A research agenda for Hybrid Intelligence:

augmenting human intellect by collaborative, adaptive, responsible and explainable AI

Zeynep Akata, University of Amsterdam; Dan Balliet, Vrije Universiteit Amsterdam; Frank Dignum, Utrecht University; Virginia Dignum, TU Delft; Guszti Eiben, Vrije Universiteit Amsterdam; Antske Fokkens, Vrije Universiteit Amsterdam; Linda van der Gaag, Utrecht University; Davide Grossi, University of Groningen, Frank van Harmelen, Vrije Universiteit Amsterdam; Koen Hindriks, Vrije Universiteit Amsterdam; Herke van Hoof, University of Amsterdam; Holger Hoos, Leiden University; Hayley Hung, TU Delft; Catholijn Jonker, TU Delft; Christof Monz, University of Amsterdam; Mark Neerincx, TU Delft; Frans Oliehoek, TU Delft; Henry Prakken, Utrecht University; Birna van Riemsdijk, TU Delft; Maarten de Rijke, University of Amsterdam; Stefan Schlobach, Vrije Universiteit Amsterdam; Rineke Verbrugge, University of Groningen, Bart Verheij, University of Groningen, Piek Vossen, Vrije Universiteit Amsterdam; Max Welling, University of Amsterdam; Aimee van Wynsberghe, TU Delft.
(Authors are listed in alphabetical order)

Abstract

We define Hybrid Intelligence (HI) as the combination of human and machine intelligence, augmenting human intellect and capabilities instead of replacing them, and achieve goals that were unreachable by either humans or machines alone. We identify Hybrid Intelligence as an important new research focus for the field of Artificial Intelligence, and we set a research agenda for Hybrid Intelligence by formulating four challenges: how do we develop AI systems that work in synergy with humans (Collaborative HI); how can these systems learn from and adapt to humans and their environment (Adaptive HI); how do we ensure that they behave ethically and responsibly (Responsible HI); and how can AI systems and humans share and explain their awareness, goals and strategies (Explainable HI). For each of these four challenges, we survey the state of the art and define a research agenda.

The Need for Hybrid Intelligent Systems

Over the course of history, the use of tools have played a crucial role in enabling human civilisations, 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 human thought to previously unachievable levels. Over the past decades, Artificial Intelligence (AI) techniques have become the latest addition to this toolset that allows humans to "scale up", by providing increasingly intelligent decision support. However, until now, these tools are mostly used by experts. Hybrid Intelligence 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 the potential of AI to amplify human intelligence instead of replacing it. This perspective on Artificial Intelligence as Hybrid Intelligence 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 maintaining democratic institutions, global pandemics, resource scarcity, environmental conservation, and climate change. 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 favouritism and others. 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 by 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 organisations) 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: AI systems are deployed right now that are 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 large-scale manipulation of elections. This lack of alignment with human values is impacting us increasingly. Now that AI technologies affect our everyday lives at an ever-increasing pace, there is a greater need for AI systems to work synergistically with humans rather than to simply replace them. Thought leaders in AI increasingly share the conviction that in order for AI systems to augment our abilities and to compensate for our weaknesses, we need a new understanding of AI that takes humans and humanity explicitly into account (Kambhampati, 2018, AAAI presidential address). It is better to view AI systems not as “thinking machines,” but as cognitive prostheses that can help humans think and act better (Guszcza, 2018).

What is Hybrid Intelligence?

We define Hybrid Intelligence (HI) as the combination of human and machine intelligence, augmenting human intellect and capabilities instead of replacing them, in order to make meaningful decisions, perform appropriate actions, and achieve goals that were unreachable by either humans or machines alone. Hybrid Intelligence requires the interaction between artificial intelligent agents and humans, taking human expertise and intentionality into account, together with ethical, legal and societal considerations. The main HI research challenge is therefore:

How to build adaptive intelligent systems that augment rather than replace human intelligence,
that leverage our strengths and compensate for our weaknesses while taking into account ethical, legal and societal considerations?

Developing HI needs fundamentally new solutions to core research problems in AI: modern AI technology surpasses humans in many pattern recognition, machine learning, reasoning and optimisation tasks, but it falls short on general world knowledge, common sense, and the human capabilities of collaboration, adaptability, responsibility in terms of norms and values, and explanation. Humans on the other hand, excel in collaboration; they flexibly adapt to changing circumstances during 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?

For each of these challenges we will now discuss the state of the art, leading to a set of research questions to be addressed in order to achieve Hybrid Intelligent systems as envisaged above.

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 exploit this complementarity in human-machine collaboration. There is a great need to bridge the gap between ‘social animal and unsocial machine’ which is being partially addressed by the emerging field of Social Signal Processing. In addition, early results in successful complementary human-machine collaboration in cognitive tasks are known from negotiation tasks (Hindriks et al., 2008), planning (Sycara et al., 2010), behaviour change support systems (Schouten et al., 2017) and in `centaur' chess. Promoting machines from tools to partners faces key challenges: a computational understanding of human actors, a theory of mind, an understanding of joint actions in teams, and the social norms such as reciprocity that are crucial in such teamwork. Hybrid intelligent machines will need to both perceive social behaviour by collaborators, and communicate with their collaborators using multiple modalities.

Understanding human actors*.* In order 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 bias towards short-term rewards, confirmation bias, entrenchment, in-group favouritism, limited attention span and limited short-term memory. Solutions can build on the substantial research on how to mitigate cognitive biases (e.g., Cook & Smallman, 2008).

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; the formalisms presented in (Jonker, van Riemsdijk, et al., 2011), present meta-reasoning techniques allowing an agent to recursively apply a ToM to detect anomalies in its state of mind. De Weerd, Verbrugge & Verheij, (2013) 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. Vossen et al. (2018) describe a robot implementation that stores the results of perceptions and communication within a ToM model and captures uncertainties, gaps and alternative or conflicting information.

Teamwork, joint actions, plans and tasks: In Multi-Agent Systems (MAS), substantial work has been performed on distributing tasks and monitoring plan progression (Dunin-Keplicz & Verbrugge, 2010). Much used systems such as TAEMS only consider software agent teams but no hybrid teams of humans and agents. Thus, many results might not carry over to hybrid teams, as humans typically react differently from agents in unexpected situations, and are not likely to accept orders from agents in all circumstances, etc. Recent work on an agreement framework proves to support human-agent teams when they dynamically adapt their task allocation and coordination (Mioch et al., 2018). Cooperation and teamwork have been extensively studied in economic disciplines and specifically in game theory, including within MAS (Grossi & Turrini, 2012). 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 trade-offs 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 (Romano & Balliet, 2017). Further, people can quickly and efficiently interpret social situations along various parameters (e.g., mutual dependence, power, 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 (Ranjbar-Sahraei et al., 2014). In order for such agents to interact with humans in ways to 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 to develop new architectures for social agents (Dastani et al., 2009). That designing for interdependencies and co-activity makes the system more effective was proved by the success of the IHMC team ending overall at the 2nd place in the DARPA challenge (Johnson et al., 2015), where the team capabilities and interaction design were based on the co-active design method.

Multimodal interaction: There is a long tradition of research on multimodal communication, human-computer interfacing, and other component technologies such as facial expression analysis , and gesture detection that shows the importance of multimodal interaction for collaboration (Rauschert et al., 2002). The same can be said about multimodal dialogue systems, and more recently, around chatbot systems using neural networks (Serban et al., 2016). 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 & Chai (2017) describe a system that is instructed through multimodal interaction to perform a physical task. This system deals with uncertainties of perceived sensor data and the 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 behaviour: In the growing branch of multimodal interaction concerned with human social behaviour, the fields of affective computing and social signal processing have made great leaps with respect to the machine perception, modelling and synthesis of social cues, individual and social constructs, and emotion. There has been a paradigm shift in research on the perception of human behaviour 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 approach. As argued by Hung et al. (2018), we need to move away from expecting clearly visible video footage of frontal faces, and using other sensing modalities to exploit the arsenal of social signals that are being emitted by humans.

Research questions

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

- What are appropriate models for negotiation, agreements, planning and delegation in hybrid teams?
- How to design a computational theory of mind (based on social and psychological concepts) that can be used to investigate and design collaboration between humans and artificial agents?
- How can Hybrid Intelligence exploit experience sharing for the purpose of establishing common ground, resolving uncertainties and conflicts, adjusting tasks, 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 to understand and generate multimodal messages, expressions, gestures, and semi- or unstructured representations for the purpose of collaboration?

Adaptive HI

In HI settings, artificial agents 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 (e.g., available resources and environment) can vary over time. Thus, competences cannot be fixed before deployment and the 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, the agents need to deploy machine learning techniques to learn from data, from experience and from 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 undesirable from the point of view of ethical and social acceptability. Such constraints may be encoded in the reward/loss functions of the learning system; they may be symbolically encoded; or they may be implemented through modification of the adaptive exploration process. Highly adaptive systems also pose a challenge for 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 in order to be able 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 this information to a variety of users, in order 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 example, *multi-task 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 *meta-learning* for rapid adaptation. Meta-learning methods try to learn a solution strategy from a collection of previously-solved tasks to, e.g., discover optimal exploration strategies. Adapting to changing preferences of the user can be addressed with multi-objective models and methods, which model different reward functions for different desirable features of a solution. Recently, so-called automated machine learning (or *AutoML*) methods have been developed in order to select and optimize learning algorithms for specific tasks or data sets.

Various aspects and sub-problems 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 multi-objective 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 multi-attribute negotiation systems. However, none of these approaches combine techniques for learning from data streams or from dialogues. Furthermore, there is no explicit strategic reasoning on what the best learning techniques would be, given the task and circumstances. The sub-problem 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 (Eiben & Smith, 2015). Finally, fully automated procedures have been developed for selecting and configuring algorithms for a given supervised machine learning task (Hutter et al., 2019), and are rapidly gaining traction.

Research questions

The above state of the art 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, e.g., by communicating intent, asking for and handling complex feedback?
- How can learning systems respect 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 re-learn each time something changes?
- How can the learning mechanism itself be adapted to improve efficiency and effectiveness in highly dynamic Hybrid Intelligence 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, analyse and design AI systems with the capabilities 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 the recent reports by the IEEE Ethically Aligned Design of Autonomous Systems, UNESCO, the French government , the UK House of Lords, and the European Commission.

State of the art

We distinguish a dual approach to dealing with the challenges concerning legal and ethical HI systems:

Ethical reasoning about HI systems. Where it concerns 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 (Floridi & Sanders., 2004)? To ensure responsibility, deliberation should ideally include grounding in moral concepts, allowing explanations based in, and coordination over values (such as privacy), social norms and relationships, commitments, habits, motives, and goals. Underlying all of the above, there is a need to analyse the social, ethical and legal characteristics of the domain. The ‘design for values’ approaches (van den Hoven et al., 2015) and methods to identify and align the possibly conflicting values of all stakeholders (Verdiesen et al., 2018) 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 is an even more controversial issue. When creating artificial moral agents– machines that are embedded with ethical reasoning capabilities - questions arise such as: Can machines comprehend the world of ethics? How to decide on which ethics to program? Can machines be assigned moral roles or moral capacities? Should machines be made accountable or responsible for consequences? Methods and tools to design ethical behaviour of intelligent agents are either descriptive (Wallach & Allen, 2010) or focus on modelling moral reasoning (Bonnemains et al., 2016) as a direct translation of some well-known moral theory, on modelling moral agency in a general way (Lorini, 2012) or on designing an ethical agent architecture (Cointe, 2016). 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 run-time (Van Riemsdijk et al., 2015). 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 & Law on **artificial legal reasoning** is reasonably well developed. Deductive techniques have been practically successful, especially in the application of knowledge-based systems in 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 (Prakken & Sartor 2015). However, while in constrained applications deductive techniques can be very useful, they face two serious limitations. First, determining the facts requires common sense knowledge and probabilistic reasoning, which go beyond deduction. Second, since rigid deductive application of legal rules often leads to immoral, unfair or socially undesirable outcomes, judges and lawyers often resort to non-deductive forms of reasoning about the rules, including argumentation with cases, analogical reasoning and reasoning about purpose. AI & Law models of these non-deductive kinds of reasoning exist (for an overview see Ashley 2017). However, they have not yet scaled up to practical applications, since they are critically dependent on the possibility of acquiring and computationally representing large amounts of information, including common sense knowledge and knowledge about legal, ethical and societal values. This is an instance of the well-known ‘knowledge acquisition bottleneck’, which has proved a major barrier to the practical exploitation of intelligent techniques in many domains. Recent success of deep learning, data science and natural language processing applied to huge amounts of unstructured legal information that is currently available 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 & Law and AI & Ethics do not clearly take the collective and distributed dimension of interaction into account. Work on norms and institutions in multi-agent systems (Dignum, 1999) can be used to prove that specific rules of behaviour are observed when making decisions. There is also relevant research on theoretical frameworks for ethical plan selection that can be formally verified and research on how to guide institutional design to be coordinated by institutions, while not imposing unacceptable limits on agents’ rights (Dennis et al., 2016). However, these approaches do not yet integrate the flexible and context-dependent ways in which people are used to interpreting social and ethical norms.

Research questions

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

- How to include ethical, legal and societal (ELS) considerations in the HI

development process? (ethics in design)
- How to verify the agent's architecture and behaviour to prove their ethical 'scope' (ethics in design)
- How to measure ELS performance and compare designed systems vs 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 design, development and use of ELS-aware HI systems? (ethics by design)
- What new computational techniques are required for ELS in 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 that can 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 (e.g. identifying correlations between layers in a neural network for language analysis and linguistic knowledge). An explanation provides information that provides insights to users as to how a model came to a decision or interpretation. Models of how humans explain decisions and behaviour 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 (2017) argue that anthropomorphisation of agents causes users to expect explanations using the same conceptual framework used to explain human behaviours. This suggests a focus on *everyday explanations*, that is, explanations of why particular facts (events, properties, decisions, etc.) occurred, rather than explanations of more general relationships, such as in a scientific explanation. Trust is lost when users cannot understand observed behaviour or decisions, and effective solutions must combine AI with insights from the social sciences and human-computer interaction.

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 causes 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 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 (2002) survey methods of explanation for Bayesian networks and distinguish between the reasoning, the model, and the 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, e.g. 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 (Ustun & Rudin, 2016).

Today, there is considerable attention for work on interpreting and explaining the predictions of complex ("black box") models. Methods for improving 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 instance, that lower-layers of models 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 SVCCA (singular value canonical correlation analysis) and RSA (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 is 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 LRP (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 specific input propagates through the model. The insights provided by these methods help identifying how the model comes to specific decisions and are thus typical examples of explanatory features.

Previously, many studies that focus on the explainability of machine learning algorithms have been conducted from a Human Computer Interaction angle. That is, questions are asked such as “how do users interact with the system and how can explanations help with this?” These studies do not focus on how to construct faithful explanations to describe the underlying decisions of the algorithm. Recently, the focus is changing (a) towards describing the training process, (b) towards explaining the outcomes and the relation to the training material, and (c) towards the underlying algorithm. As to the first, Ross et al. (2017) uses 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 & Liang (2017) 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., (2016) introduce LIME, a method to locally explain the classifications of any classifier.

Research questions

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

- How to build shared representations to be 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 the explanations improve the user’s trust and leads to a successful agent-user collaboration?
- How can explanations be personalised so that they align with the users’ needs and capabilities?
- How can the quality and strength of the explanations be evaluated?

Example Applications of Hybrid Intelligence

Hybrid intelligence techniques can be applied across many domains, and we expect them to bring major economic and societal benefits in those applications. Here, we outline three concrete scenarios that illustrate the use of hybrid intelligence - namely, healthcare, education and science - to illustrate the potential of Hybrid Intelligence and we point the interested reader to further sources for additional details.

Education: A child with learning difficulties is supported by a team in which her remedial teacher, an educational therapist and a Nao robot collaborate. Together, they design a targeted learning programme, monitor progress, and provide encouragement. The robot combines expertise from the human team members with its own observations, and gives advice on possible adjustments of the programme. Interacting with the Nao robot helps the child to stay concentrated and have fun for longer (see www.robotsindeklas.nl for an early example of how robots can be deployed in the classroom).

Healthcare: A teenage leukemia patient is accompanied 24/7 by a robot dog during her multiple prolonged stays in hospital. A large medical team collaborates with this HI agent to answer the patient’s questions. Simple ones, e.g., on diet and daily schedule, are autonomously answered by the embodied agent. More complex medical questions are routed to medical staff, 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 wellbeing. (See 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, she 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 that work at competing big pharma companies, and consulting the HI system of a sister lab in China. The scientist and her HI agent analyse the findings and conclude that the compound has been investigated before, and failed to show the required inhibitory activity. Thanks to HI, all this could be done in a day rather than weeks. (See https://goo.gl/CajqnM for an early example of our work)

Conclusions

In this short position paper we have argued that AI research should include the quest for systems that collaborate with people, instead of (implicitly) often focusing on systems that replace people. We have defined the notion of Hybrid Intelligence, and formulated the main research challenge to be faced. We have identified four central properties that are required for such Hybrid Intelligent systems: they should be collaborative, adaptive, responsible and explainable (CARE). For each of these desiderata, we have discussed the state of the art and have formulated a number of key research questions to be addressed, and we have briefly illustrated the use of Hybrid Intelligent systems in 3 example application scenario’s.

Literature References

Kambhampati, S. (2018). Challenges of Human-Aware AI Systems. AAAI 2018 Presidential Address.

Guszcza, J. (2018). "Smarter together: Why artificial intelligence needs human-centered design." Deloitte Review(22).

Hutter, F., Kotthoff, L., and Vanschoren J. (Eds.), Automated Machine Learning Methods, Systems, Challenges, Springer Verlag 2019, https://www.automl.org/book/

Cook, M. B. and H. S. Smallman (2008). "Human factors of the confirmation bias in intelligence analysis: decision support from graphical evidence landscapes." Human Factors 50(5): 745-754.

de Weerd, H., R. Verbrugge and B. Verheij (2013). "How much does it help to know what she knows you know? An agent-based simulation study." Artificial Intelligence 199: 67-92.

Vossen, P., S. Baez, L. Bajčetić and B. Kraaijeveld (2018). "Leolani: a reference machine with a theory of mind for social communication." 21st International Conference on Text, Speech, and Dialogue, pgs 15-25, 11107 LNAI, Springer Verlag.

Grossi, D. and P. Turrini (2012). "Dependence in Games and Dependence Games. J. of Autonomous Agents and Multi-Agent Systems." Springer 25(2): 284–312.

Romano, A. and D. Balliet (2017). "Reciprocity Outperforms Conformity to Promote Cooperation." Psychological Science 28(10): 1490-1502.

Johnson, M., B. Shrewsbury, S. Bertrand, T. Wu and et al (2015). "Team IHMC's lessons learned from the DARPA robotics challenge trials." J. of Field Robotics 32(2): 192-208.

Rauschert, I., P. Agrawal, R. Sharma, S. Fuhrmann, I. Brewer and A. MacEachren (2002). Designing a human-centered, multimodal GIS interface to support emergency management. Proc. of the 10th

ACM Int. Symp. on Advances in geographic information systems.

She, L. and J. Y. Chai (2017). Interactive Learning of Grounded Verb Semantics towards Human-Robot Communication. Proc. of the 55th Annual Meeting of the Assoc. for Computational Linguistics.

Hung, H., E. Gedik and L. Cabrera-Quiros (2018). Complex Conversational Scene Analysis Using Wearable Sensing. Multi-modal Behavior Analysis in the Wild: Advances and Challenges.

van den Hoven, J., P. Vermaas and I. van de Poel (2015). Sources, Theory, Values and Application Domains. Handbook of Ethics, Values, and Technological Design: . Dordrecht, Springer.

Verdiesen, I., V. Dignum and J. Van Den Hoven (2018). "Measuring Moral Acceptability in E-deliberation: A Practical Application of Ethics by Participation." ACM Trans. Internet Technol. 18(4).

Ashley, K. D. (2017). Data-centric and logic-based models for automated legal problem solving. AI and Legal Analytics. New Tools for Law Practice in the Digital Age. Cambridge University Press. 25: 5-27.

How People Explain Action (and Autonomous Intelligent Systems Should Too), Maartje M. A. de Graaf, Bertram F. Malle, AAAI Fall Symposia 2017

Lacave, C., and Diez, F., A review of explanation methods for Bayesian networks, The Knowledge Engineering Review, Vol. 17, No. 2, 2002.

Ribeiro, M. T., S. Singh and C. Guestrin (2016). Why should I trust you?: Explaining the predictions of any classifier. Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining.

Ross, S., M. C. Hughes and F. Doshi-Velez (2017). "Right for the right reasons: Training differentiable models by constraining their explanations." arXiv preprint arXiv:1703.03717.

Koh, P. W. and P. Liang (2017). "Understanding black-box predictions via influence functions." arXiv preprint arXiv:1703.04730.