

University of Groningen

## A Research Agenda for Hybrid Intelligence

Akata, Zeynep; Balliet, Dan; de Rijke, Maarten; Dignum, F; Dignum, V.; Eiben, Guszti; Fokkens, Antske; Grossi, Davide; Hindriks, Koen; Hoos, Holger

*Published in:*  
Computer

*DOI:*  
[10.1109/MC.2020.2996587](https://doi.org/10.1109/MC.2020.2996587)

**IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.**

*Document Version*  
Publisher's PDF, also known as Version of record

*Publication date:*  
2020

[Link to publication in University of Groningen/UMCG research database](#)

### *Citation for published version (APA):*

Akata, Z., Balliet, D., de Rijke, M., Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K., Hoos, H., Hung, H., Jonker, C., Monz, C., Neerincx, M., Oliehoek, F., Prakken, H., Schlobach, S., van der Gaag, L., van Harmelen, F., ... Welling, M. (2020). A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *Computer*, 53(8), 18-28. <https://doi.org/10.1109/MC.2020.2996587>

### **Copyright**

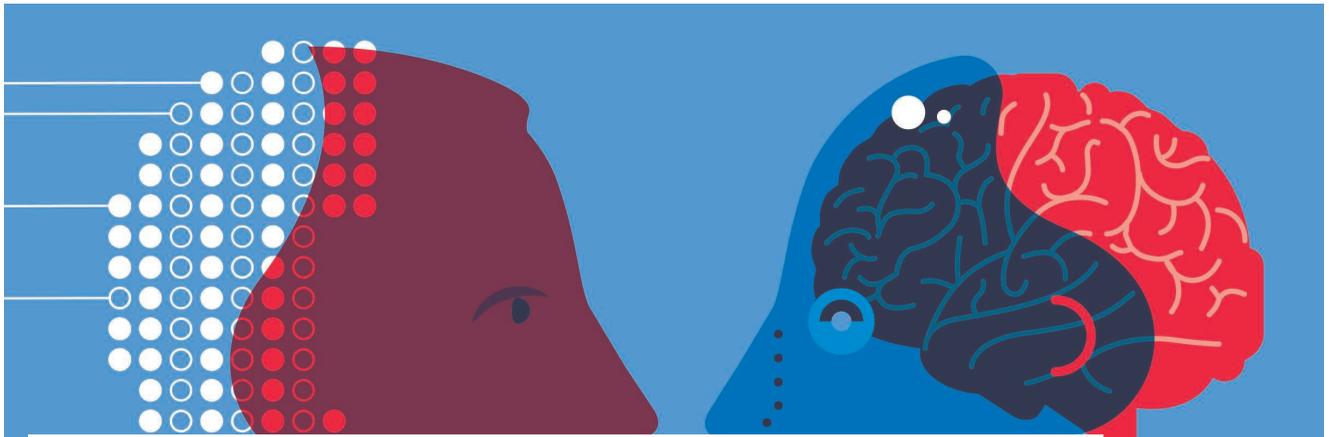
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

### **Take-down policy**

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

*Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.*



# A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence

**Zeynep Akata**, University of Amsterdam and University of Tübingen

**Dan Balliet**, Vrije Universiteit Amsterdam

**Maarten de Rijke**, University of Amsterdam

**Frank Dignum**, Utrecht University

**Virginia Dignum**, TU Delft

**Gusztai Eiben and Antske Fokkens**, Vrije Universiteit Amsterdam

**Davide Grossi**, University of Groningen

**Koen Hindriks**, Vrije Universiteit Amsterdam

**Holger Hoos**, Leiden University

**Hayley Hung and Catholijn Jonker**, TU Delft

**Christof Monz**, University of Amsterdam

**Mark Neerincx and Frans Oliehoek**, TU Delft

**Henry Prakken**, Utrecht University

*Digital Object Identifier 10.1109/MC.2020.2996587*  
*Date of current version: 30 July 2020*

**Stefan Schlobach**, Vrije Universiteit Amsterdam

**Linda van der Gaag**, Utrecht University and Dalle Molle Institute

**Frank van Harmelen**, Vrije Universiteit Amsterdam

**Herke van Hoof**, University of Amsterdam

**Birna van Riemsdijk and Aimee van Wynsberghe**, TU Delft

**Rineke Verbrugge and Bart Verheij**, University of Groningen

**Piek Vossen**, Vrije Universiteit Amsterdam

**Max Welling**, University of Amsterdam

*We define hybrid intelligence (HI) as the combination of human and machine intelligence, augmenting human intellect and capabilities instead of replacing them and achieving goals that were unreachable by either humans or machines. HI is an important new research focus for artificial intelligence, and we set a research agenda for HI by formulating four challenges.*

**O**ver the course of history, the use of tools has played a crucial role in enabling human civilizations, cultures and economies: fire, the wheel, the printing press, the computer, and the Internet are just a few of humanity's crucial innovations. Such tools have augmented human skills and thought to previously unachievable levels. Over the past several decades, artificial intelligence (AI) techniques, which allow humans to "scale up" by providing increasingly intelligent decision support, have become the latest addition to this toolset. Until now, however, these tools have been mostly used by experts. Hybrid intelligence (HI) can go well beyond this by creating systems that operate as mixed teams, where humans and machines cooperate synergistically, proactively, and purposefully to achieve shared goals, showing AI's potential for amplifying instead of replacing human intelligence. This perspective on AI as

HI is critical to our future understanding of AI as a way to augment human intellect as well as to our ability to apply intelligent systems in areas of crucial importance to society.

Contemporary societies face problems that have a weight and scale novel to humanity, such as global pandemics, resource scarcity, environmental conservation, climate change, and maintaining democratic institutions. To solve these problems, humans need help to overcome some of their limitations and cognitive biases: poor handling of probabilities, entrenchment, short termism, confirmation bias, functional fixedness, stereotypes, in-group favoritism, and so forth. We need help from intelligent machines that challenge our thinking and support our decision making, but we do not want to be ruled by machines and their decisions, nor do we want to supplant human biases with those of machines. Instead, we need cooperative problem-solving

approaches in which machines and humans contribute through a collaborative conversation, where machines engage with us, explain their reasoning, behave responsibly, and learn from their mistakes.

AI systems tend to be "idiots savants," reaching or exceeding the performance of human experts in a very narrow range. There is a danger that users (be they individuals or organizations) will overestimate the range of expertise of an automated system and deploy it for tasks at which it is not competent, with potentially catastrophic consequences. Human experts are needed in the loop to ensure that this does not happen. This is an urgent problem; at present, there are deployed AI systems that were not designed with societal values such as fairness, accountability, and transparency in mind. This contributes to today's problems of "fake news," Facebook messages leading to ethnic and religious violence, and the

large-scale manipulation of elections. This lack of alignment with human values is impacting us more frequently. Now that AI technologies affect our everyday lives at an ever-increasing pace, there is a greater need for AI systems that work synergistically with humans rather than ones that simply replace them. Thought leaders in AI increasingly share the conviction that, for AI systems to augment our abilities and compensate for our weaknesses, we need a new understanding of AI that takes humans and humanity explicitly into account.<sup>1</sup> It is better to view AI systems not as “thinking machines” but as cognitive prostheses that can help humans think and act better.<sup>2</sup>

### WHAT IS HYBRID INTELLIGENCE?

We define HI as the combination of human and machine intelligence, augmenting human intellect and capabilities instead of replacing them, to make meaningful decisions, perform appropriate actions, and achieve goals that were unreachable by either humans or machines alone. HI requires interaction between artificial intelligent agents and humans, taking human expertise and intentionality into account, together with ethical, legal, and societal (ELS) considerations. The main HI research challenge is as follows: how to build adaptive intelligent systems that augment rather than replace human intelligence, leverage our strengths, and compensate for our weaknesses while taking into account ethical, legal, and societal considerations.

Developing HI requires fundamentally new solutions to core research problems in AI. Modern AI technology surpasses humans in many pattern recognition, machine learning, reasoning, and optimization tasks, but it

falls short on general world knowledge; common sense; and the human capabilities of collaboration, adaptability, and responsibility in terms of norms and values and explanation. Humans, on the other hand, excel in collaboration, flexibly adapting to changing circumstances during the execution of a task. An essential element in our collaboration is the capability to explain motivations, actions, and results. And humans always operate in a setting where norms and values (often implicitly) delineate which goals and actions are desirable or even permissible. We therefore unpack the challenge of building HI systems into four research challenges:

- › *Collaborative HI*: How do we develop AI systems that work in synergy with humans?
- › *Adaptive HI*: How can these systems learn from and adapt to humans and their environment?
- › *Responsible HI*: How do we ensure that they behave ethically and responsibly?
- › *Explainable HI*: How can AI systems and humans share and explain their awareness, goals, and strategies?

In the following sections, we discuss the state of the art for each of these challenges, leading to a set of research questions to be addressed to achieve hybrid intelligent systems as envisaged previously.

### COLLABORATIVE HI

#### State of the art

Collaboration in human teams is vital, pooling different skills to solve more difficult problems than any of the members could alone. The skills that computer systems excel in are different from those

of humans. A key question is therefore how to best exploit this complementarity in human-machine collaboration. Early results in successful complementary human-machine collaboration in cognitive tasks are known from negotiation tasks, planning, behavior change support systems, and “centaur” chess. There are key challenges when promoting machines from tools to partners: a computational understanding of human actors, a theory of mind, an understanding of joint actions in teams, and social norms such as reciprocity, which are crucial in such teamwork. Hybrid intelligent machines will need to both perceive social behavior by collaborators and communicate with their collaborators using multiple modalities. Our notion of collaborative HI goes beyond the established notions of human-in-the-loop machine learning<sup>3</sup> or interactive AI by aiming for reciprocity between human and computer agents, as discussed in the following sections.

**Understanding human actors.** To exploit skill differences, we need models that make machines aware of these differences and enable them to proactively provide support by exploiting skill complementarity. In addition, machines can help prevent common human biases and limitations, such as a bias toward short-term rewards, a confirmation bias, entrenchment, in-group favoritism, a limited attention span, and limited short-term memory. Solutions can build on the substantial research of how to mitigate cognitive biases.<sup>4</sup>

**Theory of mind.** Maintaining the beliefs, goals, and other mental attitudes of other people in a theory of mind (ToM) is essential for effective cooperation. In complex social interactions,

people also need to apply a second-order ToM (“She thinks that I plan to go right”). There is substantial theory on people’s use of and difficulties with ToM. A relatively unexplored area is the use of recursive ToM in hybrid groups containing humans, robots, and software agents, allowing an agent to recursively apply a ToM to detect anomalies in its state of mind. de Weerd et al.<sup>5</sup> show how second-order ToM is beneficial in competitive, cooperative, and mixed-motive situations and how software agents of different ToM levels can support humans to achieve better negotiation outcomes.

**Teamwork, joint actions, plans, and tasks.** In multiagent systems (MASs), substantial work has been performed on distributing tasks and monitoring plan progression. Frequently used systems such as TAEMS consider only software agent teams, not hybrid teams of humans and agents. Thus, many results might not carry over to hybrid teams because humans typically react differently from agents in unexpected situations and are not likely to accept orders from agents in all circumstances and so on. Recent work on an agreement framework proves to support human-agent teams when they dynamically adapt their task allocation and coordination. Cooperation and teamwork have been extensively studied in economic disciplines and specifically in game theory, including within MASs.<sup>6</sup> Game theory has already had several high-impact ramifications in the MAS field and will provide ways to inform artificial agents in hybrid teams of the tradeoffs involved in collaborative tasks and how to best manage them.

**Reciprocity, social norms, and culture.** The social and biological sciences

have converged on a common understanding that kinship, direct reciprocity, indirect reciprocity, and the social learning of norms can explain why and how humans cooperate.<sup>7</sup> Further, people can quickly and efficiently interpret social situations along various parameters (for example, mutual dependence,

communication, human-computer interfacing, and other component technologies, such as facial expression analysis and gesture detection, that show the importance of multimodal interaction for collaboration.<sup>9</sup> The same can be said about multimodal dialogue systems and, more



**RECENT WORK ON AN AGREEMENT FRAMEWORK PROVES TO SUPPORT HUMAN-AGENT TEAMS WHEN THEY DYNAMICALLY ADAPT THEIR TASK ALLOCATION AND COORDINATION.**



power, and conflict), and this can shape their willingness to cooperate. Computational theories of reciprocity show that the effect of reciprocity has similar effects on artificial agents. For such agents to interact with humans in ways that promote collaboration, HI systems should be aware of these traits in humans and use this knowledge to engage in actions that can positively influence human collaboration. Initial work has been done to incorporate social norms in agents and develop new architectures for social agents. That designing for interdependencies and coactivity makes the system more effective was proved by the success of the Florida Institute for Human and Machine Cognition team that secured second place in the DARPA challenge,<sup>8</sup> where its team capabilities and interaction design were based on the coactive design method.

**Multimodal interaction.** There is a long tradition of research on multimodal

recently, chatbot systems using neural networks. In all these studies, the assumption is made that systems process signals correctly. They also consider tasks separately and not systems as a whole. There are few systems that combine natural language communication and perception for the purpose of task-oriented learning. She and Chai<sup>10</sup> describe a system that is instructed through multimodal interaction to perform a physical task. This system deals with the uncertainties of perceived sensor data and interpretation of the instructions, but it does not assume that humans and AI systems work together and is limited to very basic physical actions.

**Machine perception of social and affective behavior.** In the growing branch of multimodal interaction concerned with human social behavior, the fields of affective computing and social signal processing have made great leaps with respect to the machine

perception, modeling, and synthesis of social cues; individual and social constructs; and emotion. There has been a paradigm shift in research on the perception of human behavior, going away from training machine learning models using data collected in the lab to settings in controlled, real-life settings. However, moving from controlled laboratory studies to real-life settings requires a fundamental change in experimental approaches. As argued by Hung et al.,<sup>11</sup> we need to transition from expecting clearly visible video footage of frontal faces and use other sensing modalities to exploit the arsenal of social signals that are emitted by humans.

### Research questions

The aforementioned state of the art leads to the following research questions for collaboration in hybrid systems:

- › What are the appropriate models for negotiation, agreements, planning, and delegation in hybrid teams?
- › How can a computational ToM (based on social and psychological concepts) be designed to plan collaboration between humans and artificial agents?
- › How can HI exploit experience sharing for the purpose of establishing common ground, resolving uncertainties and conflicts, adjusting tasks and goals, and correcting actions?
- › Which specific challenges and advantages arise when groups of humans and agents collaborate, given the complementarities in their skills and capabilities?
- › How can multimodal messages, expressions, gestures, and semi- or unstructured representations

be understood and generated for the purpose of collaboration?

### ADAPTIVE HI

In HI settings, artificial and human agents work together in complex environments. Such environments are seldom static: team composition and tasks can change, interpersonal relations evolve, preferences can shift, and external conditions (for example, available resources and environment) can vary over time. Thus, competences cannot be fixed before deployment, and agents will have to adapt and learn during operation. As such, the ability of HI systems to adapt or learn is a prerequisite not only to perform well but to function at all. To accomplish such adaptivity, agents need to deploy machine learning techniques to learn from data, experiences, and dialogues with other agents (human or artificial).

### State of the art

There is an inherent tension between the adaptive nature of HI systems and the desire for their safety and reliability. Constraints on the adaptivity of a system are needed to avoid adaptations that are undesirable from the point of view of safety, either for the agent or the environment, or from the standpoint of ethical and social acceptability. Such constraints may be encoded in the reward/loss functions of the learning system, symbolically encoded, or implemented through the modification of the adaptive exploration process. Highly adaptive systems also pose a challenge to the transparency and explainability of a system's actions or advice. Data, settings, concepts, and competences all interact in the decision-making process. The system's architecture thus needs to keep track of all these changes to trace back why a specific decision was made

at a specific point in time. Furthermore, these systems must not only keep track of such information but also be able to effectively communicate it to a variety of users to elicit necessary feedback.

Several research directions within AI have focused on learning models that can adapt to either changing users, tasks, resources, or environments. For instance, multitask learning aims to find models for a range of tasks. Transfer learning approaches try to adapt learned models from source tasks to target tasks that could differ in either environment or objective. A growing body of work has also studied the use of metalearning for rapid adaptation. Metalearning methods attempt to learn a solution strategy from a collection of previously solved tasks to, for example, discover optimal exploration strategies. Adapting to the changing preferences of the user can be addressed using multiobjective models and methods, which model different reward functions for different desirable features of a solution. Recently, so-called automated machine learning methods have been developed to select and optimize learning algorithms for specific tasks or data sets.

Various aspects and subproblems of the challenge of adaptive HI have already been addressed in the literature. For example, to handle user preferences that change over time, different preference-elicitation strategies have been compared, and multiobjective optimization has been used to adapt an information retrieval system to the current user preferences. Incomplete knowledge about the preferences of negotiation parties has also been used to inform multiattribute negotiation systems. However, none of these approaches combine techniques for learning from data streams

or dialogues. Furthermore, there is no explicit strategic reasoning on what the best learning techniques would be, given the task and circumstances. The subproblem of adaptivity to changes in the environment has been studied in the form of robot controllers that adapt depending on the environmental conditions, and even the morphology of robots can be adapted to the environment. Finally, fully automated procedures have been developed for selecting and configuring algorithms for a given supervised machine learning task<sup>12</sup> and are rapidly gaining traction.

### Research questions

The state of the art discussed in the previous section leads to the following research questions for adaptivity in hybrid systems:

- › How can interaction in a mixed group of agents (humans and machines) be used to improve learning systems, for example, by communicating intent and asking for and handling complex feedback?
- › How can learning systems respect the societal, legal, ethical, safety, and resource constraints that might be expressed symbolically?
- › How can learning systems accommodate changes in user preferences, environments, tasks, and available resources without having to completely relearn each time something changes?
- › How can the learning mechanism itself be adapted to improve efficiency and effectiveness in highly dynamic HI settings based on task experience as well as human guidance?
- › How can the adaptivity of machine learning techniques

be integrated with the precision and interpretability of symbolic knowledge representation and reasoning?

### RESPONSIBLE HI

Modern AI techniques often put users in situations in which information about their decisions is unknown or unclear, and the ability to dispute a decision is not possible. Advances in AI increasingly lead to concerns about the ability of such systems to behave according to legal constraints and moral values. Models and techniques are needed to evaluate, analyze, and design AI systems with the capability to reason about and act according to legal constraints and moral values as well as to understand the consequences of their decisions. The urgency of these questions is increasingly acknowledged by researchers and policy makers alike, as shown from recent reports by the IEEE Ethically Aligned Design of Autonomous Systems; the United Nations Educational, Scientific and Cultural Organization; the French government; the U.K. House of Lords; and the European Commission. In the following sections, we describe a dual approach for dealing with the challenges concerning legal and ethical HI systems.

### State of the art

#### **Ethical reasoning about HI systems.**

Where it concerns the legal and regulatory governance of HI systems, current research focuses on whether existing legal systems can deal with the consequences of introducing artificial systems. However, the liability of and for any (semi)autonomous system remains a challenge, requiring a better understanding between lawyers and computer scientists

of concepts such as legal personhood (which does not require moral agency), human autonomy (which does not stand in the way of strict liability), and machine autonomy (which does not imply self-consciousness, let alone moral agency).

Many different solutions have been developed and discussed: from strict liability for manufacturers, to reversing the burden of proof, to compulsory certification or automated compensation in the case of smart contracts. This relates to the position of AI systems: are they tools or (anthropocentric) moral entities with moral patience and distribution of responsibility? To ensure responsibility, deliberation should ideally include a grounding in moral concepts, allowing for explanations based in and coordinated over values (such as privacy), social norms and relationships, commitments, habits, motives, and goals. Underlying all of these is the need to analyze the social, ethical, and legal characteristics of the domain. The “design for values” approaches<sup>13</sup> and methods used to identify and align the possibly conflicting values of all stakeholders<sup>14</sup> are well-known candidates for these tasks. Translating abstract values to more concrete design requirements is an important area where more research is needed to make these approaches effective in designing responsible HI.

#### **Ethical reasoning by HI systems.**

Ethical reasoning is an even more controversial issue. When creating artificial moral agents, that is, machines that are embedded with ethical reasoning capabilities, the following questions arise: Can machines comprehend the world of ethics? Which ethics should be programmed? Can machines be assigned moral roles or capacities?

Should machines be made accountable or responsible for consequences? The methods and tools used to design the ethical behavior of intelligent agents are either descriptive or focus on modeling moral reasoning as a direct translation of some well-known moral theory, modeling moral agency in a general way, or designing an ethical agent architecture. Other approaches take a fundamentally interactive approach to normative reasoning by HI systems, allowing users to express their norms and values to the system at runtime. Ethical decision making then emerges from the resulting human-machine interaction. This is motivated by the observation that, in particular for personal and intimate technologies, the choice of how to support a person is highly context dependent.

On the other hand, research in AI and the law on artificial legal reasoning is reasonably well developed. Deductive techniques have been practically successful, especially in the application of knowledge-based systems in the large-scale processing of administrative law, such as social benefit law and tax law, and, more recently, for legal advice and regulatory compliance. Such systems apply computational representations of legislation to the facts as interpreted by the human user. However, such systems often suffer from the well-known “knowledge acquisition bottleneck,” which has proved a major barrier to the practical exploitation of intelligent techniques in many domains. The recent success of deep learning and natural language processing applied to huge corpora of unstructured legal information may provide opportunities, but employing them in the right way to obtain the necessary knowledge to overcome this barrier is highly challenging. Finally,

most approaches to AI and the law and AI and ethics do not clearly take the collective and distributed dimension of interaction into account. Work on norms and institutions in multiagent systems<sup>22</sup> can be used to prove that specific rules of behavior are observed when making decisions.

### Research questions

The aforementioned state of the art leads to the following research questions:

- › How can ELS considerations be included in the HI development process (ethics in design)?
- › What is the best way to verify the agent’s architecture and behavior to prove their ethical “scope” (ethics in design)?
- › What is the best way to measure ELS performance and compare designed versus learning systems (ethics in design)?
- › What are the ELS concerns around the development of systems that can reason about ELS consequences of their decisions and actions (ethics by design)?
- › Which methodology can ensure ELS alignment during the design, development, and use of ELS-aware HI systems (ethics by design)?
- › What new computational techniques are required for ELS in the case of HI systems where humans and artificial agents work together?

### EXPLAINABLE HI

People look for explanations to improve their understanding of someone or something so that they can derive a stable model to be used for prediction and control. By building more transparent, interpretable, or explainable

artificial agents, human agents will be better equipped to understand, trust, and work with intelligent agents. A recent trend is to distinguish between interpretation and explanation. In the case of interpretation, abstract concepts are translated into insights that are useful for domain knowledge (for example, identifying correlations between layers in a neural network for language analysis and linguistic knowledge). An explanation provides information that gives insights to users as to how a model came to a decision or interpretation. Models of how humans explain decisions and behavior can be used to design and implement intelligent agents that provide explanations, including how people employ biases and social expectations when they generate and evaluate an explanation.

de Graaf and Malle<sup>15</sup> argue that the anthropomorphization of agents causes users to expect explanations utilizing the same conceptual framework used to explain human behaviors. This suggests a focus on everyday explanations, that is, explanations of why particular facts (events, properties, decisions, and so on) occurred rather than of more general relationships, such as in a scientific explanation. Trust is lost when users cannot understand observed behavior or decisions, which necessitates effective solutions that must combine AI with insights from the social sciences and human-computer interactions.

Everyday explanations are contrastive; people do not ask why an event happened but rather why it happened instead of another event. Moreover, explanations are selective (in a biased manner); people rarely expect a complete causal chain of events as explanation. Humans are adept at selecting one or two causes from a large chain of

them to be the explanation; however, this selection is influenced by certain cognitive biases. In addition, explanations are social, that is, they are a transfer of knowledge as part of an interaction, and thus are presented relative to the explainer's beliefs about the explainee's beliefs.

### State of the art

AI has a long history of work on explanation. In early work on expert systems, users rated the ability to explain decisions as the most desirable feature of a system design to assist in decision making. Studies consistently show that explanations significantly increase users' trust as well as their ability to correctly assess whether an algorithmic decision is accurate. The need for explaining the decisions of expert systems was discussed as early as the 1970s, with early work already stressing the importance of explanations that are not merely traces but also contain justifications. Lacave and Díez<sup>16</sup> survey methods of explanation for Bayesian networks and distinguish between the reasoning, model, and evidence for the decision. Recommender systems have long had facilities to produce justifications to help users decide whether to follow a recommendation.

Studies from the early 2000s show that users are much more satisfied with systems that contain some form of justification. Early work on explanations in machine learning focused on visualizing predictions to support experts in assessing models. This line of work continues to this day, for example, with techniques for producing visualizations of the hidden states of neural networks. Another line of work on explainability in machine learning develops models that are intrinsically interpretable and

can be explained through reasoning, such as decision lists or trees. Other approaches have created sparse models via feature selection or extraction to optimize interpretability.

Today, considerable work is focused on interpreting and explaining the predictions of complex ("black box") models. Methods for improving the interpretability of neural networks aim at identifying what information is captured in various layers of the neural network. Diagnostic probing methods, for instance, investigate which properties can be predicted from individual layers of a neural network by testing whether these properties can be predicted by a regression model. These methods have shown, for example, that lower layers of models used for interpreting natural language perform reasonably well on syntactic categories such as part-of-speech tasks whereas higher layers are more successful for more semantic-oriented properties.

Correlation-based methods such as singular value canonical correlation analysis and representation similarity analysis can be used to identify correlations between layers in different models. Here, the inner layers of a more complex model under investigation are typically compared to the output layer of a model trained on a more basic task that identifies information likely to be relevant for the complex task as well. Examples of methods that support explanation of the output of a neural network include layerwise relevance propagation, which uses the gradients of the network to determine the relevance of previously seen input. Contextual decomposition, on the other hand, computes how information from a specific input propagates throughout the model. The insights provided by

these methods help identify how the model arrives at specific decisions and are thus typical examples of explanatory features.

Previously, many studies that focused on the explainability of machine learning algorithms were conducted from a human-computer interaction angle, that is, questions such as how users interact with the system and how explanations can help with this are asked. These studies do not focus on how to construct faithful explanations to describe the underlying decisions of the algorithm. Recently, the focus has shifted toward 1) describing the training process, 2) explaining the outcomes and the relationship to the training material, and 3) the underlying algorithm. As to the first, Ross et al.<sup>17</sup> use the gradients of the output probability of a model with respect to the input to define feature importance in a predictive model, but this is restricted to differentiable models. Concerning the second, Koh and Liang<sup>18</sup> deal with finding the most influential training objects so as to make a model's prediction more understandable. And concerning the third, Ribeiro et al.<sup>19</sup> introduce LIME, a method used to locally explain the classifications of any classifier.

### Research questions

The state of the art described in the previous section leads to the following research questions for explainability in hybrid systems:

- ▶ How can shared representations be built and used as the basis for explanations, covering both the external world and the internal problem-solving process?
- ▶ What are the different types of explanations that make the

decision-making process more transparent and understandable?

- › How can explanations be communicated to users such that they improve the user's trust and leads to a successful agent-user collaboration?
- › How can explanations be personalized so that they align with the users' needs and capabilities?
- › How can the quality and strength of the explanations be evaluated?

## EXAMPLE APPLICATIONS OF HI

HI techniques can be applied across many domains, and we expect them to bring major economic and societal benefits to those applications. In the following sections, we outline three potential scenarios that illustrate the use of HI (namely, health care, education, and science) in demonstrating its potential, and we direct the interested reader to additional sources for more details. Although implementations of all of these scenarios have been tested, HI is a new research focus, and the results described are preliminary examples of what future HI systems may look like.

**Education.** A child with learning difficulties is supported by a team in which the child's remedial teacher, an educational therapist, and a Nao robot collaborate. Together, they design a targeted learning program, monitor progress, and provide encouragement. The robot combines expertise from the human team members with its own observations and gives advice on possible adjustments to the program. Interacting with the Nao robot helps the child to stay focused and have fun for a longer period of time. (Visit [www.robotsindeclas.nl](http://www.robotsindeclas.nl) for an early example

of how robots can be deployed in the classroom.)

**Health care.** A teenage leukemia patient is accompanied 24/7 by a robot dog during multiple prolonged stays in the hospital. A large medical team collaborates with this HI agent to answer the patient's questions. Simple questions, for example, on diet and daily schedule, are autonomously answered by the embodied agent. More complex medical questions are routed to medical staff members according to their medical discipline, available knowledge, and rapport with the patient. The dog explains the inevitable medical terminology, remembering what has been explained before. It monitors the teenager's mood and advises the specialists on the patient's psychological well-being. (Visit <https://goo.gl/CNN8iM> for an early example of how robots can support children during long-term hospital stays.)

**Science.** A scientist in a commercial pharmaceutical lab is investigating a chemical compound expected to have an inhibitory effect on neurodegeneration. Overwhelmed by the enormous amounts of data available online, the scientist turns to the lab's HI virtual assistant. Data volume is not a problem for this assistant, who searches through dozens of databases, scans the recent literature, and fires off a few emails to authors of relevant papers while making sure not to include scientists who work at competing big pharma companies and consulting with the HI system of a sister lab in China. The scientist and the HI agent analyze the findings and conclude that the compound has been investigated before and failed to show the required inhibitory activity. Thanks to HI, all of this could be done in

a day rather than weeks. (Visit <https://goo.gl/CajqnM> for an early example of our work.)

Based on these case studies, we are formulating generalizable design patterns that capture reusable patterns in both the HI architecture as well as reusable interaction patterns with these systems. For example, Lighthart et al.<sup>20</sup> identify five interaction patterns for the "getting acquainted" phase of an HI system, including open-ended and closed questions and prechoreographed turn taking. An evaluation of 75 8–11-year-old children shows substantially different efficacy between the various behaviors of the HI system. Similarly, van Harmelen and Ten Teije<sup>21</sup> describe how a large number of hybrid system architectures can be captured in a limited number of design patterns.

In this article, we argued that AI research should include the quest for systems that collaborate with people instead of focusing mainly on systems that replace people. We defined the notion of HI and formulated the main research challenges to be faced. We identified four central properties that are required for such hybrid intelligent systems: collaborative, adaptive, responsible, and explainable. For each of these, we discussed the state of the art and formulated a number of key research questions to be addressed. We also briefly illustrated the use of hybrid intelligent systems in three example application scenarios. 

## REFERENCES

1. S. Kambhampati, Challenges of human-aware AI systems. 2019. [Online]. Available: [arXiv:1910.07089](https://arxiv.org/abs/1910.07089)

## ABOUT THE AUTHORS

**ZEYNEP AKATA** is a professor of computer science at the University of Tübingen. Contact her at zeynepakata@gmail.com.

**DAN BALLIET** is a professor of experimental and applied psychology and the head of the Amsterdam Cooperation Lab at the Vrije Universiteit Amsterdam. Contact him at d.p.balliet@vu.nl.

**MAARTEN DE RIJKE** is a professor of artificial intelligence at the University of Amsterdam. Contact him at derijke@uva.nl.

**FRANK DIGNUM** is a professor of socially aware artificial intelligence at Umeå University. Contact him at f.p.m.dignum@uu.nl.

**VIRGINIA DIGNUM** is a professor at Umeå University. Contact her at m.v.dignum@tudelft.nl.

**GUSZTI EIBEN** is a professor of artificial intelligence at the Vrije Universiteit Amsterdam. Contact him at a.e.eiben@vu.nl.

**ANTSKE FOKKENS** is an assistant professor at the Vrije Universiteit Amsterdam. Contact her at antske.fokkens@vu.nl.

**DAVIDE GROSSI** is an associate professor of artificial intelligence at the University of Groningen. Contact him at d.grossi@rug.nl.

**KOEN HINDRIKS** is a professor of social artificial intelligence at the Vrije Universiteit Amsterdam. Contact him at k.v.hindriks@vu.nl.

**HOLGER HOOS** is a professor of machine learning at Leiden University. Contact him at hh@liacs.nl.

**HAYLEY HUNG** is an associate professor at the Technical University Delft. Contact her at H.Hung@tudelft.nl.

**CATHOLIJN JONKER** is a professor of interactive intelligence at the Technical University Delft. Contact her at c.m.jonker@tudelft.nl.

**CHRISTOF MONZ** is an associate professor at the University of Amsterdam. Contact him at c.monz@uva.nl.

**MARK NEERINCX** is a professor of human-centered computing at the Technical University Delft. Contact him at mark.neerincx@tno.nl and m.a.neerincx@tudelft.nl.

**FRANS OLIEHOEK** is an associate professor at the Technical University Delft. Contact him at f.a.oliehoek@tudelft.nl.

**HENRY PRAKKEN** is a professor of legal informatics at the University of Groningen. Contact him at h.prakken@uu.nl.

**STEFAN SCHLOBACH** is an associate professor at the Vrije Universiteit Amsterdam. Contact him at k.s.schlobach@vu.nl.

**LINDA VAN DER GAAG** is a senior researcher at the Dalle Molle Institute, Lugano, Switzerland. Contact her at l.c.vanderGaag@uu.nl.

**FRANK VAN HARMELEN** is a professor of artificial intelligence at the Vrije Universiteit Amsterdam. Contact him at frank.van.harmelen@vu.nl.

**HERKE VAN HOOF** is an assistant professor at the University of Amsterdam. Contact him at h.c.vanhoof@uva.nl.

**BIRNA VAN RIEMSDIJK** is an associate professor at the University of Twente. Contact her at m.b.vanriemsdijk@utwente.nl.

**AIMEE VAN WYNSBERGHE** is an associate professor of ethics and technology at the Technology University Delft. Contact her at A.L.Robbins-vanWynsberghe@tudelft.nl.

**RINEKE VERBRUGGE** is a professor of logic and cognition at the University of Groningen. Contact her at l.c.verbrugge@rug.nl.

**BART VERHEIJ** is a professor of artificial intelligence and argumentation at the University of Groningen. Contact him at bart.verheij@rug.nl.

**PIEK VOSSSEN** is a professor of computational lexicology at the Vrije Universiteit Amsterdam. Contact him at piek.vossen@vu.nl.

**MAX WELLING** is a professor of machine learning at the University of Amsterdam. Contact him at m.welling@uva.nl.

2. J. Guszczka, "Smarter together: Why artificial intelligence needs human-centered design," Deloitte, London, no. 22, 2018. [Online]. Available: [https://www2.deloitte.com/content/dam/insights/us/articles/4214\\_Smarter-together/DI\\_Smarter-together.pdf](https://www2.deloitte.com/content/dam/insights/us/articles/4214_Smarter-together/DI_Smarter-together.pdf)
3. S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza, "Power to the people: The role of humans in interactive machine learning," *AI Mag.*, vol. 35, no. 4, pp. 105–120, 2014. doi: 10.1609/aimag.v35i4.2513.
4. M. B. Cook and H. S. Smallman, "Human factors of the confirmation bias in intelligence analysis: Decision support from graphical evidence landscapes," *Hum. Factors*, vol. 50, no. 5, pp. 745–754, 2008. doi: 10.1518/001872008X354183.
5. H. de Weerd, R. Verbrugge, and B. Verheij, "How much does it help to know what she knows you know? An agent-based simulation study," *Artif. Intell.*, vol. 199–200, pp. 67–92, June–July 2013. doi: 10.1016/j.artint.2013.05.004.
6. D. Grossi and P. Turrini, "Dependence in games and dependence games," *J. Auton. Agents Multi-Agent Syst.*, vol. 25, no. 2, pp. 284–312, 2012. doi: 10.1007/s10458-011-9176-3.
7. A. Romano and D. Balliet, "Reciprocity outperforms conformity to promote cooperation," *Psychol. Sci.*, vol. 28, no. 10, pp. 1490–1502, 2017. doi: 10.1177/0956797617714828.
8. M. Johnson et al., "Team IHMC's lessons learned from the DARPA robotics challenge trials," *J. Field Robot.*, vol. 32, no. 2, pp. 192–208, 2015. doi: 10.1002/rob.21571.
9. I. Rauschert, P. Agrawal, R. Sharma, S. Fuhrmann, I. Brewer, and A. MacEachren, "Designing a human-centered, multimodal GIS interface to support emergency management," in *Proc. 10th ACM Int. Symp. Advances Geographic Information Systems*, 2002, pp. 119–124. doi: 10.1145/585147.585172.
10. L. She and J. Y. Chai, "Interactive learning of grounded verb semantics towards human-robot communication," in *Proc. 55th Annu. Meeting Assoc. Computational Linguistics (ACL)*, 2017, pp. 1634–1644. [Online]. Available: <https://dblp.uni-trier.de/pers/hd/s/She:Lanbo>
11. H. Hung, E. Gedik, and L. Cabrera-Quiros, "Complex conversational scene analysis using wearable sensing," in *Multi-modal Behavior Analysis in the Wild: Advances Challenges*, X. Alameda-Pineda, E. Ricci, N. Sebe, Eds. New York: Academic, 2018, pp. 225–245.
12. F. Hutter, L. Kotthoff, and J. Vanschoren, Eds., *Automated Machine Learning Methods, Systems, Challenges*. New York: Springer Verlag 2019. [Online]. Available: <https://www.automl.org/book/>
13. J. van den Hoven, P. Vermaas, and I. van de Poel, "Sources, theory, values and application domains," in *Handbook of Ethics, Values, and Technological Design*. Dordrecht: Springer-Verlag, 2015. [Online]. Available: <https://link.springer.com/reference-work/10.1007%2F978-94-007-6994-6>
14. I. Verdiesen, V. Dignum, and J. Van Den Hoven, "Measuring moral acceptability in E-deliberation: A practical application of ethics by participation," *ACM Trans. Internet Technol.*, vol. 18, no. 4, Art. no. 43. 2018, doi: 10.1145/3183324.
15. M.A. de Graaf, and B.F. Malle, "How people explain action (and Autonomous Intelligent Systems should too)," in *Proc. Association Advancement Artificial Intelligence Fall Symp. (AAAI)*, 2017, pp. 19–26.
16. C. Lacave and F. Diez, "A review of explanation methods for Bayesian networks," *Knowl. Eng. Rev.*, vol. 17, no. 2, pp. 107–127, 2002. doi: 10.1017/S026988890200019X.
17. S. Ross, M. C. Hughes, and F. Doshi-Velez, Right for the right reasons: Training differentiable models by constraining their explanations. 2017. [Online]. Available: [arXiv:1703.03717](https://arxiv.org/abs/1703.03717)
18. P. W. Koh and P. Liang, Understanding black-box predictions via influence functions. 2017. [Online]. Available: [arXiv:1703.04730](https://arxiv.org/abs/1703.04730)
19. M. T. Ribeiro, S. Singh and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *Proc. 22nd ACM SIG KDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.
20. M. Lighthart, T. Fernhout, M. A. Neerincx, L. A. Van Bindsbergen, M. A. Grootenhuis, and K. V. Hindriks, "A child and a robot getting acquainted: Interaction design for eliciting self-disclosure," in *Proc. 18th Int. Conf. Autonomous Agents and MultiAgent Systems*, May 2019, pp. 61–70.
21. F. Van Harmelen and A. Ten Teije, "A boxology of design patterns for hybrid learning and reasoning systems," *J. Web Eng.*, vol. 18, no. 1, pp. 97–124, 2019. doi: 10.13052/jwe1540-9589.18133.
22. V. Dignum and F. Dignum, "Modelling agent societies: Co-ordination frameworks and institutions." in *Portuguese Conference on Artificial Intelligence (EPIA 2001) (Lecture Notes in Computer Science series 2258)*. Berlin: Springer-Verlag, 2001, pp. 191–204.