Understanding Our Robots With the Help of Human-Centered Explainable AI

Insights from the field of human factors can help us design human-centered explanations that enable effective human-robot interaction. Studying explanation techniques according to these human factors will be critical in understanding their efficacy across diverse contexts.

By Lindsay Sanneman

DOI: 10.1145/3611686

OPEN ACCESS

Over the last decade, we have seen incredible growth in the number of robots operating throughout workplaces, communities, and homes worldwide. Sales of industrial robots for applications, like manufacturing, have doubled over the last five years [1]; the first autonomous vehicles have begun to drive alongside human drivers on roads in a growing number of U.S. cities; and the number of service robots purchased for use in industries from hospitality to medical applications grew by 37% in 2021 alone [2]. These robots promise to make our work less mundane and more rewarding, our roads safer, and our homes and communities easier and more efficient to manage and navigate—and the trend of growth we have seen in recent years shows no sign of slowing down. But there are novel risks and challenges associated with the deployment of these systems. One such challenge is ensuring the robot's behavior is sufficiently transparent to the humans who interact with these robots to enable both useful and productive interactions. Without this transparency, at best, humans may not know how to effectively use or interact with robots, rendering them useless to the humans they aim to serve. At worst, humans' lack of understanding of how these robots will behave across different scenarios could pose a threat to human safety. To avoid these potential pitfalls, ensuring transparency of robot behavior will be critical to the safe and successful integration of robots into our communities as the number of applications continue to grow in the coming years.

Explainable artificial intelligence (XAI) poses one potential solution to the transparency challenge, both for

Image by HGT Graphics / Shutterstock.com

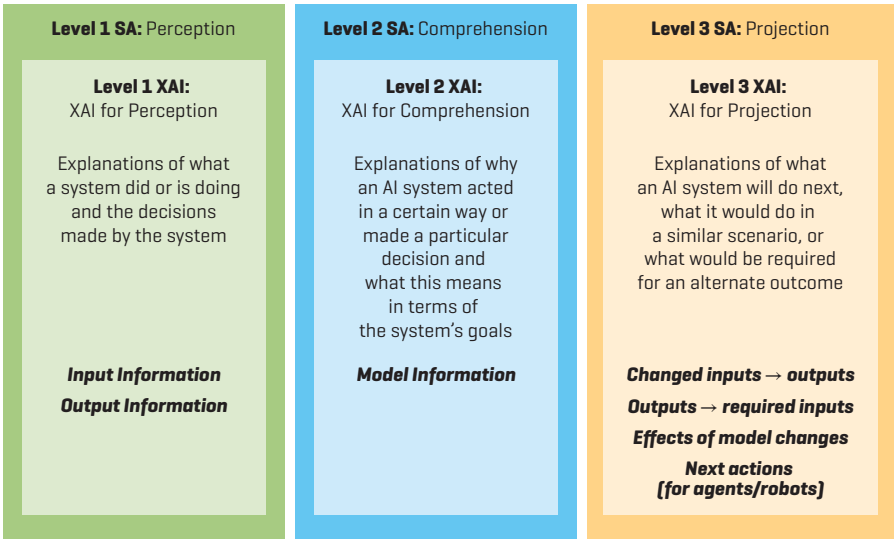
AI and for robotics. The recent wave of XAI research began to emerge around 2016 when the Defense Advanced Research Projects Agency (DARPA) launched its Explainable AI program in response to advances in machine learning that were picking up around that time [3]. The program highlighted the need for interpretable and transparent AI systems in response to the increasingly opaque and inscrutable, but powerful, approaches for machine learning that were being developed at the time. They identified XAI, which they defined as "AI systems that can explain their rationale to a human user, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future," as a potential solution.

However, even with the resurgence of research on XAI that accompanied the introduction of DARPA's XAI program, many open questions about how XAI systems should be designed remained. Many of the most important unanswered questions at the time were related to the humans who would be receiving the explanations across a variety of domains and contexts. For example, which specific information does a person need? How does this depend on their tasks, role, or context? How much information is too much to provide? How do explanations impact a person's trust in automated systems, and what does this mean in terms of how they will use or rely on such systems?

In 2017, as many new papers on XAI were being published in leading AI conferences, Tim Miller, one of the leading experts in the field, noticed these human-centric questions were not being addressed by most researchers in the field. In a paper titled "Explainable AI: Beware of the Inmates Running the Asylum," he stated, "while the re-emergence of explainable AI is positive [...] most of us as AI researchers are building explanatory agents for ourselves, rather than for the intended users" [4]. In other words, the XAI research community was developing techniques that explained information considered useful for programmers of AI systems but not necessarily for the lay users who would most often interact with these systems in their deployment.

Figure 1. The Situation Awareness Framework for Explainable AI (SAFE-AI).

Included in the figure are examples of the types of information that might be provided through techniques for XAI at each of the three levels.



So how can we develop XAI for everybody, including programmers, researchers, lay users, and even bystanders who might not directly interact with these systems but are still impacted by their deployment? This question becomes even more critical to address when considering embodied AI systems such as robots that act in the real world and have implications for human safety where they are deployed.

The field of human factors has long addressed human-centric questions like these in the context of complex human–autonomy interaction domains such as pilot interactions with cockpit automation systems. Human factors studies the application of psychological and physiological principles to the engineering and design of products, processes, and systems in order to enhance human-system performance and thus addresses many of the questions that must be answered within the XAI space. We can therefore draw lessons from human factors to enable a human-centric approach to the design and evaluation of XAI systems and explainable robots.

The widely-studied human factors concepts of situation awareness, cognitive workload, and trust can provide particularly valuable insights for the development of XAI systems [5]. Next, we delve into how each of these concepts can be leveraged to inform the design and evaluation of XAI, including for explainable robots.

SUPPORTING HUMAN SITUATION AWARENESS THROUGH EXPLAINABILITY

The human factors concept of situation awareness (SA) relates to a human's awareness of their environment as it relates to the tasks they must perform. It therefore dictates informational needs for humans performing any role in any scenario. The most common definition for SA from human factors literature includes three levels: "the perception of elements in the environment within a volume of time and space (level 1), the comprehension of their meaning (level 2), and the projection of their status in the near future (level 3)" [6]. According to human factors literature, a human must have relevant information at all three levels in order to perform their tasks successfully, and in fact, SA has been shown to correlate with human-autonomy team performance in domains as diverse as autonomous driving and search-and-rescue missions.

AI designers must ask which information is absolutely necessary to explain and what a person's overall workload will look like over the course of an interaction...

Critically, when there is an AI system or a robot that is operating within a particular environment where a human is performing tasks, part of the human's SA includes information about this autonomous agent. XAI systems, as systems that provide information about AI behavior, can contribute to the subset of a human user's SA that is related to AI behavior in particular. The Situation Awareness Framework for Explainable AI (SAFE-AI) formalizes the relationship between SA and XAI, and just as SA is divided into three levels (see Figure 1). SAFE-AI includes three levels of XAI that map closely to the levels of SA. These include the following:

1. **Level 1. XAI for perception**—explanations of what an AI system did or is doing, and the decisions made by the system.
2. **Level 2. XAI for comprehension**—explanations of why an AI system acted in a certain way or made a particular decision and what this means in terms of the system's goals.
3. **Level 3. XAI for projection**—explanations of what an AI system will do next, what it would do in a similar scenario, or what would be required for an alternate outcome.

These levels define the subset of a person's overall SA, which relates to AI behavior in particular. As with SA overall, information at all three levels is necessary to support individuals who are performing goal-oriented tasks in human-AI or human-robot teams. For a given human user and context, required information at each of the three levels can be comprehensively enumerated, and these informational needs can then be matched with XAI techniques that can be applied to meet them. In cases where there are no existing techniques that can meet a particular requirement, research must be performed to develop a new technique or set of techniques.

Importantly, the SAFE-AI framework enables the definition of informational needs for humans playing different roles, spanning the spectrum of robot programmers to bystanders. Consider an industrial robotics setting where a robot is integrated onto a factory floor to assist workers with manufacturing and logistics tasks such as palletizing boxes. Here, we can consider three distinct human roles: a robot programmer, a factory worker who interacts directly with the robot to prepare pallets for shipping, and a factory worker who does not work directly with the robot but still navigates within the robot's workspace (i.e., a bystander). Figure 2 includes examples of information that an XAI system, which explains the robot's behavior, must provide to each of these humans according to the three levels of the SAFE-AI framework.

ACCOUNTING FOR HUMAN WORKLOAD

While SA defines which information an XAI system should provide to users about an AI system, the concept of mental workload dictates how and when this information can be delivered. Mental workload can be defined as the relationship between the mental resources demanded by a task and those resources available to be supplied by the human. When an explanation is provided to a human, that human must have sufficient mental resources to process it. Therefore, workload considerations inform both the ideal frequency of information sharing and the ideal amount of information to provide through a single explanation within a given context.

In addressing considerations related to mental workload, XAI practitioners can leverage the multiple resource model (MRM) from human factors literature [7]. The MRM defines different "pools" of cognitive resources that people have available for information processing over dimensions including the modality of information representation, the form of information encoding in the brain, stages of information processing, and response modality. Explanations must be designed such that no individual pool is overwhelmed when these explanations are provided, either due to a single explanation on its own (which we can call a "local" workload consideration) or due to the frequency and content of explanations in the context of the other tasks a user is performing simultaneously (which we can call a "global" workload consideration).

Insights from human factors can help us design more transparent autonomous agents and robots through the application of human-centered explainable AI.

For example, the worker who directly collaborates with the robot in the previously-introduced manufacturing example observes the robot operating in the environment (visual information modality), reasons about how this robot will move within the workspace (spatial information coding), decides how to act in response to the robot's motions (spatial central information processing stage), and takes these actions in the workspace that is shared between the human and the robot (manual response modality). Any explanations that are provided to the worker within this context must account for the total workload they amass across various modalities in their perception, reasoning, and response processes.

CALIBRATING HUMAN TRUST IN ROBOTS THROUGH EXPLANATIONS

Human trust in automation also has important implications for the development of XAI systems. A commonly applied definition of "trust in automation" is the attitude that an autonomous agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability [8]. In XAI literature, increasing user trust in automation is often suggested as a motivation for providing explanations to humans. However, human factors literature has instead focused on appropriately calibrating trust in autonomous systems in order to ensure appropriate use and reliance on such systems. A person who over-trusts an autonomous system might over-rely on it or under-monitor it, which could compromise safety, especially in human-robot interaction scenarios. On the other hand, a person who under-trusts an autonomous system might not use it at all or might over-monitor it, thereby diverting the person's attention from other tasks they are performing, which

Figure 2. Example of SAFE-AI applied to a manufacturing task.

Here, the robot picks up boxes from the conveyor belt and places them on pallets that the collaborating worker wraps for shipment. The table above shows examples of information each person might need in this scenario, which could be provided by XAI techniques.

	Role: **Robot Programmer**	Role: **Worker (Robot Collaborator)**	Role: **Worker (Bystander)**
Level 1: XAI for Perception	Given the current robot program does the robot succeed or fail at picking up a box?	Where did the robot place the box it just picked up?	Is the robot detecting me now that I am near it?
Level 2: XAI for Comprehension	Why does the robot fail at picking up the box in some circumstances?	Why does the robot place the box in given locations over the course of the task?	Does the robot's model cause it to avoid me if I am detected?
Level 3: XAI for Projection	How could the robot's model change to enable it to pick up the box successfully?	If I move the pallet, will the robot place the boxes in different locations?	If I moved in the space, would the robot still detect and avoid me?

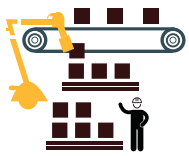
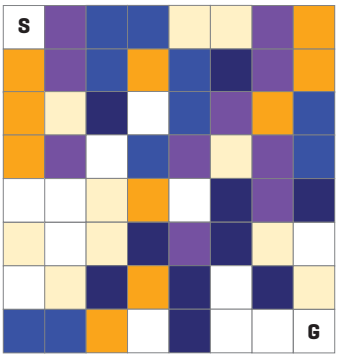
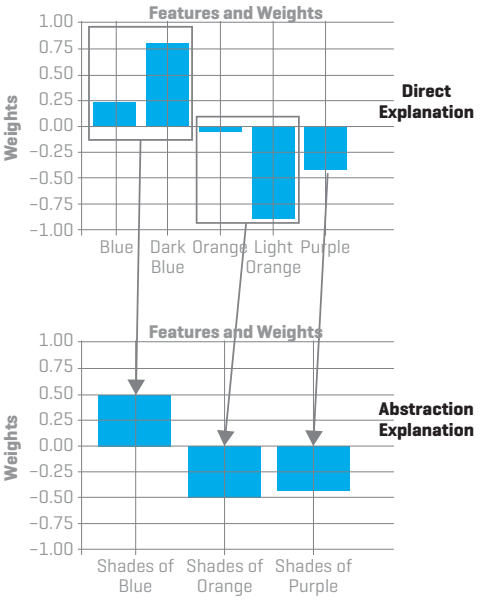


Figure 3. Example of an abstraction-based explanation approach for a grid world domain.

Here, there is a reward function consisting of various features (colors) and their weights, which could be communicated directly, as shown in the bar graph at the top of the figure. The corresponding abstract explanation which groups individual colors into shades of these colors and assigns these abstractions new weights is depicted in the bar graph at the bottom of the figure.





might hurt task performance overall. XAI systems must therefore provide information that supports a human user's "calibrated trust" or, in other words, appropriate trust in an AI system or robot.

In order to calibrate user trust in AI systems, XAI systems should provide information related to the bases of trust: purpose, process, and performance [8]. Purpose refers to the degree to which the system is being used within the realm of the designer's intent, process refers to the appropriateness of a system's algorithm for the situation in which it is working and the extent to which it can contribute to the human-AI team's goals, and performance refers to the system's demonstrated operations, including characteristics such as reliability, predictability, and ability. In the manufacturing scenario for example, the worker who interacts directly with the robot would need to know which tasks the programmer programmed the robot to perform (purpose), whether the robot can adapt to perform these tasks across all relevant possible factory conditions (process), and how accurately the robot will perform the set of tasks it has been programmed to perform (performance).

Trust specificity, which is the differentiation of trust between functions, subfunctions, and modes of automation, is also important to consider in the development of XAI. XAI systems must facilitate both functionally-specific ("local") trust and overall ("global") trust in the AI system which enables users to generalize to new contexts and scenarios. In the manufacturing scenario, to enable local trust calibration, the worker might need to know not only how successfully the robot can pick and place boxes, but also how successfully it can detect humans and plan safe motions accordingly. In terms of global trust calibration, it would be useful for the worker to know how successfully the robot can perform all of its individual tasks together as an integrated system.

Increasing user trust in automation is often suggested as a motivation for providing explanations to humans.

ASSESSING EXPLANATION QUALITY

Within human factors literature, assessments of situation awareness, workload, and trust have previously been proposed and validated, and these can be leveraged by XAI researchers in order to rigorously assess the efficacy of XAI techniques across different domains and contexts [5]. While such assessments cannot provide direct measures of explanation quality in themselves, adequate SA, appropriate workload, and calibrated trust are necessary (but not sufficient) components of human-AI team performance, which is also the ultimate aim of XAI. Therefore, these evaluations can tell us how well XAI systems achieve these intermediate ends.

For example, the situation awareness-based global assessment technique (SAGAT) is a widely-applied approach for objectively measuring a human's SA by probing their understanding of key information at various points throughout an interaction. In the context of XAI evaluation, it can be applied to assess human SA as it relates to AI behavior. Approaches for assessing workload subjectively (such as the commonly-applied NASA task load index [9]) and objectively (such as primary and secondary task measures that track how well a person performs at a task while the number or complexity of simultaneously-performed tasks varies) can further be applied to understand how XAI techniques impact a person's cognitive workload. To assess trust, behavior-based metrics, such as human reliance on and compliance with an AI system or robot, can be applied as objective assessments, and validated trust scales can be applied as subjective measures. Together, assessments like these can shed light on the quality of explanations across various contexts.

HUMAN FACTORS TRADEOFFS IN EXPLANATION DESIGN

A recent study compared a wide variety of explanation techniques through human-subject experiments and found tradeoffs that exist be-

tween the various human factors discussed previously [10]. In this particular study, information about an autonomous agent's objectives (in the form of a reward function) was explained to humans through a variety of XAI techniques that either provided information about the reward function directly or through examples of agent behavior that corresponded to the reward function. The study found techniques that provided comprehensive information about the objectives corresponded to improved understanding (i.e., situation awareness) over those that did not but resulted in higher workload. In designing and implementing XAI techniques in the future, considering this tradeoff will be critical to ensuring effective human–autonomy interaction. XAI designers must ask which information is absolutely necessary to explain and what a person's overall workload will look like over the course of an interaction, when deciding how much or how little information to provide and at what frequency to provide this information.

STRIKING A BALANCE WITH ABSTRACTIONS

In the same study, one explanation technique in particular struck a nice balance between all of the factors that were studied: Providing abstract representations of the agent's objectives corresponded to improved understanding of the communicated information over all other approaches (besides the approach which provided the objectives directly to participants), lower workload than all other approaches, and improved subjective assessment over all other approaches (where the scale applied for subjective assessment included trust-related questions). This suggests abstraction-based explanation approaches might be particularly effective for supporting human SA without increasing workload to an unmanageable degree.

Figure 3 is an example of an abstraction-based explanation approach for a grid world domain. In the aforementioned experiment, abstractions were hand-designed based on the domain and task.

Although the abstraction approach was shown to be effective in recent experiments, a number of open questions related to the design of ideal abstractions remain. For example, the best abstract representation of any given information might depend on the particular domain or the task the person receiving the explanation must perform. In addition, user expertise most likely plays an important role in the design of ideal abstractions: Experts in a particular domain may prefer to receive information at lower levels of abstraction, while novices may prefer higher levels of abstraction. Beyond this, it is likely possible to abstract away too much information, and determining the ideal balance of complexity reduction with information fidelity will be critical. Future research should investigate the design of ideal abstractions at greater length.

Ensuring transparency of robot behavior will be critical to the safe and successful integration of robots into our communities...

WHAT'S NEXT FOR ROBOT EXPLAINABILITY?

As we have seen, insights from human factors can help us design more transparent autonomous agents and robots through the application of human-centered explainable AI. Situation awareness can help us to define which information must be communicated to humans in order to support them with their tasks, workload considerations can help us understand how much information to communicate and at what frequency, and trust considerations can help us ensure that information is provided in order to appropriately calibrate user trust in these systems to prevent over- or under-reliance. Recent experimental results have demonstrated that tradeoffs between these various factors exist, and these must be accounted for in the design and evaluation of explanation techniques in the future. While there are still many unexplored directions and questions in the field of XAI and explainable robots, one thing that is clear is we should strive to keep humans at the center in our future research.

References

[1] Global industrial robot sales doubled over the past five years. International Federation of Robotics. Oct. 18, 2018; https://ifr.org/ifr-press-releases/news/global-industrial-robot-sales-doubled-over-the-past-five-years

[2] Sales of robots for the service sector grew by 37% worldwide. International Federation of Robotics. Oct. 26, 2022; https://ifr.org/ifr-press-releases/news/sales-of-robots-for-the-service-sector-grew-by-37-worldwide

[3] Gunning, D. and David A. DARPA's explainable artificial intelligence (XAI) program. *AI Magazine 40*, 2 (2019), 33–58; https://doi.org/10.1609/aimag.v40i2.2850

[4] Miller, T., Howe, P., and Sonenberg, L. Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. In *IJCAI-17 Workshop on Explainable AI (XAI)*. 2017; https://doi.org/10.48550/arXiv.1712.00547

[5] Sanneman, L. and Shah., J. A. The situation awareness framework for explainable AI (SAFE-AI) and human factors considerations for XAI systems. *International Journal of Human–Computer Interaction* 38, 18–20 (2022), 1772–88; https://doi.org/10.1080/10447318.2022.2081282

[6] Endsley, M. R. Toward a theory of situation awareness in dynamic systems. *Human Factors* 37, 1 (1995), 32–64.

[7] Wickens, C. D. Multiple resources and mental workload. *Human Factors* 50, 3 (2008), 449-55; https://doi.org/10.1518/0018720957790495

[8] Lee, John D., and See, K. A. Trust in automation: Designing for appropriate reliance. *Human Factors* 46, 1 (2004), 50-80; https://doi.org/10.1518/hfes.46.1.50_30392

[9] Hart, S. G. and Staveland, L. E. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology*, 52 (1988), 139–83; https://doi.org/10.1016/S0166-4115(08)62386-9

[10] Sanneman, L. and Shah, J. A. An empirical study of reward explanations with human-robot interaction applications. *IEEE Robotics and Automation Letters* 7, 4 (2022), 8956-63; https://doi.org/10.1109/LRA.2022.3189441

Biography

Lindsay Sanneman recently received her Ph.D. in autonomous systems in the Department of Aeronautics and Astronautics at the Massachusetts Institute of Technology. Her research focuses on the development of models, metrics, and algorithms for explainable AI (XAI) and AI alignment in complex human–autonomy interaction settings.

1528-4972/23/09 $15.00