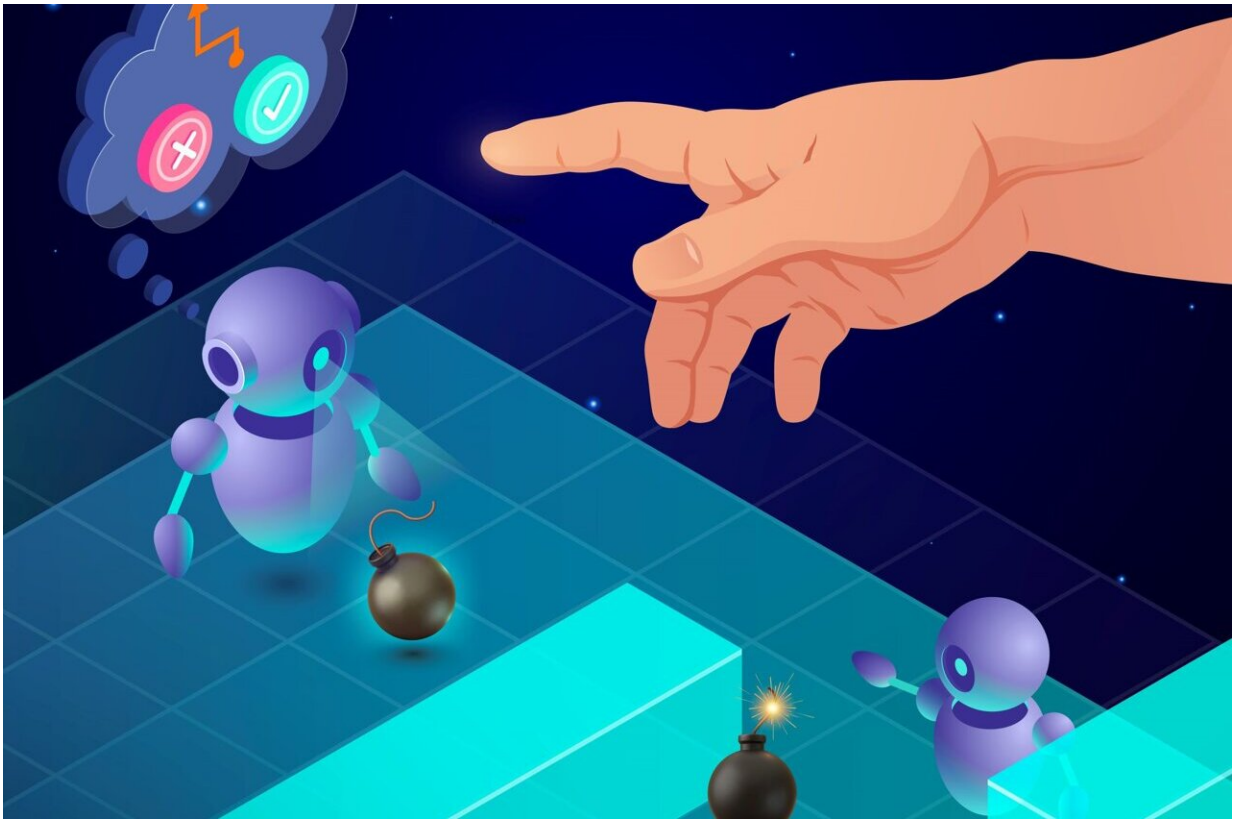
A new explainable AI paradigm that could enhance human-robot collaboration

August 10 2022, by Ingrid Fadelli



In the scout exploration game, to align with human values, the robot learns from human feedback to the proposals. Image Credit: Ms. Zhen Chen@BIGAI.

Artificial intelligence (AI) methods have become increasingly advanced over the past few decades, attaining remarkable results in many real-

world tasks. Nonetheless, most existing AI systems do not share their analyses and the steps that led to their predictions with human users, which can make reliably evaluating them extremely challenging.

A group of researchers from UCLA, UCSD, Peking University and Beijing Institute for General Artificial Intelligence (BIGAI) has recently developed a new AI system that can explain its decision-making processes to human users. This system, introduced in a paper published in *Science Robotics*, could be a new step toward the creation of more reliable and understandable AI.

"The field of explainable AI (XAI) aims to build collaborative trust between robots and humans, and the DARPA XAI Project served as a great catalyst for advancing research in this area," Dr. Luyao Yuan, one of the first authors of the paper, told TechXplore. "At the beginning of the DARPA XAI project, research teams primarily focus on inspecting models for classification tasks by revealing the decision process of AI systems to the user; for instance, some models can visualize certain layers of CNN models, claiming to achieve a certain level of XAI."

Dr. Yuan and his colleagues participated in the DARPA XAI project, which was specifically aimed at developing new and promising XAI systems. While participating in the project, they started reflecting on what XAI would mean in a broader sense, particularly on the effects it might have on collaborations between humans and machine.

The team's recent paper builds on one of their previous works, also published in *Science Robotics*, where the team explored the impact that explainable systems could have on a user's perceptions and trust in AI during human-machine interactions. In their past study, the team implemented and tested an AI system physically (i.e., in the real-world), while in their new study they tested it in simulations.

"Our paradigm contrasts with almost all of those proposed by teams in the DARPA XAI program, which primarily focused on what we call the passive machine–active user paradigm," Prof. Yixin Zhu, one of the project's supervisors, told TechXplore. "In these paradigms, human users need to actively check and attempt to figure out what the machine is doing (thus 'active user') by leveraging some models that reveal the AI models' potential decision-making process."

XAI systems that follow what Prof. Zhu refers to as the "passive machine-active user" paradigm require users to constantly check-in with the AI to understand the processes behind its decisions. In this context, a user's understanding of an AI's processes and trust in its predictions does not impact the AIs future decision-making processes, which is why the machine is referred to as "passive."

In contrast, the new paradigm introduced by Dr. Yuan, Prof. Zhu and their colleagues follows what the team refers to as an active machine-active user paradigm. This essentially means that their system can actively learn and adapt its decision-making based on the feedback it receives by users on the fly. This ability to contextually adapt is characteristic of what is often referred to as the third/next wave of AI.

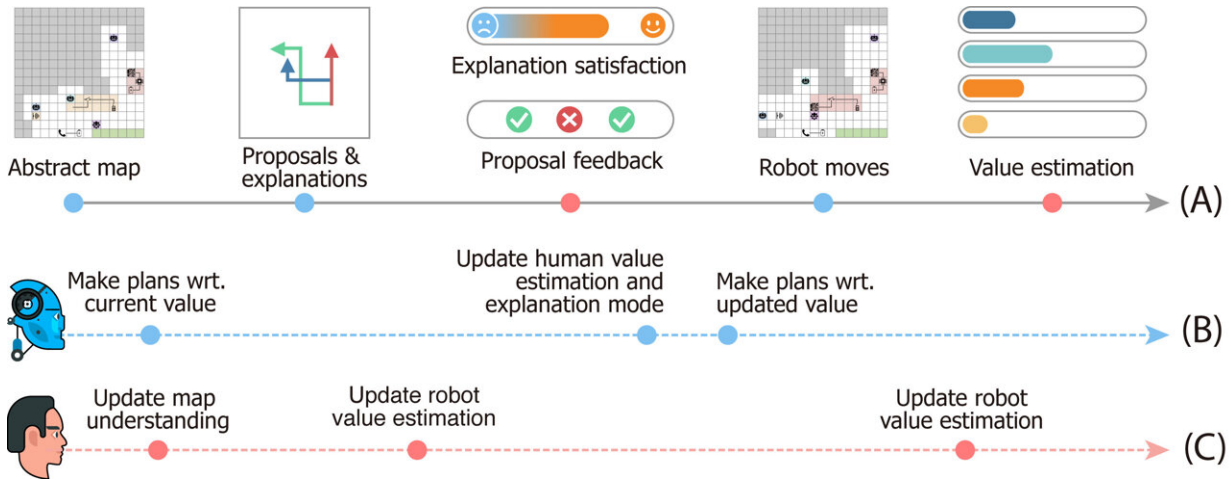
"To have AI systems assist their users as we expect them to, current systems require the user to code in expert-defined objectives," Dr. Yuan said. "This limits the potential of human-machine teaming, as such objectives can be hard to define in many tasks, making AI systems inaccessible to most people. To address this issue, our work enables robots to estimate users' intentions and values during the collaboration in real-time, saving the need to code complicated and specific objectives to the robots beforehand, thus providing a better human-machine teaming paradigm."

The goal of the system created by Dr. Yuan and his colleagues is to

attain so-called "value alignment." This essentially means that a human user can understand why a robot or machine is acting in a specific way or coming to specific conclusions, and the machine or robot can infer why the human user is acting in specific ways. This can significantly enhance human-robot communication.

"This bidirectional nature and real-time performance are the biggest challenges of the problem and the highlight of our contributions," Prof. Zhu said. "Putting the above points together, I think you'll now understand why our paper's title is "In situ bidirectional human-robot value alignment."

To train and test their XAI system, the researchers designed a game called "scout exploration," in which humans need to complete a task in teams. One of the most important aspects of this game is that the humans and robots need to align their so-called "value functions."



Study design of the Scout Exploration Game. Timeline (A) denotes events happening in a single round of the game. Timelines (B) and (C) depict the mental dynamics of the robots and the user, respectively. Image Credit: Ms. Zhen Chen@BIGAI.

"In the game, a group of robots can perceive the environment; this emulates real-world applications where the group of robots is supposed to work autonomously to minimize human interventions," Prof. Zhu said. "The human user, however, cannot directly interact with the environment; instead, the user was given a particular value function, represented by the importance of a few factors (e.g., the total time to complete the time, and resources collected on the go)."

In the scout exploration game, the team of robots do not have access to the value function given to human users, and they need to infer it. As this value cannot be easily expressed and communicated, to complete the task the robot and human team must infer it from one another.

"The communication is bidirectional in the game: on one hand, the robot proposes multiple task plans to the user and explains the pros and cons for each one of them, and on the other the user gives feedback on the proposals and rates each explanation," Dr. Xiaofeng Gao, one of the first authors of the paper, told TechXplore. "These bidirectional communications enable what is known as value alignment."

Essentially, to complete tasks in "scout exploration," the team of robots must understand what the human users' value function is simply based on the human's feedback. Meanwhile, human users learn the robots' current value estimations and can offer feedback that helps them to improve, and ultimately guides them towards the correct response.

"We also integrated theory of mind into our computational model, making it possible for the AI system to generate proper explanations to reveal its current value and estimate users' value from their feedback in real-time during the interaction," Dr. Gao said. "We then conducted extensive user studies to evaluate our framework."

In initial evaluations, the system created by Dr. Yuan, Prof. Zhu, Dr. Gao and their colleagues achieved remarkable results, leading to the alignment of values in the scout exploration game on the fly and in an interactive way. The team found that the robot aligned with the human user's value function as early as 25% into the game, while users could gain accurate perceptions of the machine's value functions about half way into the game.

"The pairing of convergence (i) from the robots' value to the user's true values and (ii) from the user's estimate of the robots' values to robots' current values forms a bidirectional value alignment anchored by the user's true value," Dr. Yuan said. "We believe that our framework highlights the necessity of building intelligent machines that learn and understand our intentions and values through interactions, which are critical to avoiding many of the dystopian science fiction stories depicted in novels and on the big screen."

The recent work by this team of researchers is a significant contribution to the area of research focusing on the development of more understandable AI. The system they proposed could serve as an inspiration for the creation of other XAI systems where robots or smart assistants actively engage with humans, sharing their processes and improving their performance based on the feedback they receive from users.

"Value alignment is our first step towards generic human-robot collaboration," Dr. Yuan explained. "In this work, value alignment happens in the context of a single task. However, in many cases, a group of agents cooperates in many tasks. For example, we expect one household robot to help us with many daily chores, instead of buying many robots, each only capable of doing one type of job."

So far, the researchers XAI system has attained highly promising results.

In their next studies, Dr. Yuan, Prof. Zhu, Dr. Gao and their colleagues plan to explore instances of human-robot value alignment that could be applied across many different real-world tasks, so that human users and AI agents can accumulate information that they acquired about each other's processes and capabilities as they collaborate on different tasks.

"In our next studies, we also seek to apply our framework to more tasks and physical robots," Dr. Gao said. "In addition to values, we believe that aligning other aspects of mental models (e.g., beliefs, desires, intentions) between humans and robots would also be a promising direction."

The researchers hope that their new explainable AI paradigm will help to enhance collaboration between humans and machines on numerous tasks. In addition, they hope that their approach will increase humans' trust in AI-based systems, including smart assistants, robots, bots and other virtual agents.

"For instance, you can correct Alexa or Google Home when it makes an error; but it will make the same error the next time you are using it," Prof. Zhu added. "When your Roomba goes somewhere you don't want it to go and tries to fight it, it doesn't understand as it only follows the pre-defined AI logic. All these prohibit modern AI from going into our homes. As the first step, our work showcases the potential of solving these problems, a step closer to achieving what DARPA called 'contextual adaptation' in the third wave of AI."

More information: Luyao Yuan et al, In situ bidirectional human-robot value alignment, *Science Robotics* (2022). DOI: 10.1126/scirobotics.abm4183

Project website: yzhu.io/publication/mind2022scirob