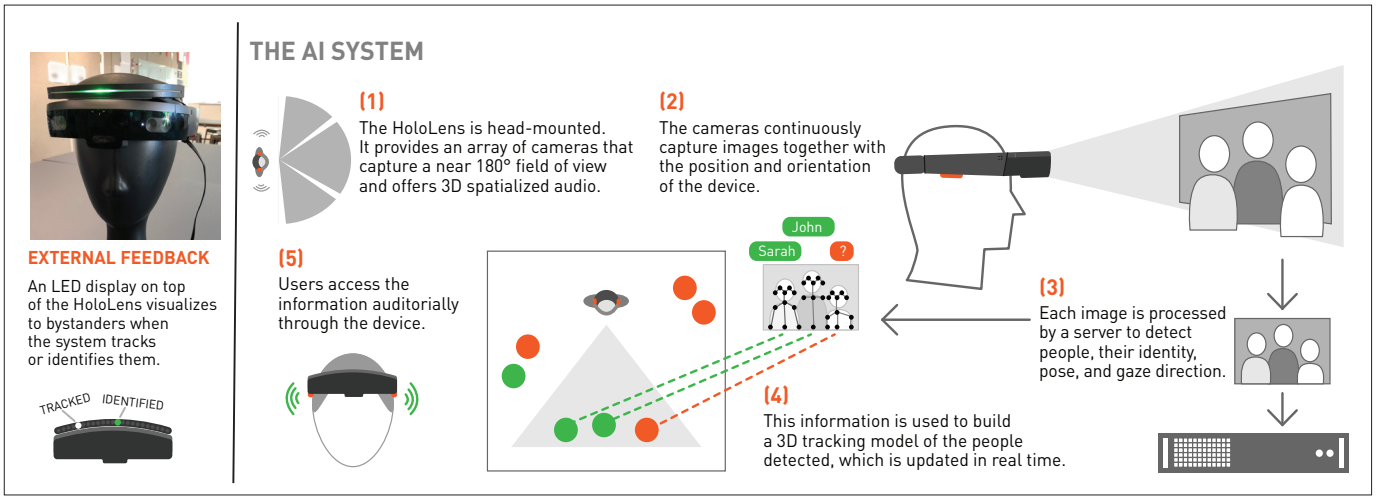


Figure 1. Left: Image of the adapted HoloLens device. Right: Schematic description of the core functionality of the AI system.

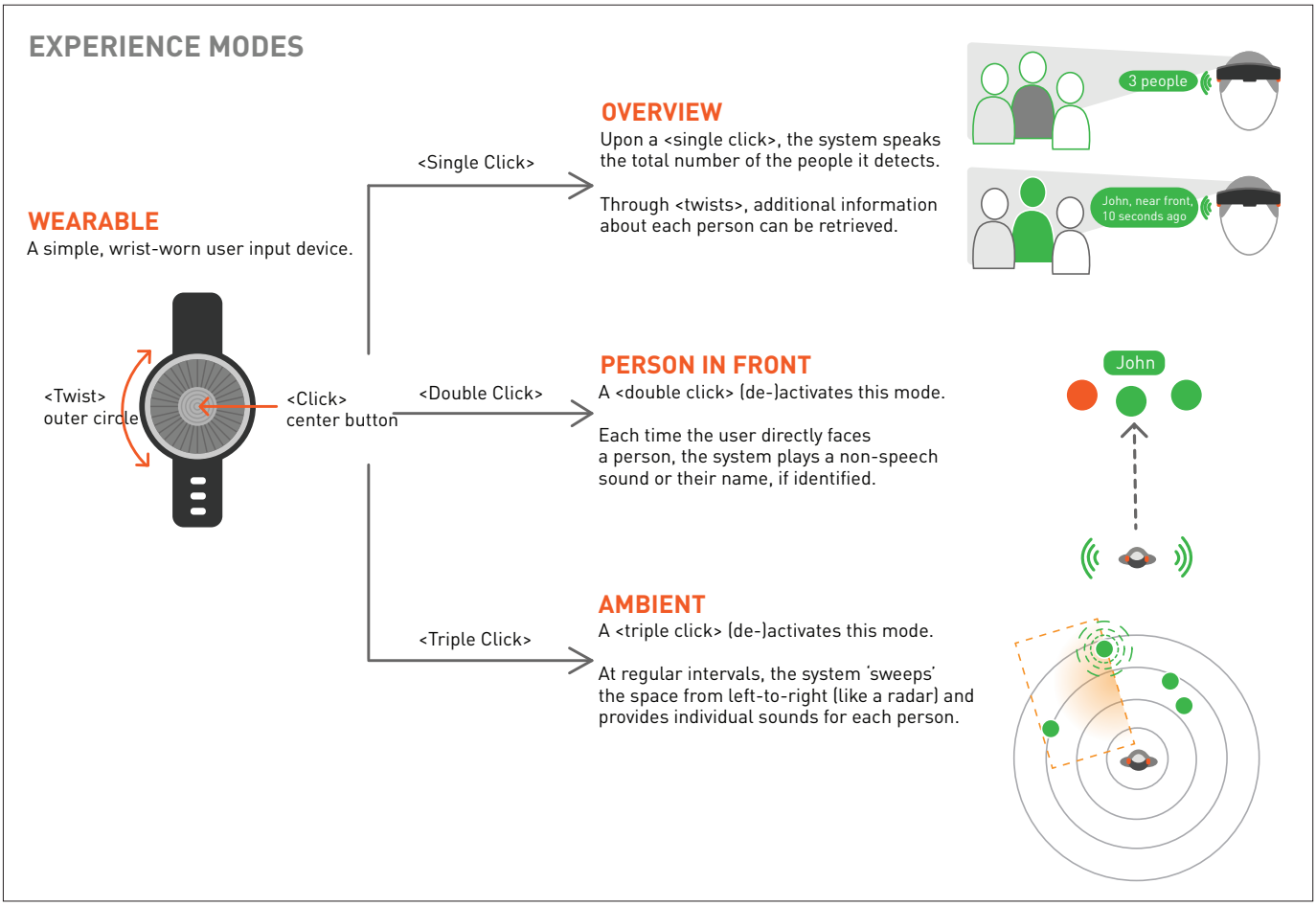


Figure 2. Description of the input modalities of the wrist-worn controller and the outputs of each of the three experience modes.

Foreground human agency and understanding. The way information is communicated by the system reflects a number of design decisions explicitly intended to reflect a human-AI partnership that places agency with the person rather than the system. First, the experience is user controlled, allowing the user to determine when and what kind of information would suit them in a given situation. This stands in contrast to a technical trend toward integrated AI systems (also called end-to-end systems). Such an approach provides complex inferences about a situation, for example, determining when to provide social information. Engagements with our user team illustrated that such inferences are very specific to personal preference, existing skills, and a complex set of contextual cues. Attempting to infer these is likely to lead to a fallible system that loses users' trust. In contrast, the design approach described here forefronts the user as "doing the understanding," helping them maintain agency of their own lived experience.

Second, the information provided, presented through a range of experience modes, is simpler than what the system is technically capable of. Participants were best able to incorporate information into their experience that was immediately useful to, but not disruptive of, their interactions with those around them. These simple pieces of information, such as "who is in the room," support the user to initiate interaction, such as approaching a particular person. This type of experience clearly positions the person as the entity with agency and the AI as a resource. The result is an experience that is less prescriptive, leaving space for users to identify meaningful appropriations of this AI resource within their lives beyond anticipated use cases.

In order to achieve this design vision of an AI system as an information resource for people that can help extend their capabilities, it is further essential to support an appropriate user understanding of the system functionality. Next, we consider how such *interpretability* can be achieved through continuous interactions with a dynamic AI system.

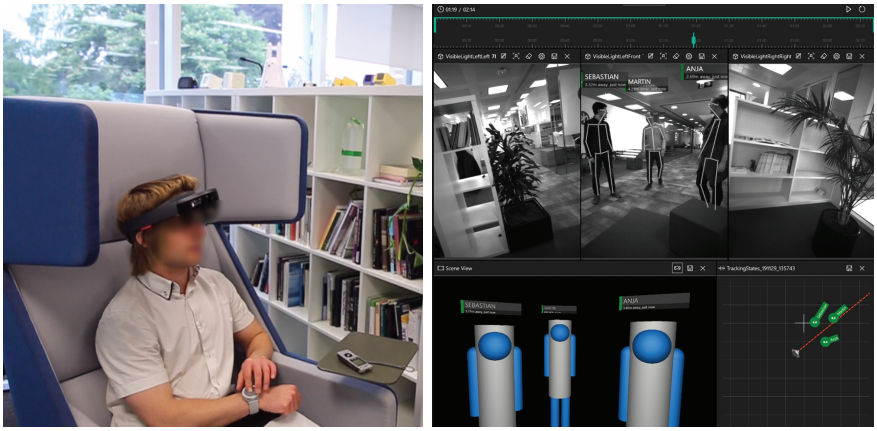
INTERPRETABILITY THROUGH CONTINUOUS INTERACTIONS WITH A DYNAMIC AI SYSTEM

The real-time human tracker of our AI system is based on a combination of multiple technically sophisticated machine learning (ML) models, each of which is trained on diverse datasets and based on many model parameters. The complexity of this integrated model can make it difficult, if not impossible, for humans to understand *how* certain predictions are achieved, or *why* a model may perform more or less accurately in different contexts. Furthermore, as a system that is continuously worn by the user, there are many contextual factors that can complicate the reliable detection of other people nearby. For example, different light conditions or the occlusion of a face can hinder the system's ability to recognize people's identity—above and beyond any probabilistic uncertainties of its algorithmic outputs [4]. Considering that these limitations are difficult to overcome fully, how can we provide

users tools that 1) enable them to *work with the AI system* to help improve its detection performance, and thereby, practical usefulness; and 2) help them *develop an appropriate mental model of an AI system's state or behavior* so that they can have a clearer understanding of whether they can rely on, or need to further verify, system-derived information.

Work with the AI to improve its performance and usefulness. Among the experience modes we created (Figure 2), the "person in front" functionality was popular. In this mode, whenever the user of the AI system was looking directly at another person, they would hear that person's name, or a spatialized sound if the system detected a person but did not identify them. When the system had a robust internal representation of nearby people, this audio feedback was instant and presented as a short, comprehensive information cue that was easy to place in the environment due to its positioning in relation to the user's head orientation and their control of head movement. As such, it served as a useful resource to quickly build up an understanding of the surrounding social landscape.

In building up the system's representation of people, different computer vision models were employed. While the recognition of people's bodies via pose detection was fairly robust, even when parts of the body were occluded, identity recognition required clear images of full faces [4]. As it can be difficult for those with vision impairments to know how to best direct their head to help frame other people's faces such that the system can perceive them, we developed *orientation cues*. For those people who are detected by the system but not identified by name, an additional *woodblock* sound is played. This sound changes pitch if the user's head is tilted too high or too low, and "snaps" to faces, helping the user orient to the nearest face.



User interacts with overview and twist inputs to build up his understanding of others nearby:

User: [Twists] "Martin, near-front, 10 seconds ago." I'm pretty sure he's not there anymore.
[Twist] "Anja, near-front, just now"
[Twist] "Sebastian, near-front, just now." Yeah, that's all good.
[Keeps twisting] "Unknown, really far away, 2 seconds ago."
[Twist] "Martin, near-front, 10 seconds ago."

Anja: Did he just creep up?

User: Maybe, yeah. I think maybe the system put him down as unknown to begin with. "Sebastian, still there."
[Twist] "Unknown, near-front, just now."
[Twist] "Martin, near-front, just now."
[Clicks overview] "3 people." OK, so the unknown was Martin. OK, that's cool.

Figure 3. Top left: Using the wrist controller, one of our blind participants switched and triangulated across different system outputs to build up his understanding of others nearby, which he articulates out loud to the researchers around him (text below). Top right: AI system view of the surroundings, showing the 180-degree field of view of the HoloLens cameras as well as the real-time model of three people that it detected: their location, identity, and orientation to the user.

Seamlessly integrated within AI system interactions, the different types of audio feedback provide insights about the system state in recognizing people while giving users the means, through body-orientation adjustments, to work with the system to improve its recognition performance.

Furthermore, we found that interactions extended beyond the user and the AI system to a three-way reciprocal relationship between the user, the system, and the people with whom the user is interacting. Designed as a social system that detects others nearby, it was important to consider how those bystanders would come to understand the purpose and functionality of the visible head-mounted device, as well as become an active participant in the social sensemaking in which the user was engaged.

We affixed a semicircular LED interface to the top of the HoloLens (Figure 1, left) that communicates the system state visually to bystanders. A moving white light tracks the location of the nearest detected person and flashes green when that person is identified (auditorily) to the user. The visual feedback enables the development of common ground between all parties and enables bystanders to test out the workings of the system. Bystanders can use that understanding to physically orient themselves to the system—to make themselves more detectable to the system or, conversely, to evade it if they do not want to be captured. Creating *transparency through a more open sharing of the system state* can further help manage bystander expectations and ameliorate concerns that they otherwise might have about a system that would try to detect them "unobtrusively."

Allow users to triangulate different system information. Each of the experience modes that we created is

We found that interactions extended beyond the user and the AI system to a three-way reciprocal relationship between the user, the system, and the people with whom the user is interacting.

enabled by a different set of computer perception models that capture different types of information. For example, while an initial overview (outputting the number of people detected) relies only on a robust pose detection of people's bodies, the identification of a person's identity in the "person in front" mode requires input from multiple perception models, starting with *pose*, then *face*, and finally, *identity* recognition. In providing users access to these differently derived types of system information via the experience modes and a simple controller to switch between these modes, we created interactions that would enable users to more easily identify consistencies and ambiguities across system outputs. This can aid confidence in the information offered or invite more caution in the interpretation of system feedback and further inspection.

Figure 3 exemplifies this interaction. Here, one of our blind participants was using the AI system to better understand who was present. Using the wrist controller, he kept triangulating between system outputs. By identifying discrepancies between unknown and identified people as well as temporal information (the time passed since a person was last detected), he was able to build up a more accurate understanding of system ambiguities. Having access to differently derived types of information, combined with the user's own common-sense understanding (e.g., that it is not possible for people to suddenly disappear or be in multiple places at once), our participant was able to better interpret the configuration of people around him. This suggests that, rather than developing highly complex, integrated perception models with multiple interdependencies that are difficult to disentangle, there might be value in deliberately enabling access to decomposed model outputs.

Facilitate system use in the context of people's other ways of knowing. As illustrated through our example in Figure 3, a user's understanding of their social surrounding is not solely informed by, or reliant on, system feedback. Humans are not "turned-off receivers"; they bring other existing senses and ways of understanding the world around them to their interactions with technology. By embedding uses of the AI system within people's everyday lives and situated interactions with others, we enable them to evaluate the system outputs in the context of their "other ways of knowing" about a social situation. This allows users to better scrutinize the validity of system information. For example, the user may be actively holding a conversation with someone whom they know well, or have clear expectations of who will attend a particular meeting. These other ways of knowing who is likely to be present are therefore instrumental in helping users more carefully interpret system outputs, and through this, to confirm or reject some of the assumptions that they may hold about their social surroundings. This mediates how much users would trust, or be cautious about relying on, system-derived information.

CONCLUSION

Positioning the AI system as a resource for people and foregrounding human agency and sensemaking, we created a set of experiences with a dynamic perception system that enables people with vision impairments to build up a richer understanding of their social surroundings. This can extend their capabilities by creating more opportunities for them to socially connect and be more confident in their interactions with others nearby. In this context, interactive features such as orientation cues, an external display of the system state, and access to different experience modes can support the formation of collaborative partnership(s) between human(s) and the AI system. It is through the dynamic back-and-forth interactions between system feedback and other human sensemaking capabilities, situated within people's everyday lives, that users can develop a better understanding of the AI system's functionality and derive practical and meaningful uses from it, despite some ambiguity in its outputs.

ENDNOTES

1. Grayson, M., Thieme, A., Marques, R., Massiceti, D., Cutrell, E., and Morrison, C. A dynamic AI system for extending the capabilities of blind people. *CHI EA'2020*. ACM, 2020, 1–4; https://doi.org/10.1145/3334480.3383142
2. Morrison, C., Cutrell, E., Dhareshwar, A., Doherty, K., Thieme, A., and Taylor, A. Imagining artificial intelligence applications with people with visual disabilities using tactile ideation. *Proc. ASSETS 2017*. ACM, 2017, 81–90; https://doi.org/10.1145/3132525.3132530
3. Thieme, A., Bennett, C.L., Morrison, C., Cutrell, E., and Taylor, A.S. 'I can do everything but see!'—How people with vision impairments negotiate their abilities in social contexts. *Proc. CHI 2018*. ACM, 2018, Paper 203; https://doi.org/10.1145/3173574.3173777
4. Stearns, L. and Thieme, A. Automated person detection in dynamic scenes to assist people with vision impairments: An initial investigation. *Proc. ASSETS 2018*. ACM, 2018, 391–394; https://doi.org/10.1145/3234695.3241017

Anja Thieme is a senior researcher in the Healthcare Intelligence group at Microsoft Research Cambridge. She takes a human-centered approach to the study of machine learning applications for health and accessibility, aiming to design responsible AI systems that provide useful, interpretable, and actionable insights that can positively transform people's lives.
→ anthie@microsoft.com

Edward Cutrell is a senior principal researcher at Microsoft Research, where he explores computing for disability, accessibility, and inclusive design in the MSR Ability group. He has worked in the field of HCI since 2000, on topics ranging from intelligent notifications and disruptions to technology for global development (ICTD).
→ cutrell@microsoft.com

Cecily Morrison is a principal researcher in the Future of Work group at Microsoft Research Cambridge. Her research lies at the intersection of human-computer interaction and artificial intelligence. Working in a cross-disciplinary collaboration, she is currently focused on AI applications for those with visual disabilities.
→ cecilym@microsoft.com

Alex Taylor is a sociologist working in the Centre for Human Computer Interaction Design at City, University of London. Showing a broad fascination with the entanglements between social life and machines, his research ranges from empirical studies of technology in everyday life to speculative design interventions—both large and small scale.
→ alex.taylor@city.ac.uk

Abigail Sellen is deputy director at Microsoft Research Cambridge. She has published on many different topics in HCI, always seeking to put human aspiration front and center in designing new technology. A recent focus is on the intelligibility of AI systems, viewed through the lens of both HCI and philosophy.
→ asellen@microsoft.com

DOI: 10.1145/3411286