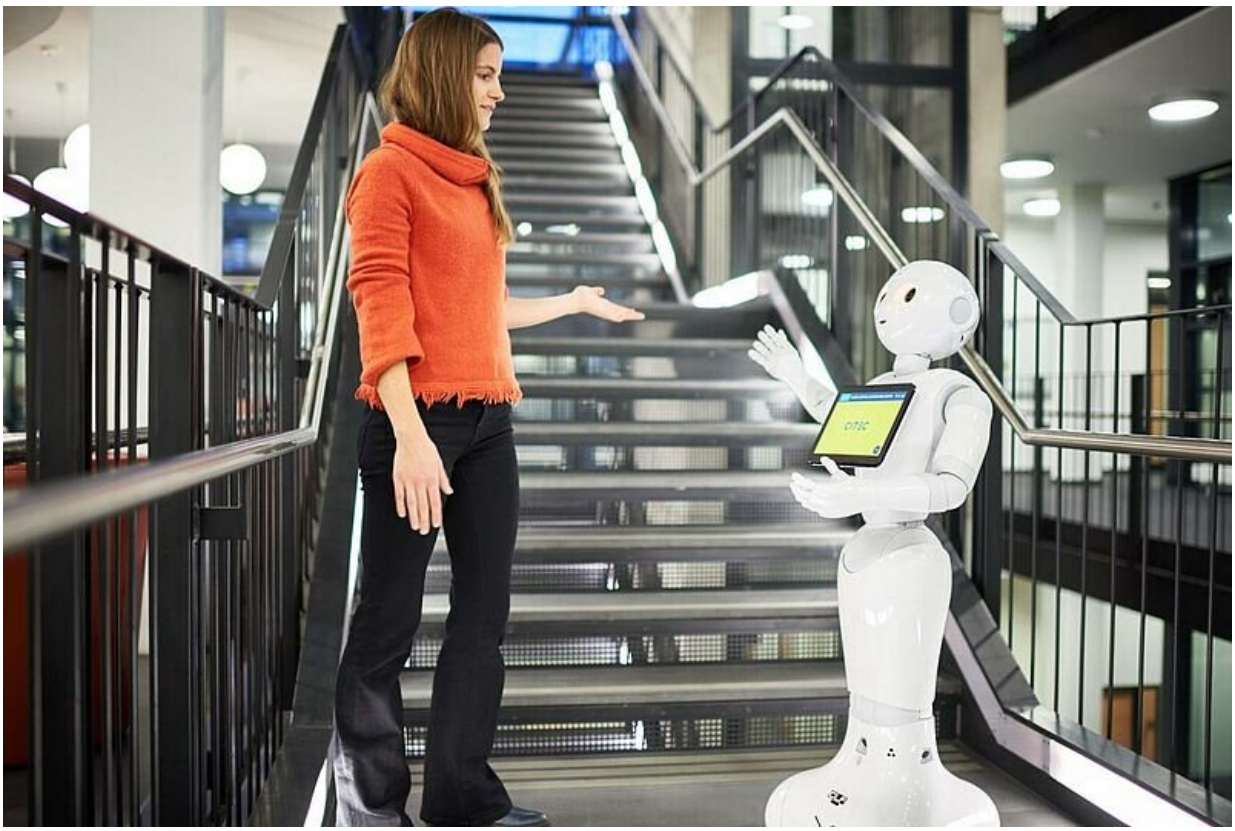
Making artificial intelligence understandable: Constructing explanation processes

March 8 2021



Human-machine interaction is complex. Explanations are needed to understand computer-based decisions. Credit: Paderborn University

Sifting through job applications, analyzing X-ray images, suggesting a

new track list—interaction between humans and machines has become an integral part of modern life. The basis for these artificial intelligence (AI) processes is algorithmic decision-making. However, as these are generally difficult to understand, they often prove less useful than anticipated. Researchers at Paderborn and Bielefeld University are hoping to change this, and are discussing how the explainability of artificial intelligence can be improved and adapted to the needs of human users. Their work has recently been published in the respected journal *IEEE Transactions on Cognitive and Developmental Systems*. The researchers describe explanation as a social practice, in which both parties co-construct the process of understanding.

Explainability research

"Artificial systems have become complex. This is a serious problem—particularly when humans are held accountable for computer-based decisions," says Professor Philipp Cimiano, a computer scientist at Bielefeld University. Particularly in the area of medical prognosis or legal contexts, we need to understand how machine-driven decisions are made, continues Cimiano. He points out that while there are already some approaches that address the explainability of such systems, they do not go far enough. Professor Katharina Rohlfing at Paderborn University agrees that further action is urgently needed: "Citizens have the right for algorithmic decisions to be made transparent. There are good reasons why this issue is specifically mentioned in the European Union's General Data Protection Regulation." The goal of making algorithms accessible is central to what is known as "eXplainable Artificial Intelligence (XAI)": "In explainability research, the focus is currently on the desired outcomes of transparency and interpretability," says Rohlfing, describing the latest research.

Understanding how decisions are made

The team involved in this research study go one step further and are investigating computer-based explanations from various different perspectives. They start from the assumption that explanations are only understandable to the users if they are not just presented to them, but if the users are involved in formulating them: "As we know from many everyday situations, good explanations are worth nothing if they do not take account of the other person's knowledge and experience. Anyone who wonders why their application was rejected by an algorithm is not generally interested in finding out about the technology of machine learning, but asks instead about how the data was processed with regard to their own qualifications," explains Rohlfing.

"When people interact with one another, the dialogue between them ensures that an explanation is tailored to the understanding of the other person. The dialogue partner asks questions for further explanation or can express incomprehension which is then resolved. In the case of artificial intelligence there are limitations to this because of the limited scope for interaction," continues Rohlfing. To address this, linguists, psychologists, media researchers, sociologists, economists and computer scientists are working closely together in an interdisciplinary team. These experts are investigating computer models and complex AI systems as well as roles in communicative interaction.

Explanation as a social practice

The Paderborn and Bielefeld researchers have developed a conceptual framework for the design of explainable AI systems. Rohlfing says: "Our approach enables AI systems to answer selected questions in such a way that the process can be configured interactively. In this way an explanation can be tailored to the dialogue partner, and social aspects can be included in decision-making." The research team regards explanations as a sequence of actions brought together by both parties in a form of social practice.

The aim is for this to be guided by 'scaffolding' and 'monitoring.' These terms come originally from the field of developmental studies. "To put it simply, scaffolding describes a method in which learning processes are supported by prompts and guidance, and are broken down into partial steps. Monitoring means observing and evaluating the reactions of the other party," explains Rohlfing. The researchers' objective is to apply these principles to AI systems.

New forms of assistance

This approach aims to expand on current research and provide new answers to societal challenges in connection with artificial intelligence. The underlying assumption is that the only successful way to derive understanding and further action from an explanation is to involve the dialogue partner in the process of explanation. At its core, this is about human participation in socio-technical systems. "Our objective is to create new forms of communication with genuinely explainable and understandable AI systems, and in this way to facilitate new forms of assistance," says Rohlfing in summary.

More information: Katharina J. Rohlfing et al. Explanation as a social practice: Toward a conceptual framework for the social design of AI systems, *IEEE Transactions on Cognitive and Developmental Systems* (2020). DOI: 10.1109/TCDS.2020.3044366

Provided by Universität Paderborn