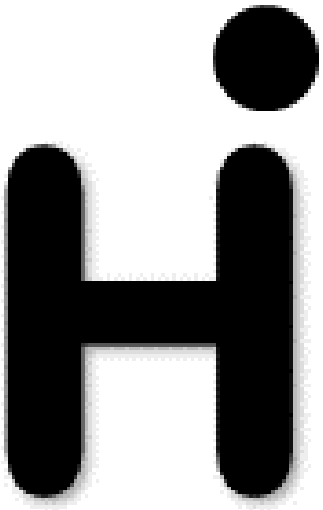
Application form

Gravitation Programme
2018 – 2019

Hybrid Intelligence (HI): augmenting human intellect



This is the public version of the NWO proposal submitted in September 2018.
Personal information and financial information about individual institutions have been removed.

September 25, 2018

Content

1. General information

Scientific summary

Over the past decade, researchers in Artificial Intelligence (AI) have made ground-breaking progress on long-standing problems. Now that AI is becoming increasingly part of our daily lives, we need to avoid being ruled by machines and their decisions. *Hybrid Intelligence (HI)* is the combination of human and machine intelligence, expanding human intellect instead of replacing it. It takes human expertise and intentionality into account when making meaningful decisions and perform appropriate actions, together with ethical, legal and societal values. Our goal is to design Hybrid Intelligent systems, an approach to Artificial Intelligence that puts humans at the centre, changing the course of the ongoing AI revolution. By providing intelligent artificial collaborators that interact with people we amplify human capacity for learning, reasoning, decision making and problem solving.

The challenge is to build intelligent systems that augment and amplify rather than replace human intelligence, that leverage our strengths and compensate for our weaknesses. Such Hybrid Intelligence requires meaningful interaction between artificial intelligent agents and humans to negotiate and align goals, intentions and implications of actions. Developing HI needs fundamentally new solutions to core research problems in AI: current AI technology surpasses humans in many pattern recognition and machine learning tasks, but it falls short on general world knowledge, common sense, and the human capabilities of (i) collaboration, (ii) adaptivity, (iii) explanation and (iv) awareness of norms and values. These challenges will be addressed in four interconnected research lines:

Collaborative HI: How to design and build intelligent agents that work in synergy with humans, with awareness of each other's strengths and limitations? We will develop shared mental models for communication between humans and agents, computational theories of mind to enable collaboration, and exploit multimodal interaction for seamless dialogues.
Adaptive HI: The world in which Hybrid Intelligent systems will operate is dynamic, as are the teams of humans and agents that make up such HI systems. HI systems will thus need to operate in situations not anticipated by their designers, and cope with variable team configurations, preferences and roles. This will require progress in online reinforcement learning, auto ML, and the integration of learning and reasoning.
Explainable HI: Intelligent agents and humans need to be able to mutually explain to each other what is happening (shared awareness), what they want to achieve (shared goals), and what collaborative ways they see of achieving their goals (shared plans and strategies). The challenge is to generate appropriate explanations in different circumstances and for different purposes, even for systems whose internal representations are vastly different from human cognitive concepts. We will use causal models for shared representations, develop methods for contrastive, selective and interactive explanations, and combine symbolic and statistical representations.
Responsible HI: Addressing and mitigating some of the perceived risks of Artificial Intelligence technologies requires ethical and legal concerns to be an integral part of the design and operation of HI systems. Values such as transparency, accountability, trust, privacy and fairness can no longer be relegated to regulations that apply after system's deployment. We will develop methods to include ethical, legal and societal considerations into the design process ("ethics in design") and into the performance ("ethics by design") of HI systems.

Applications in healthcare, education and science will demonstrate the potential of Hybrid Intelligence: virtual agents and robots will help children with concentration problems to study better; virtual agents and robots will support children in paediatric oncology wards by providing them with entertainment and information during prolonged hospital stays; virtual agents will collaborate with scientists on large scale analysis of the literature, formulate new hypotheses and help design experiments to test them.
The team brings together top AI researchers from across the Netherlands in machine learning, knowledge representation, natural language understanding & generation, multi-agent systems, human collaboration, cognitive psychology, multimodal interaction, social robotics, AI & law and ethics of technology. We will initiate a Hybrid Intelligence Centre (HI Centre) to host joint research facilities, multidisciplinary PhD programs, and training and exchange programs. Sustainability of the HI Centre is ensured through tenure track positions with guaranteed long-term funding from the participating universities.

Summary of the research proposal in layman's terms

Hybrid Intelligence (HI) combines human and artificial intelligence. Six Dutch universities develop theories and methods for intelligent systems that cooperate with humans, that adapt to dynamic circumstances and that can explain their actions. Ethical and legal values, such as transparency, accountability and trust, are taken into account during the design of such HI systems. We demonstrate applications of HI systems in healthcare, education and science to show the potential of artificial intelligence to amplify human intelligence instead of replacing it.

Key words
Artificial Intelligence
Socially aware agents
Explainable AI
Responsible AI

Hybrid Intelligence

2. Research proposal

Hybrid Intelligence (HI) is the combination of human and machine intelligence in order to make meaningful decisions and perform adequate actions. Hybrid Intelligence requires the interaction between artificial intelligent agents and humans, taking human expertise and intentionality into account, together with ethical, legal and societal values. Our main research challenge is:

How to build adaptive intelligent systems that augment rather than replace human intelligence, that leverage our strengths and compensate for our weaknesses?

Addressing this challenge requires a multidisciplinary research effort, that combines AI research with insights from social psychology, multimodal interaction, and ethics of technology. To facilitate this research, we will create a Hybrid Intelligence Centre (HI Centre).

The HI Centre

What? A collaboration of top AI researchers from across the Netherlands in areas such as machine learning, knowledge representation, natural language understanding & generation and multi-agent systems, collaboration, cognitive psychology, multimodal interaction, social robotics, AI & law and ethics of technology.

Why? The HI centre will create a national and international focus point for research on all aspects of Hybrid Intelligent systems. By creating intelligent machines that interact with humans, we aim to give people new, intelligent artificial collaborators for joint reasoning in order to optimize decision making and problem solving. This interaction has the potential to amplify both human and machine intelligence by combining their complementary strengths. Hybrid Intelligence (HI) focuses on the assistive and collaborative role of Artificial Intelligence, emphasizing its potential to enhance human intelligence instead of replacing it.

How? The HI-Team will achieve this through:

- Four interconnected research lines, each with focused questions led by a team of AI experts
- Attracting top international talent to the HI Centre
- Multidisciplinary PhD programs (AI, psychology, interaction, ethics)
- Training and exchange programs
- International collaboration with related programmes (EU Flagship, CLAIRE, UK HLC programme)
- Collaborative research and engineering facilities
- Joint dissemination and promotion activities
- Sustainable embedding of the programme in each of the participating universities through cash co-funding for tenure track and assistant professor positions in HI.

Where? The organizational core is based at the VU in Amsterdam, with dedicated facilities; the HI Centre operates as a virtual research centre, in a unique collaboration between 6 Dutch Universities, comprising major AI research lines in Amsterdam (VU & UvA), Delft, Groningen, Leiden and Utrecht.

The result: The HI Centre will be a sustainable centre of excellence where researchers can meet, collaborate and use the shared laboratories and infrastructure, with a lasting impact on the research programmes of each of the participating universities. Sustainability of the HI Centre is ensured through tenure track positions with guaranteed long-term funding from the participating universities.

Hybrid Intelligence:
augmenting human intellect

"Augmenting human intellect would warrant full pursuit by any enlightened society"

Doug Engelbart, 1962.

How to build adaptive intelligent systems that augment rather than replace human intelligence, leverage our strengths and compensate for our weaknesses?

Over the course of history, human civilisations, cultures and economies have scaled up through the use of tools: 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) has 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 typically passive, and mostly used by experts. We propose that Hybrid Intelligence should go well beyond this: we will study and develop HI 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 tool to augment human intellect, as well as to our ability to apply intelligent systems in areas of crucial importance to society.

Societal relevance and urgency

Contemporary societies face problems that have a weight and scale novel to humanity, such as maintaining democratic institutions, nuclear proliferation, 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 (Plous, 1993; de Martino et al., 2006; Efferson et al., 2008). We need new methods for deliberating these challenges and ways to consolidate opinions. 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 (Angwin, 2016; Flores, 2016). Instead, we need cooperative problem solving in which machines and humans contribute through a collaborative conversation, where machines engage with us, explain their reasoning, and learn from their mistakes.

> "The guiding principle of all AI-related research will be the development of responsible AI, putting the human at the centre".
>
> Communication from the European Commission on Artificial Intelligence for Europe

AI will either empower our ability to make more informed choices or reduce human autonomy; expand the human experience or replace it; create new forms of human activity or make existing jobs redundant; expand democracy in our societies or put it in danger. This is not a problem for the distant future: AI systems are deployed right now that do not take into account social values such as fairness, accountability, and transparency at the centre. This leads to today's problems of fake news at the top of our news feed, Facebook messages leading to ethnic and religious violence, and trolls influencing elections. This lack of alignment with human values is impacting us now.

Autonomous AI systems tend to be "idiot savants": world-class in a very narrow range. Humans are needed to deal with the borderline cases that none of the artificial systems is competent to handle. There is a danger that users 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.

Given the huge scientific challenges and the urgency to meet societal needs, now is the time to capitalise on the rapid advances in AI over the past decade and to influence the development of the field. By investing in the considerable AI expertise available in the Netherlands, and in our innovative approach of HI, the Netherlands will be able to match the investments made in AI in other countries[1] and to co-determine the rapid developments. The Netherlands needs to advance its scientific knowledge in this area to develop technology conforming to the ethics and values of our society and to avoid becoming dependent on other countries and companies. We will need to enhance the power of machine learning with the strength of human reasoning and the precision of automated reasoning. Economically, our results will allow organizations to innovate faster and more creatively, using understandable and trustable systems.

Designing Hybrid Intelligence

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 simply replacing them. Thought leaders in AI increasingly share the conviction that in order for AI systems to help humans and humanity, we need a new understanding of AI that takes humans and humanity explicitly into account (Kambhampati, 2018). They argue that it is better to view AI systems not as "thinking machines," but as cognitive prostheses that can help humans think better (Guszcza, 2018).

We aim to design and build agents that work in synergy with humans. Such synergy is productive if we can leverage the complementary strengths and weaknesses of humans and machines, see Figure 3. Humans excel in *collaboration*; we flexibly *adapt* to changing circumstances during executing of a task; an essential element in our collaboration is the capability to *explain* motivations, actions and results; and we always operate in a setting where *norms and values* (often implicitly) delineate which goals and actions are desirable or even permissible. Current AI technology surpasses humans in many pattern recognition and machine learning tasks, but it falls short on general world knowledge, common sense, and the human capabilities of collaboration, adaptivity, explanation and awareness of norms and values. We will address these challenges in four interconnected research lines of the HI Centre: Collaborative HI, Adaptive HI, Explainable HI and Responsible HI.

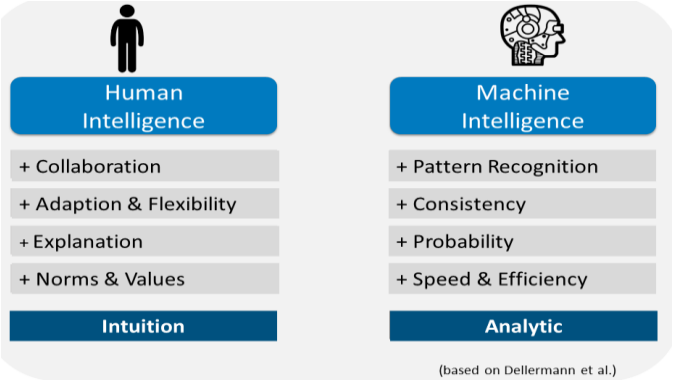


Figure 3: Strengths of human and machine intelligence

Scenarios for Hybrid Intelligence

Figure 4 describes 3 illustrative scenarios for HI. Simple versions of all of these scenarios can already be done with current state of the art, and indeed members of our team have built demonstrators for each of them. However, essential elements are missing and require new breakthroughs (see fig. 3): collaborating in teams of different sizes and expertise, flexibly adapting to changing circumstances, knowing what other team members know and expect, providing explanations at the appropriate level of detail, understanding social norms and values.

[1] Canada is investing $125M with 3 new institutes and 50 new professorial chairs; in Germany and China the expected level of investment is above $500M; South-Korea is investing $800M over 5 years; the UK is investing £300m in AI and funding 1000 PhD positions <https://goo.gl/T8go3y>; France has pledged over €1000M <https://goo.gl/xSzTsv>; Sweden is investing €100M on AI <http://wasp-sweden.org/ai/>.

Figure 4: Illustrative scenarios for HI

Core scientific challenges

We will address these challenges in four interconnected research lines of the HI Centre: Collaborative HI, Adaptive HI, Explainable HI and Responsible HI (see Figure 5).

Collaborative HI: How to build intelligent systems with a computational theory of mind and collaboration models, through which they (a) can reason about the situated cognitive capabilities, beliefs, desires and intentions of other agents, and (b) can initiate and join collaborative activities. How can agents communicate effectively for this purpose using both natural language and by exploiting physical embodiment using appropriate physical clues and signals, such as gestures, facial expressions and voice intonation, both by recognising them, and by producing them?

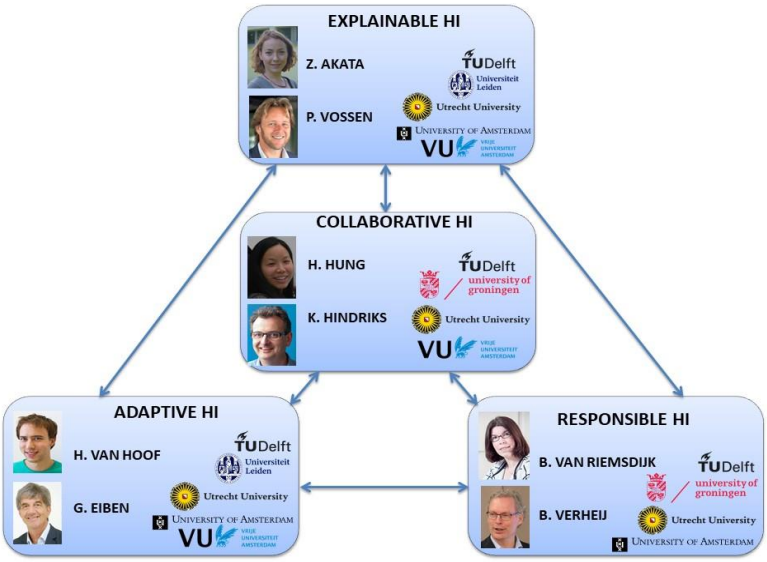
Our central questions are:

What are hybrid intelligent systems? How can human insights be combined with AI to come to better decisions, and how to characterize the tasks for which this is true?

How do we build HI systems? How to formulate a design theory for HI systems, telling us which technical components, which representation formalisms and which computational methods should be combined in which collaboration architecture to achieve a desired functionality?

How to build them responsibly? How to ensure that the behaviour of artificial agents in Hybrid Intelligent systems is aligned with ethical values and social norms?

Figure 5:Interconnected research lines of HI Centre



Adaptive HI: How to build intelligent agents that are aware of the context in which they operate, allowing them to adapt to changes in that context? And how to reconcile such behavioural adaptation with requirements on safety, transparency and predictability? How to integrate the adaptivity of machine learning techniques with the precision and interpretability of symbolic knowledge representation and reasoning?

Explainable HI: What constitutes an appropriate explanation in different circumstances and for different purposes, and how to generate such appropriate explanations even for systems whose internal representations are vastly different from human cognitive concepts?

Responsible HI: How to ensure that the behaviour of artificial agents in HI systems is aligned with ethical values social norms and legal constraints? And how to guarantee that the way we build HI systems is in line with values, social norms and legal constraints?

What is not in Scope?

Our goals are to understand HI, to learn how to build HI systems, and how to build and use them responsibly. Even with these ambitious goals, there are topics which are outside the boundaries of our programme.

- *Interface technology*: We will use state-of-the-art software and hardware, both academic and commercial, for interface technology, embedded systems and Internet of Things. We expect to make extensive use of state-of-the-art software for speech recognition and generation, posture and gesture recognition and communication by avatars.
- *Robot platforms*: We will use robots for experiments with embodied communication using state of the art commercial robot platforms.
- *Cognition*: A thorough understanding of aspects of human cognition is crucial for the design of HI systems (e.g. theories on human attention, multitasking, perception etc), and we will make use of state-of-the-art scientific theories and insights on these topics. Expert knowledge on these is available in the consortium (Neerincx, Balliet).

Tracking progress

Each of the research lines defines measurable functional capabilities of HI systems that are used to track our progress. Together, these capabilities define a design space of HI systems. We will follow a spiral approach, where we start studying and building HI systems in simple regions of this design space, and gradually move to investigating more sophisticated regions of the design space of HI systems. We will *track our progress* in three different ways. First, we will perform a regular series of **controlled experiments** that are based on experimental designs from psychology for studying human collaboration patterns by members of our consortium (Romano et al., 2017; Arslan et al., 2017). These experiments test human behaviour for achieving shared or competing goals under a variety of reward and penalty schemes.

Secondly, we will test the collaboration and effectiveness of hybrid human-AI systems in **real-world domains** where our team members already have a track record: healthcare (Boumans et al., 2018), disaster response (de Greeff et al., 2018), negotiation (Jonker et al., 2017) education (Xu et al., 2014) and information retrieval (Renet al., 2017; ter Hoeve et al., 2017). Finally, we will test the performance of HI systems in realistic settings through **experiments**

with external partners, where we provide resources to external parties to deploy the insights and technologies from the consortium in areas such as healthcare, policy making, computer security and education.

The next sections provide for each research line an overview and motivation, a discussion of the state of the art, the research questions and corresponding research activities, and the expected results of these activities, including a matrix that identifies the improved capabilities of HI systems (Figure 6).

2.2.1 Research line: Collaborative HI

Aim: To develop HI systems that collaborate effectively with humans by: understanding social contexts, norms and practices; recognizing and reasoning about the abilities, beliefs and goals of other agents through a computational theory of mind; communicating appropriately, including via the use of physical embodiment for cues and signals.
Coordinators: Hindriks & Hung
Core participants: Hung, Neerincx (Delft); Verbrugge (RUG); F Dignum (Utrecht); Balliet, Vossen, Hindriks (VU)

Overview and motivation

Collaboration in human teams is vital, pooling different skills to solve more difficult problems than any of the members could alone (Hong & Page, 2001). 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. Using insights from social and psychological sciences, we aim to enhance human cognitive, perceptual and decision-making skills by combining them with those of AI systems. Early results in successful complementary human-machine collaboration in cognitive tasks are known from negotiation tasks (Hindriks et al., 2008; Bosse & Jonker, 2005), planning (Sycara et al., 2010), behaviour change support systems (Schouten et al., 2017; Shamekhi et al., 2017) and in `centaur' chess (Kasparov, 2010), although dysfunctional behaviour is also seen (van Wissen et al., 2012). Promoting machines from tools to partners faces two key challenges: a computational understanding of the social context of collaborative tasks, and a computational understanding of actors engaging in such tasks.

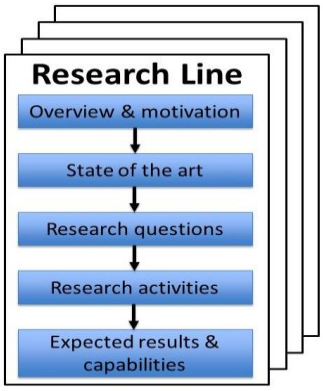


Figure 6: Elements of research line description

Understanding the social context. How can machines adapt to the ways humans typically collaborate? How can machines amplify human capacity to collaborate effectively? We know that human collaboration follows certain patterns, social practices, and strategies such as kinship, direct reciprocity, indirect reciprocity, and ingroup favouritism. These behaviours are based on their values, intentions, and beliefs, are shaped by previous experiences, their social context and their current mental state. Can we build machines that can learn what these are, adapt over time via social learning (similar to humans)? Could machines become part of groups and have a positive influence on people's willingness and competence to cooperate, and so producing better outcomes for the group as a whole? Can machines learn to recognise social practices and collaborative strategies of humans? Humans build relationships and make agreements, sometimes implicitly, to address task dependencies adequately (e.g., for coordination). Can machines contribute to the learning, setting and execution of work agreements?

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. Here, we can build on the substantial research on how to mitigate cognitive biases (e.g., Cook & Smallman, 2008; Kayhan, 2013; Rothmund, Mojzisch & Schulz-Hardt, 2011). We will focus on developing new models to make machines aware of which biases apply to certain tasks and the use of machine learning techniques for detecting bias in collaboration.

Awareness is key for such collaboration. We will use the extensive research on (a) Theory of Mind (ToM) in developmental and cognitive psychology, and (b) Joint Cognitive Systems (JCS) in social-work psychology and human factors as models for mutual (self-other) understanding and task harmonization as our starting point for developing a computational theory of hybrid collaboration. Key concepts in ToM and JCS are shared in the work on cognitive agents in AI (e.g., beliefs, desires, goals, intentions, team coherence, shared situation awareness, workload, trust and agreements), and are starting points for a computational theory (Jennings et al., 2014, Harbers & Neerincx, 2017). The key question that we will address is how to formulate and operationalize a computational theory that supports collaboration in hybrid teams.

State of the art

Theory of mind: Maintaining the beliefs, goals and other mental attitudes of other people in a theory of mind 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 et al., 1997; Jonker, van Riemsdijk, et al., 2010), present meta-reasoning techniques allowing an agent to recursively apply a ToM to detect anomalies in its state of mind. In de Weerd et al., (2013, 2017, 2018) we have shown 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. Previous studies on three components of the experience-sharing model show the potential to accommodate the social, cognitive and affective processes in teamwork: the reciprocal child-robot self-disclosure module of Burger et al. (2016) to enhance bonding and information sharing, the psychologically-inspired Artificial Empathic Memory architecture for structuring and interpreting user experiences (Dudzik et al., 2018), and the Cognitive Affective Agent Framework (CAAF) to enhance goal-setting and goal-adherence (Burger et al., 2016; Kaptein et al., 2017).

Teamwork, joint actions, plans and tasks: In Multi-Agent Systems (MAS) research, substantial work has been performed on distributing tasks and monitoring plan progression in MAS (Dunin-Keplicz et al., 2010). Much used systems such as TAEMS only consider software agent teams and 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. (Harbers et al., 2012; 2014). Our 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 (Sinha et al., 2018), 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 (Balliet et al., 2011, 2013), indirect reciprocity (Wu et al., 2016), 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 (Balliet et al., 2017; Gerpott et al., 2018). Computational theories of reciprocity show that the effect of reciprocity has similar effects on artificial agents (Ranjbar-Sahraei et al., 2014; Polevoy et al., 2016). 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. Fortunately, 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 (Johnson et al., 2014). A critical issue here is intercultural collaborations. Some preliminary

work on culture-aware agent systems (Hofstede, Jonker & Verwaart, 2012) and research on game theory from a multiagent perspective offers some starting points.

Multimodal interaction: There is a long tradition of research on multimodal communication (Partan & Marler 2005), human-computer interfacing (Card, 2017), and other component technologies such as facial expression analysis (Gunes et al., 2011; Burgoon et al., 2017), and gesture detection (Mitra & Acharya 2007) that shows the importance of multimodal interaction for collaboration (Rauschert et al., 2002). The same can be said about multimodal dialogue systems (Wahlster, 2006), 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 (Poria et al., 2017) and social signal processing (Burgoon et al., 2017) 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, in a HI setting, where multiple people may be interacting with multiple agents in a dynamic environment, we cannot expect to have good video footage for each person (the group size may be unknown beforehand). As discussed by Gunes and Hung (2016), moving from controlled laboratory studies to real life settings requires a fundamental change in experimental approach. As argued by Hung et al. (2018), in challenging the notion of understanding and ultimately influencing human social behaviour, 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:

1. What are appropriate models for negotiation, agreements, planning and delegation in hybrid teams? This challenge is addressed in Activities C1 and C4.
2. 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? (Activity C2, C5)
3. How can Hybrid Intelligence exploit experience sharing for the purpose of establishing common ground, resolving uncertainties and conflicts, adjusting tasks, goals and correcting actions? (C2 and C4)
4. Which specific challenges and advantages arise when groups of humans and agents collaborate, given the complementarities in their skills and capabilities? (Activity C4)
5. How to understand and generate multimodal messages, expressions, gestures, and semi- or unstructured representations for the purpose of collaboration? (Activities C1, C3)

Research activities

These research questions will be investigated in the following research activities.

C1. Human-AI Interactions: People use practices, norms, values, context etc. to cope with the complexity of human-human interactions. We will investigate how these concepts can be translated to the setting of human-AI interaction. Starting from models of these concepts taken from human interactions, we will investigate how they apply for such new type of interactions and how they can be adapted to changing contexts and for users with different skills, needs and characters.

C2. Computational theory of mind: Using cognitive agent technology (Dunin-Keplicz & Verbrugge, 2006; 2010) we will develop capabilities for modelling self and others, based on our work theory of mind and shared mental models

(Jonker et al., 2002; 2010), and on joint cognitive systems (Harbers and Neerincx, 2017). A cognitive agent architecture will be designed including mechanisms for storage of beliefs, based on communication and intention recognition techniques (Han et al., 2011, Vossen et al., 2018, Fokkens et al., 2017). The main contribution will be a computational higher-order intentional system enabling reasoning about individual and collective cognitive and motivational states in the hybrid team. Such theories of mind will be used in C4 and C5.

C3. Multimodal interaction dynamics: We will build machines that detect the behaviour of individuals in order to understand and respond to others' behaviour. This involves understanding and generating natural language within the context of the collaboration, the partners and the assumed goals, as well as sensing and modelling of behaviour for smooth coordination. For the language understanding and generation (English and Dutch), we start from dialogue and communication models and redesign these for the purpose of HI. We will take the ToM model as a basis for correcting errors, filling gaps, acquiring knowledge, resolving conflicts and negotiating about goals and their implications. We will build on our methods of understanding the timing of body language using social signal processing techniques. Specifically, we will investigate the role of gestures and speech for (i) estimating how effective an individual is in communicating to either an artificial agent or other human, (ii) modelling how conversational turn taking occurs, (iii) developing strategies for turn management such as interruption or turn moderation (Xu et al., 2015; Xu, Broekens, Hindriks et al., 2015). We will also investigate the generation of expressive multimodal interaction for agents embodied as robots. Whereas C1 is about socio-cognitive aspects of interactions, C3 focuses on observable interactive behaviour.

C4. Managing group dynamics: In hybrid teams, agents must understand the group dynamics in order to influence coordination and collaboration. Machines must form a mental model of a group in order to influence and adapt collaboration patterns. This adaptivity benefits from machine learning techniques such as AutoML techniques, are elaborated in the Adaptive HL research line. Moreover, to support team formation, we will build machines that can select and initiate collaborative relationship. We will use a machine perception approach combined with social practice theories to understand how alliances in groups are initiated, maintained, and evolve over time. We will compare hybrid teams with human teams for group cohesion and team performance such as creativity, efficiency, and productivity. We will apply analysis, modelling, and synthesis perspectives to understand how group dynamics in hybrid teams change and evolve over time (Dunin-Keplicz & Verbrugge, 2003; Grosz & Hunsberger, 2006; Dunin-Kęplicz, et al., 2011). This work on group dynamics is closely linked to the work on interaction dynamics in C3.

C5. Effective and scalable computational theory of mind: Using a ToM during deliberation is necessary but also costly, especially in large teams when several orders of ToM about all team members are being kept. Whereas in C2 we will investigate some fundamental properties of the use of ToM, in this activity we will investigate the practical computational aspects. What level of detail is necessary when modelling the minds of other team members? How does this level of detail depend on role and task in the team? How we can we switch between detailed and abstract models of ToM, based on what we can deduce from the ToM? Similar questions pertain to the maintenance of a ToM. When are elements of a ToM updated, and with which frequency? A final aspect to be investigated is the required depth of nesting of ToM models, and how to anticipate the fact that other team members will also use a ToM to anticipate the actions of the system.

Connections with other research lines:

	Explainable	Adaptive	Responsible
C1			
C2			
C3			
C4			
C5			

Activity C1 produces models to be used when generating explanations (Explainable HI), uses techniques from Adaptive HI, and must of course follow the ethical guidelines of Responsible HI. The work on theory of mind (C2) is necessary for generating tailored explanations. The interaction dynamics from C3 are feeding into the research line on Explainable HI and may use results from Adaptive HI. Because of the social nature of the interactions from C4, insights from the work on Responsible HI are highly relevant in C4. C5 interacts with all other research lines as the level of detail of a ToM is essential for both explanations, adaptivity and responsible behaviour.

Expected results

This research line will result in theory, methods and tools for the following:

- Perceiving, modelling, and managing the *internal state of an individual*, including the building blocks for a computational theory of mind.
- Perceiving, modelling, and managing *Interpersonal Relationships at a dyadic level*, i.e. human-human or human-agent.
- Perceiving, modelling, and managing *internal state of the group and intra-group relationships*, e.g., cohesion, delegation, willingness to collaborate and short-term goals.
- Perceiving, modelling, and managing how collaborations are initiated and maintained in an HI system
- Understanding and generating expressions, multimodal messages, and semi- and unstructured data for the purpose of collaborating with humans.

The activities lead to agents that will have the following core capabilities, with measurable dimensions of progress:

Core capabilities	Level 1	Level 2	Level 3
Initiating relationships	Form a goal	Select a partner or group appropriate for a goal	Initiate relationships and group formation appropriate for a goal
Establishing shared situational awareness	Individual knowledge of situation	Shared knowledge of a situation	Common knowledge of a situation
Personalised multi-modal user interaction	Perception of social cues to infer partner characteristics)	Management of social signals to communicate to partner to coordinate effectively	Collaborative strategies based on long term memory of group experiences
Collaborative group support	Communicate and respond to other's needs	Identify inefficiencies and better solutions to effective coordination and cooperation	Strategies for managing conflict, power asymmetries, hierarchy

2.2.2 Research line: Adaptive HI

Aim: To make the system able to adjust to previously unknown situations, changing circumstances, new user requirements and goals, and variable team configurations. To reconcile such behavioural adaptation with requirements on safety, transparency and predictability.
Coordinators: Eiben & van Hoof
Core participants: Oliehoek (Delft); Hoos (Leiden); van der Gaag (Utrecht); Welling, van Hoof (UvA); van Harmelen, Eiben, Roijers (VU).

Overview and motivation

In HI settings, artificial agents and human agents work together, possibly on multiple tasks, in complex environments. Such settings are seldomly static: team composition and tasks can change, preferences can shift and external conditions (e.g., available resources, and environment) can vary over time. Thus, the competences in terms of theory of mind, collaboration, task execution, et cetera cannot be fixed before deployment. The agents will have to adapt and learn in operation. Moreover, members of the team get to know each other better over time. As such, the ability of HI systems to adapt or learn is a prerequisite not only for them to perform well, but for them 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).

The temporal aspects of adaptivity are highly important in this respect: on one end of the spectrum, learning only takes places in an ('offline') *design stage*. This typically requires the system designer to collect a dataset from which a model is learnt, which is then merely used for prediction during the ('online') *operational stage* (Haasdijk et al., 2013). At the other end of the spectrum, we see systems that, while in operation, continually adapt their models. In between these extremes, there is the `versioned' approach, which operates in update cycles (e.g., if version 1 of the

system becomes inadequate, then the system is adjusted and version 2 is launched). This proposal aims at systems in which the agents learn online and combine all forms of adaptivity: online learning from direct and indirect feedback from its environment (including dialogues with other agents and humans) to adapt its rules, knowledge, formalism, and even its learning strategies.

Well established forms of adaptivity are rule-based, where the agent has rules that govern how to act in different circumstances, and feedback-based, in which the agent adapts its decision rules over time based on (explicit or implicit) feedback on its behaviour. The ambition of this proposal is of the next level: over time, the agent also learns new formalisms with higher expressive power, and new strategies and techniques for learning, and meta-strategies to steer the learning and select appropriate mechanisms. This will enable agents to adapt to the highly changeable situations that result from the HI setting, in which at any given time the environment, tasks, resources and the team of human and artificial agents may change. To be effective in such a dynamic setting, agents will have to learn new concepts and competences, and communicate with human agents to adapt quickly and effectively. For this communication, systems will also need to be aware of the variety of the human agents it collaborates with and adapt feedback elicitation strategies accordingly (linking to theory of mind).

There is an inherent tension between the adaptive nature of the systems we investigate and the desire for safety and reliability. Constraints on the adaptivity of the 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 ethics and social acceptability. Such constraints may be encoded in the reward/loss functions of the learning system, or may be symbolically encoded, or may be implemented through modification of the adaptive exploration process (Garcia & Fernandez, 2015). Highly adaptive systems also pose a challenge for transparency and explainability of a system's actions. Data, settings, concepts and competences all interact in the decision-making process. The system's architecture thus needs to be able to keep track of all these changes in order to be able to backtrack why a specific decision was taken at a specific point in time. If data, settings, concepts and competences constantly change and evolve dynamically, possibly at different rates, capturing the required information for explaining a decision becomes a highly complex task requiring technologies going far beyond standard version control and general provenance modelling. Furthermore, these systems must not only keep track of such information, but also be able to effectively communicate this information with a variety of users, in order to receive necessary feedback.

State of the art

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 (Caruana, 1997). Changes in the environment are targeted, e.g., by *domain adaptation* (Daume & Marcu, 2006). More generally, *transfer learning* approaches try to adapt learned models from source tasks to target tasks that could differ in either environment or objective (Pan & Yang, 2010). A growing body of work has also studied the use of *meta-learning* for rapid adaptation (Vilalta & Drissi, 2002). 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 (Roijers et al., 2013), which model different reward functions for different desirable features of a solution. Because users may have different preferences over these objectives, multi-objective models allow the user to be modelled separately from the environment model (Roijers et al., 2017). Recently, so-called automated ML (or *AutoML*) methods have been developed in order to select and optimize learning algorithms for specific tasks or data sets.

None of the mentioned models fully account for learning new representations or new learning strategies, nor do they 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.

Within the consortium, various aspects and sub-problems of the challenge of Adaptive HI have already been addressed. For example, to handle user preferences that change over time (Roijers et al., 2013, 2017), different preference elicitation strategies have been compared (Zintgraf et al., 2018), and multi-objective optimization has been used to adapt an information retrieval system to the current user preferences (van Doorn et al., 2016).

Incomplete knowledge about the preferences of negotiation parties has also been used to inform multi-attribute negotiation systems (Aydogan et al., 2017; Baarslag et al., 2017; Jonker et al., 2018).

The sub-problem of adaptivity to changes in the environment has also led to preliminary results within the consortium. Robot controllers that are adapted depending on the environmental conditions were compared and developed (Barbaros et al., 2018; Heinerman et al., 2017). Furthermore, the morphology of robots can be adapted to the environment as well (Eiben & Smith, 2015). We have co-developed a co-active design method for hybrid teams of humans and robots (Johnson, 2014).

On a higher level, the adaptivity of the learning mechanism itself has been investigated by consortium members. For example, the representation of knowledge in the AI system can depend on the context (e.g., objectives) of the system (Idrissou et al., 2017). Furthermore, fully automated procedures have been developed for selecting and configuring algorithms for a given supervised machine learning task (Thornton et al., 2013; Kotthoff et al., 2017), and are rapidly gaining traction.

Research questions

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

1. 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? Addressed in activities A1 and A2.
2. How can learning systems respect societal, legal, ethical, safety, and resource constraints that might be expressed symbolically? Activity A4.
3. How can learning systems accommodate changes in user preferences, environments, tasks, and available resources without having to completely re-learn each time something changes? Mostly activity A3, but supported by activities A1, and A2.
4. How can the learning mechanism itself be adapted to improve efficiency and effectivity in highly dynamic Hybrid Intelligence settings, based on task experience as well as human guidance? Activities A1, A3, supported by A4.
5. How to integrate the adaptivity of machine learning techniques with the precision and interpretability of symbolic knowledge representation and reasoning? Activity A4.

Research Activities

These research questions will be investigated in the following research activities.

A1. Learning through interaction: Agent in hybrid systems will need to learn efficiently from interaction with others (humans and/or machines), and will have to be able to interpret complex feedback as part of these interactions. We will investigate special cases of competitive learning (as in Generative Adversarial Networks) (Goodfellow et al., 2014), learning from demonstration (where examples of good actions are given), coaching - where a mentor gives information at a more abstract level (e.g., indicating areas for improvement), and collaborative learning (learners learning in parallel, sharing information). In these, the agents will use different techniques for communication from humans to learning agents such as reward shaping (Ng et al., 1999), human rewards (Ng & Russell, 2000) and explicit preference information, possibly in symbolic form (A4).

A2. Learning how to interact: In order to learn through interaction (A1), agents need to learn how to interact. To communicate effectively with humans, it is essential that an agent learns about individuals (e.g. learning to improve a theory of mind), and about how interaction takes place in (possibly mixed) groups. Agents need to learn when and how to communicate, using shared representation formalisms, their intentions and information, especially if they have knowledge their partner does not possess, and without overloading human collaboration partners. Symbolic knowledge might be an intermediary representation for such dialogue plans (A4). We will adopt user-centric methods to identify effective communication methods using a variety of modalities for a variety of users.

A3. Incremental learning in changing conditions. In Adaptive HI tasks, the composition of collaborative groups, the tasks, the environment, user preference, and the resources may change. When incorporating information and feedback from humans, we need the adaptation to happen sufficiently fast. This requires that we keep learning

tractable, for example by avoiding that any change in one or more aspects of the environment does not require a retraining of the entire system. To address this, we will use and improve on AutoML, multi-objective reinforcement learning, and meta-learning techniques to gradually adapt the learning mechanism and the model/controller representation to most efficiently deal with such changes. This capability is essential for interactive learning tasks (A1, A2), as well as for learning how to adapt explanations to changing circumstances (the research line on Explainable HI).

A4. Integrating learning, reasoning and planning. We aim to combine machine learning techniques with symbolic knowledge representation, reasoning and planning, which are easier to interpret and specify by human partners. We will develop a system architecture that allows learning systems to benefit from symbolic knowledge as priors or constraints. We expect such mixed systems to be more robust to noise, to have better learning rates, to have a higher learning transferability, and to be more explainable and predictable (Wilcke et al., 2017). We will investigate and develop techniques for specifying and accommodating societal, ethical, legal, and safety constraints which may be expressed in symbolic form or through dialogues (A1). Such constraints may also specify which amount of risk-taking during exploration is acceptable while still allowing for sufficient adaptivity.

The idea is to couple machine learning techniques with symbolic approaches for knowledge representation, planning and search, such as Bayesian nets, argumentation systems, ontologies, and other representational forms. This coupling allows us to use automated reasoning to detect undesirable behaviour, inconsistencies and gaps in the knowledge, use that to steer the machine learning approaches to change behaviour or gather more information to tackle inconsistencies and gaps, and to provide feedback to the machine learning systems. We will apply methods such as (Eggensperger et al., 2018).

	Collaborative	Explainable	Responsible
A1			
A2			
A3			
A4			

Connections with other research lines: There are tight connections from the research line on Collaborative HI to activities A1 ("Learning through interaction") and A2 ("Learning how to interact"). A1 will produce interaction capabilities that will be used in the research line on Explainable HI. Some of the changing conditions to be accounted for in A3 will come from the Collaborative HI research line. The "integration of learning, reasoning and planning" (A4) will yield representations and algorithms that are useful in all other research lines.

Expected results

These research activities will improve our understanding of how mixed groups of agents (human and machine) can learn through social interaction, e.g., by asking for and considering complex feedback. This encompasses machines learning from humans, machines learning from machines and humans learning from machines. We will develop a theory of managing trial-and-error learning processes with the intention of excluding impermissible trials, e.g., due to safety, ethical, or legal constraints. We will develop incremental algorithms that enable learning systems to handle any combination of users, tasks, environments, or resources to change without triggering a full re-learning of all capabilities. All of these will benefit from innovative methods to integrate learning and reasoning, in particular having the capability that the learning system can interpret and respect symbolic constraints and map abstract symbolic knowledge to new sub symbolic representations of situations: two-way-learning from sub symbolic to symbolic models and from symbolic-to-sub symbolic models. We will develop next-generation AutoML techniques that will critically enable progress on Hybrid Intelligence and beyond.

The activities lead to agents with the following core capabilities with measurable dimensions of progress:

Core capabilities	Level 1	Level 2	Level 3
Learning through interaction	Ask humans for suitable and sufficient feedback	Rich interaction with one human partner	Rich interaction with multiple AI and human partners
Learning how to interact	Predict what human partner knows, wants and needs (ToM)	Anticipate how human reacts in human-agent collaboration	Anticipate how humans react in larger, mixed teams
Incremental adaptivity	Detect uncertainty of performance due to changing situations, goals and preferences; take suitable action	Predict likely upcoming situations, switching between pre-learned models	Online learning for predicted and surprising changes in situations, goals, and preferences
Integrate symbolic constraints during learning	Identify and ask for meaning of new symbols	Learn rich meaning of symbols	Shared meaning and use of symbols

2.2.3 Research line: Explainable HI

Aim: methods to generate appropriate explanation in different circumstances and for different purposes, even for systems whose internal representations are vastly different from human cognitive concepts. Methodology to evaluate explanations in a scalable and reproducible manner.
Coordinators: Vossen & Akata
Core participants: Jonker (Leiden/Delft); van der Gaag (Utrecht); Akata, Moncz (UvA); Vossen (VU)

Overview and motivation

In Hybrid Intelligent systems, human and artificial agents work together. 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 (Heider, 1958). Our hypothesis is that by building more transparent, interpretable, or explainable artificial agents, human agents will be better equipped to understand, trust and work with intelligent agents (Mercado et al., 2016; Hayes et al., 2017). We will use models of how humans explain decisions and behaviour to design and implement intelligent agents that provide explanations (Miller, 2017), including how people employ biases (Kahneman, 2011) and social expectations (Hilton, 1990) when they generate and evaluate an explanation. De Graaf and Malle (2017) argue that antromophization 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 (Mercado et al., 2016; Stubbs et al., 2007), 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 (Hilton, 1990). How can such contrastive explanations be generated, even in the absence of an explicitly mentioned counterfactual alternative? 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 (Trabasso & Bartolone, 2003). How can we capture and use such biases in automatically selecting causes? 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 explainer's beliefs (Antaki & Leudar, 1992). How can we generate interactive explanations, both for symbolic approaches and for less interpretable approaches such as neural networks?

To be able to have artificial agents generate explanations that are contrastive, selective, and social, we will pursue a number of promising solution directions. First, we will study scenario-based interactions in which human and artificial agents will refer to shared representations in which different selections of causes can be made. Second, we will consider a generative scenario (with a black box algorithm) and study different ways of identifying and describing

errors in machine learned translation models and producing contrastive explanations. Third, in a multi-modal setting we will study different ways of creating selective explanations, depending on the type of user (developers or end users). Fourth, the socio-situational nature of explanations will be studied by building different types of mental models of human and artificial agents that depend on personal traits of the human agents, and use the models to create selective explanations that fit the needs of the human agents given the current situation.

Finally, we seek to develop evaluation methodologies for explanations that assess the effectiveness of machine generated explanations and that allow us to assess the degree to which explanations are contrastive, selective, social, and situational; the big challenge is to create test beds that are repeatable and metrics that are intuitive.

State of the art

AI has a long history of work on explanation (Biran & Cotton, 2017). 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 1970's. Swartout (1983) already stressed 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 2000's show that users are much more satisfied with systems that contain some form of justification (Sinha & Swearingen, 2002). Bilgic and Mooney (2005) show that feature-based justifications are superior to neighbor- and user-history-based ones.

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 (Karpathy et al., 2015). Another line of work on explainability in ML 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. Previously, many studies that focus on the explainability of machine learning algorithms have been conducted from a Human Computer Interaction angle (e.g., Bilgic & Mooney, 2005; Herlocker et al., 2000). 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. Sharchilev et al., (2018) extend their work to tree-ensemble based methods. And concerning the third, Ribeiro et al., (2016) introduce LIME, a method to locally explain the classifications of any classifier. Three important characteristics underlie the construction of LIME: an explaining model needs to be (1) "interpretable," (2) "locally faithful," and (3) "model-agnostic".

Research questions

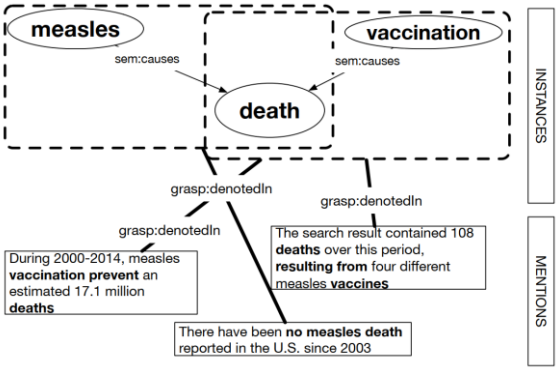
The above state of the art leads to the following research questions for explainability in hybrid systems:

1. How to build shared representations to be used as the basis for explanations, covering both the external world and the internal problem-solving process? Addressed in activities E1 and E3.
2. What are the different types of explanations that make the decision-making process more transparent and understandable? Activity E3.
3. How can explanations be communicated to users such that the explanations improve the user's trust and leads to a successful agent-user collaboration? Activities E2, E3 and E4.

4. How to personalise explanations that align with the users' needs and capabilities? Activity E4.
5. How to evaluate the quality and strength of the explanations? Activity E5.

Research Activities

These research questions will be investigated in the following research activities. The initial focus of the activities will be on the development of contrastive, selective, social and situational explanations of the decisions of artificial agents. In a later stage of the program we will increasingly emphasize *mixed-initiative* scenarios, where artificial agents may ask human agents for clarifications and explanations. This stage is more ambitious because we expect agents to have more difficulty understanding humans than the other way around.

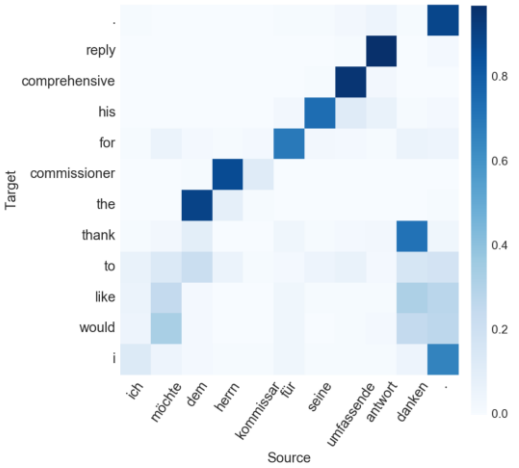


E1. Shared representations: To create contrastive explanations, we will pursue both model-driven and data-driven approaches. In a model-driven scenario, we assume that human and artificial agents are able to refer to shared representations (E4), covering both the concepts used to describe the external world, as well as a shared vocabulary to rationalize the (internal) steps and operations of the problem-solving process (E3). Our Circumstantial Event Ontology (CEO) (Segers et al., 2018; Vossen et al., 2018) can be used to recognize a variety of situations and events, by connecting events through (weak) causal and explanatory relations. Modelling weak and strong causality through CEO enables artificial agents to explain its motivation, speculate on the motivation of the human agent in relation to any actual situation or judge the risks of situations and take preventive actions.

In (Van der Waa et al., 2018; Neerincx et al., 2018) we have proposed an approach to explanation that creates *contrastive targeted explanations* by answering the question "Why this output instead of that output?". We will study how CEO-based explanations can be combined with such data-driven contrastive explanation as well as focus on understanding what information is captured in our knowledge models and how neural networks learn from the input they receive. In recent work on the Grounded Representation and Source Perspective (GRaSP) model (Fokkens et al., 2017) we proposed a provenance representation that allows human and artificial agents to trace back the entire process of coming to a certain interpretation. Our team has shown that for knowledge-based agents, efficient omniscient debugging is possible (Koeman et al., 2017a; 2017b), which means that artificial agents can efficiently go backwards in time to inspect why they came to certain conclusions, locate possible errors in the design of artificial agents, or to pinpoint the source of the faulty information they based their conclusions on.

E2. Interactive explanation generation: In interactive explanation generation human agents can ask not just for explanations but also pose scenario-based questions, such as `what if'-scenarios? By how much could the input be modified without the final result changing (scenario robustness)? What would, be the best option if the proposed option is not possible due to circumstances not known to the system (alternative scenarios)?

Communication among agents in HI systems itself is subject to explanations. Our earlier research on multi-source translation showed that the richness of the encoded information has a direct impact on the language generation quality (Garmash & Monz, 2016). In earlier work, we made a first step in tracing mistakes made by the translation models, by analysing the attention distributions used within a neural machine translation system to visualize how a translation is linked to the source sentence (Ghader & Monz, 2017). Although this approach allows one to identify the source of a mistake, it only identifies the problematic words themselves, but does not explain why these words are problematic nor does it explain the types of potentially complex interactions between words that lead to certain decisions or errors. To address this problem, we will build on Sharchilev et al. (2018) and develop a new framework that generates natural language explanations that describe in more detail the steps that have led to the current prediction.



E3. Explanation through introspection and rationalization: In any knowledge-intensive collaboration between agents, the capacity to know what you don't know (reflection) is a crucial ingredient of a successful collaboration. Artificial agents should discuss consequences that are contradictory by themselves or that contradict the norms and values that are to bind them to acceptable behaviour. These functionalities can be established by two different types of explanation: introspective reasoning, possibly visual and modular, to help an expert user to determine what happened in decision process; and post-hoc rationalization, which are typically textual and aimed at the non-expert users, e.g., to calibrate trust in an algorithmic solution, but do not necessarily present a faithful description of the system's actual internal mechanisms. In all cases, such explanations should be selective, focusing on the most pertinent factors. The concrete scenario on which we will focus is human activity recognition and visual question answering tasks. Here, multimodal explanation models offer significant benefits over unimodal approaches. Existing approaches for deep visual recognition do not provide justifications: contemporary vision-language models describe image content but do not take class-discriminative image aspects which justify visual predictions into account. Hendricks et al. (2016) focus on the discriminating properties of the visible object, jointly predict a class label, and explain why the predicted label is appropriate for the image. To further build trust with users, we integrate visual explanations to the model in the form of bounding boxes (Hendricks et al., 2018) pointing to the evidence. Furthermore, a multimodal approach yields better textual justifications, and better localizes the evidence that supports the decision (Park et al., 2018). What textual or visual features do we select? Which modality? Recent work by the team has identified showcases in which visual explanation is more insightful than textual explanation, and vice versa.

This is a Kentucky warbler because this is a yellow bird with a black cheek patch and a black crown.

E4. User-tailored explanations: Social and situational aspects of explanations are at the heart of this activity, which addresses the question of what explanations need to be communicated to various agents and how to communicate them. The kind of explanations needed by an agent depends on the intended use of the outcome. Agents that intend to draw conclusions based on automatic data analysis, for instance, need to gain insight into whether the data or applied methods introduce a bias. The question of what explanations are needed is challenging, because agents using the output may not necessarily be aware that they require specific explanations. Agents need to detect specific multimodal signals for misunderstanding or ineffective and frustrated collaboration, which provide a motivation to explain a point of view or motivation. Artificial agents thus should be designed to provide specific explanation based on end-user and task profiles. To generate explanations that can be understood by the recipient (E1), an agent needs to use a mental model for all human and artificial agents it collaborates with (linking this activity to work on interaction and theory of mind in the research line on Collaborative AI). The explanation dialogue (E2) should be based on such mental models of other agents, but can also help to develop those models. The questions posed and elucidating examples provided by the recipient should be used to further inform the mental model the system has of the recipient.

Logic-based models of theory of mind developed by the team allow for knowledge-based agents to reason about the needs and intentions of other agents (Verbrugge & Mol, 2008). Our work on formalizing the notion of shared mental models allows a shared understanding of the current task that the system has to solve together with its collaborators (Jonker et al., 2011). We will also use GRaSP as an element of a computational theory of mind that allows us to model alternative views next to each other as well as the authority and expertise levels of the owners of these views, which can be the human partners with which a common ground of understanding needs to be established.

Techniques from E3 will be applied to improve the mental models that an artificial agent has of the artificial and human agents it interacts with. In particular, we intend to combine GRaSP and other methods from E1–E3 with logic-based models of theories of mind to run-time create ever more informed models of other human and artificial agents. The logical models provide the structure of what is to be learned, GRaSP provides the content and the provenance. Similarly, and linking to Adaptive AI, for generating explanations in a changing environment, logical models will help provide (temporal) structure while methods from E1–E3 for content and provenance.

E5. Evaluating Explanations: Evaluation methodology for assessing automatically generated explanations is still an open problem. This activity connects to all other activities in Explainable AI. It will be anchored in the team's extensive experience in evaluating interactive systems (such as search engines, recommender systems, and conversational agents), both based on labelled data in an offline setting (that is, prior to actual user interactions) and based on behavioural data in an online setting (that is, while being used by actual users).

We distinguish between the following types of evaluation approach: *offline evaluation*, *user studies & user panels*, and *online evaluation*. For offline evaluation (Sanderson, 2010) no experiments with human subjects are conducted; instead some formal definition of an explanation is used as a proxy for explanation quality. For user studies and user panels (Kelly, 2009), real human subjects are involved but the tasks are typically simulated or simplified versions; e.g., human agents are presented with an explanation and an input and must correctly simulate the model's output (regardless of the true output). For online evaluation, real human agents conduct real tasks with a real application (Hofmann et al., 2016), such as our experiments in education and science. It is important that these three types of evaluation inform each other: the factors that capture the essential needs of real-world tasks should inform what simplified or simulated tasks we consider, and the performance of our methods for generating explanations in offline evaluation should reflect their real-world settings. An interesting complicating factor here is the phenomenon called *algorithm aversion* (Dietvorst et al., 2015): when human users are confronted with errors in algorithmic predictions, they are less likely to use the model; users are more likely to choose a human decision maker instead, at great cost. How do we correct for this type of bias in our evaluation methodology?

Another important dimension here concerns the development of a solid set of informative metrics. To date, most attention in the evaluation of explanations is devoted to coverage (how much does the explainer explain?) and fidelity (does the explainer produce correct explanations?) (ter Hoeve et al., 2018). E.g., we should also develop metrics that capture aspects such as the degree to which an explanation is actionable, helps to calibrate trust, is relevant (given task, user and context), etc.

Connections with other research lines:

The choice of shared representations in activity E1 may have legal and ethical implications that should be considered (some representations must or may not be shared), so connects with the research line on Responsible HI. Similarly, the design of evaluation metrics for explanations (E4) will have ethical and legal aspects. The interactive aspects of E2 strongly connect with Collaborative HI, similarly the rationalisations from E3. The user-tailored explanations (E4) will benefit from results in the research lines on Adaptive HI and Collaborative HI and will itself contribute to responsible HI. E5 (evaluation) is internal to this research line.

	Collaborative	Adaptive	Responsible
E1			
E2			
E3			
E4			
E5			

Expected results

We expect to deliver algorithms that are able to generate explanations that are contrastive, selective, and social and situational. These algorithms will come with reference implementations. In addition, we will deliver methods and models for creating shared representations as well as methods for personalizing explanations. Second, we will deliver an evaluation framework for rigorously evaluating explanations, in three complementary and related approaches: offline evaluation, user studies & user panels, and online evaluation. In addition, we will deliver a set of evaluation metrics for the assessment of automatically generated explanations, with reference implementations. In addition, we will create and release a range of training and test resources for the development, evaluation and analysis of explainers. These will range from annotated data sets to informative simulations.

Core capabilities	Level 1	Level 2	Level 3
Transferability of shared representations	Within a single domain	Between different domains	Between different domains and tasks
Quality of the explanations[2]	Partial explanations	Full explanations	User-tailored full explanations
Interactive explanations	Instruction generation (speaker model)	Awareness of instructions (listener model)	End-to-end speaker & listener model

2.2.4 Research line: Responsible HI

Aim: how to ensure that the behaviour of artificial agents in HI systems is aligned with ethical and legal values and social norms
Coordinators: Verheij & Van Riemsdijk
Core participants: V. Dignum, Jonker, Van Riemsdijk & Van Wynsberghe (Delft), Verheij (Groningen); Prakken (Utrecht); de Rijke (UvA)

Overview and motivation

AI increasingly affects our decisions in our professional lives, in our private lives and in our public lives. However, 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 often 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 the capabilities of AI systems to reason about and act according to legal constraints and moral values, and to understand the consequences of their decisions. Research on the ethical, societal and legal impact of AI is paramount and needed today. It is urgent that this research accompanies the technological advances in the field rather than waiting until it is too late. 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 and UK House of Lords, and the European Commission. We take a dual approach to dealing with these complex challenges concerning legal and ethical HI systems:

Responsible Design of Intelligent Systems (Ethics *in* Design): This concerns the methods and tools for the design and development of HI systems that ensure (a) the analysis and evaluation of ethical, legal and societal implications; (b) the participation and integrity of all stakeholders as they research, design, construct, use, manage and dismantle HI systems; and (c) the governance issues required to prevent misuse of these systems. The aim is to ensure that HI system development is mindful of the consequences for individuals and societies, by anticipating the consequences of the design choices, reflect upon the problem being solved by engaging all stakeholders, and take appropriate action to ensure social, legal and ethical acceptability of the system.

[2] Metric to be determined in E5

Ethical or Legal Reasoning by Machines (Ethics *by* Design): This refers to the computational and theoretical methods and tools that support the representation, evaluation, verification, and transparency of legal or ethical deliberation by machines with the aim of supporting and informing human responsibility on shared tasks with those machines as in Hybrid Intelligence. Even though intelligent systems are increasingly able to take decisions and perform actions that have legal or moral impact, such systems are artefacts and therefore are neither ethically nor legally responsible (or at least not yet). This research is on understanding what suitable constraints on system behaviour are and on developing the desiderata and corresponding computational models on the representation and use of legal or moral values by HI systems.

State of the art

We distinguish ethical or legal reasoning about HI systems and ethical or legal reasoning by HI systems. Concerning **ethical/legal reasoning about HI systems**, it is urgent to have answers to questions such as 'What does it mean for an HI system to make a decision?', 'What are the moral, societal and legal consequences of their actions and decisions?', ''Who is to be held accountable for the actions of an HI system?' and 'How can these systems be controlled once their learning capabilities bring them into states that are only remotely linked to their initial, designed, setup?'. 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, liability of and for any (semi-)autonomous system remains a challenge, requiring again 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 (Koops, 2010), from strict liability for manufacturers to reversing the burden of proof, to compulsory certification or automated compensation in the case of smart contracts (De Filippi & Hassan 2016). 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 (Berendt et al., 2014). Underlying all of the above, there is a need to **analyse the social, ethical and legal characteristics of the domain**. 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 (AMAs) – 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? Ought AI machines to remain tethered to humans? Can machines be assigned moral roles or moral capacities? Should machines be made accountable or responsible for consequences? Nevertheless, AI systems are already making decisions of ethical or legal consequence and there is a need to understand the relationship between the consequences of AI on the one hand and the choices of users and designers on the other. 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, Ganascia, 2007) 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 (Arkin, 2009, Coelho, 2010, Cointe, 2016).

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; Ashley 2017; Branting 2017).

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

(Verheij et al. 2016). Second, since rigid deductive application of legal rules often leads to immoral, unfair or socially undesirable outcomes and also since legal rules need to be interpreted, judges and lawyers often resort to non-deductive forms of reasoning about the rules, including reasoning with cases, analogical reasoning and reasoning about purpose, principle and values. AI & Law models of these non-deductive kinds of reasoning exist (for overviews see Ashley 2017; Prakken & Sartor 2015). However, they have not yet scaled up to practical application, 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 (Ashley 2017; Branting 2017) may provide opportunities, but employing them in the right way to obtain the necessary knowledge to overcome this barrier is highly challenging.

An important research issue here is whether these approaches can also be applied to deal with moral reasoning. Ensuring ethical behaviour of intelligent agents meets similar challenges to legal reasoning, but to a greater extent, since in ethics, unlike in law, there are no authoritative or even written sources and no institutions for promulgating or applying ethical rules and principles. This makes the need for relying on cases, analogy and values even bigger in ethics than in the law. The challenge is to make the fruits of AI & Law research on these issues applicable to AI ethics while respecting the differences between law and ethics.

Another reason why the fruits of AI & Law cannot be directly used for ensuring legal or ethical behaviour of ethical agents is that most AI & Law models are for assessing behaviour of other agents in retrospect (like judges do) but legal or ethical intelligent agents have to decide about their own future behaviour. Research on making autonomous AI systems behave lawfully is scarce and in its initial stages (Indurkhya, Hage & Broszek, 2017).

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 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 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:

1. How to include ethical, legal and societal (ELS) considerations in the HI development process? (ethics in design)
2. How to verify the agent's architecture and behaviour to prove their ethical 'scope' (ethics in design)
3. How to measure ELS performance and compare designed systems vs learning systems? (ethics in design)
4. What are the ELS concerns around the development of systems that can reason about ELS consequences of their decisions and actions? (ethics by design)
5. Which methodology can ensure ELS alignment during design, development and use of ELS-aware HI systems? (ethics by design)
6. What new computational techniques are required for ELS in case of HI systems where humans and artificial agents work together?

Research Activities

These research questions will be investigated in the following research activities.

R1. Explaining algorithmic decisions that have ethical, legal or societal implications.

For example, explaining decisions of algorithms for predictive policing, fraud detection, personalized advertisements or job candidate selection (related to research question 4). This question is related to the questions addressed in the research

line on Explainable HI and will build on results from that research line. However, in the research line on Responsible HI we want to address some specific issues that do not arise in general. With algorithms that take ethically or legally relevant decisions, the main focus is not on understanding how a decision was reached (as e.g. in medical diagnosis) but on critically examining whether it is morally or legally acceptable. Techniques for modelling such critical examination will include argumentation-based techniques of two kinds: techniques for case-based reasoning as developed in AI & Law for dealing with the specifics of cases, and dialogue models of argumentation for facilitating dialogues between the artificial agents and human users about the moral or legal quality of the decision. These techniques will be developed while keeping possible differences in background knowledge of the humans (e.g. professional lawyers versus citizens) in mind.

R2. The Algorithmic Glass box (research question 2). Rather than opening the 'algorithmic black box' we will develop verification mechanisms to guarantee that input and output of an AI system (black box) meets a given set of ethical/normative principles. In recent work of the consortium, we have shown how argumentation and scenarios can be used as tools to open the black box of statistical model, in particular Bayesian networks (Timmer et al., 2017; Vlek et al., 2014, 2016; Verheij et al., 2016; Verheij, 2017). In this work, argumentation and scenarios provide explicit explanatory structure in statistical models. Van de Gaag has been effective by coupling ontologies to Bayesian Networks for generating explanations (Helsper et al., 2007). We propose to use machine learning techniques to improve on both the ontologies, as well as the coupling of the ontologies to the Bayesian Nets, as well as to the argumentation systems and scenarios. This will improve provenance. Important research issues are how to connect logical and statistical modelling and how these are connected to decision making (using values, preferences and utilities) (Bex et al., 2017; Verheij, 2016). Argumentation can be used to guide towards responsible choices. Scenarios can be used to describe settings of responsible choices.

R3. Mining opinions, values and arguments concerning ethical, legal or social issues (research question 5 and 3). Using machine learning and natural language technology for mining opinions, values and arguments concerning ethical, legal or social issues addresses the knowledge acquisition problem for ethically, legally or socially aware agents. Ethical and social knowledge, opinions and values will not be available in ready-made rule-based form but must be mined from many natural language sources. The same holds for legal information in interpretation issues, or for the ground on which a legal rule can or should be set aside because of the social context of ethical considerations. The mined opinions, values and arguments also contribute to the design of measurements of ELS performance.

R4. Algorithmic integration of ELS reasoning and decision-making capabilities in artificial agents. Even though AI systems are artefacts and therefore are neither ethically nor legally responsible, AI systems are increasingly able to take decisions and perform actions that have legal or moral impact. This research is on understanding how HI systems can take ethical, legal and societal considerations into account while reasoning about their decisions.

A major methodological issue here is whether HI systems should be designed to reason with explicit legal or ethical knowledge about how to behave in a lawful or ethical way or whether they can be trained to do so with machine-learning techniques applied to a large number of training cases (as is currently the dominant method for making self-driving cars conform to the traffic rules), or whether some meaningful combination of both is possible. All these approaches will be investigated. In the reasoning approach formal techniques from logic, argumentation, probability theory and game- and decision theory will be further developed and applied, while the above-mentioned knowledge-acquisition bottleneck will be addressed by integrating these symbolic approaches with machine-learning approaches for providing their inputs, as well as interactive approaches in which knowledge is obtained through dialogues with users. A limitation of the training approach is that the ethical or legal knowledge on which the AI system is trained is often implicit in the learning method, which hinders validation (Leenes & Lucivero, 2014). This problem we want to address in the "ethics IN design" part of the research, by developing principled techniques for training and validating systems on the basis of explicit 'off-line' representations of the relevant ethical or legal knowledge.

R5. Context-aware ethical or legal reasoning. Straightforwardly applying clear ethical or legal rules to a case is easy to automate but often leads to immoral, unfair or socially undesirable outcomes. This topic concerns the development of computational models of normative reasoning that take the particular of a case and the social context

into account, and contributes to research questions 4 and 5 about ethics by design, and to research question 6 on ELS in case of Hybrid Intelligence where humans and artificial agents work together.

Connections with other research lines:

Activity R1 is closely linked to the research line on Explainable HI. Activity R2 has connections with the research lines on Adaptable HI because the current techniques for adaptivity are not developed for transparency. Activity R3 is a specific instance of Adaptive HI, since a generic agent architecture would have to adapt itself to reflect the acquired ELS arguments. The ambitions of research lines on agents that perform ELS reasoning (R4 and R5) will require techniques from explainability and adaptivity.

	Collaborative	Adaptive	Explainable
R1			
R2			
R3			
R4			
R5			

Expected results

Responsible Design of Intelligent Systems (Ethics *in* Design): methodology and supporting tools for (a) a design process that ensures ethical, legal and societal issues are taken into account, through (b) the participation of all stakeholders, and (c) defines governance models to prevent misuse.
Responsible Behaviour of Intelligent Systems (Ethics by Design): theoretical methods and computational tools to represent, evaluate and verify the ethical, legal and social effects of deliberation by machines, aiming to support human responsibility on shared tasks with those machines.

The activities lead to agents that will have the following core capabilities, that have measurable dimensions of progress:

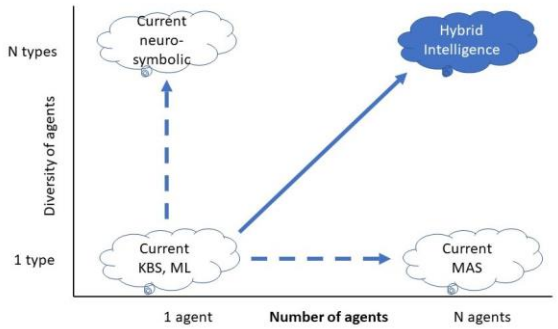
Core capabilities	Level 1	Level 2	Level 3
Critically examining algorithmic decisions of big-data applications on their legal or moral quality	Identifying the grounds on which a decision was reached	Monologically assessing the legal or moral quality of the decision	Engaging in human-machine dialogue about the legal or moral quality of the decision
Validating whether legally or morally acceptable behaviour is learned	Consistently and completely representing ethical or legal knowledge for validation purposes	Matching the representations with the learned behaviour	Improving the learned behaviour
Reasoning about the legal or ethical acceptability of intended behaviour	Combining reasoning with formalized legal or ethical knowledge with self-interested motivations	Combining reasoning with legal or ethical knowledge extracted from unstructured sources with self-interested motivations	Engaging in human-machine dialogue about the legal or moral quality of the intended behaviour

2.2.5 Evaluation

Investigating the design space

Hybrid Intelligent systems are highly complex systems with multi-faceted functionality. Their functionality can be organized in a multi-dimensional design space, and we will use this design space as an organizing principle for the project. Each of the dimensions of this design space corresponds to a specific functional dimension of HI systems, with increasing values along such a dimension representing increasing functionality of the system. Very simple HI systems would have low values on many or all of design dimensions, while increasingly advanced HI systems are characterised by increasing values on many or all of the functional dimensions. Some example dimensions that we are able to identify at the time of writing are:

Cardinality dimension: One dimension of this design space concerns the number of agents that a HI system can collaborate with, ranging from simple 1-on-1 collaborations via 1-to-many and ultimately many-to-many collaborations. This dimension would characterise some of the work from the Collaborative HI research line.



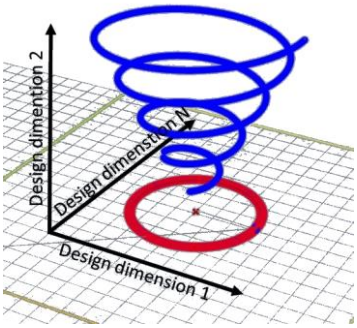
Heterogeneity dimension: A second example from the same research line concerns the degree of heterogeneity among agents in a HI system. Are all agents of a single type, or does a HI system contain multiple types of agents?

Explainability dimension: Another dimension of the design space captures the quality of the explanations generated by the systems: are they partial explanations, more complete but generic explanations, or are they both complete and tailored to specific users? This dimension characterises some of the work from the Explainable HI research line.

Interactivity dimension: Another dimension from Explainable HI covers the degree of interactivity, ranging from entirely statically presented explanations via partially interactive explanations to full dialogues.

Shared knowledge dimension: A final example concerns the degree of shared knowledge possessed by agents: do they only have individual knowledge, is some knowledge shared between some agents, or do they have complete common knowledge?

Over the course of the project, we will investigate increasingly advanced regions of the HI design space. This will also lead to the identification of entirely new dimensions of the design space of which we are currently unaware. Our results will therefore be both practical, with increasingly advanced HI functionalities along many design dimensions, as well as theoretical, with a much better understanding of the structure of the design space for HI systems, including currently unknown dimensions, an understanding of the interactions between the different dimensions, and a theory of which types of HI systems inhabit which regions of the space.



Measuring progress through controlled experiments

While traversing the HI design space, we will use a process of controlled experiments to continuously measure our progress. Members of the consortium are experts on the understanding of social collaboration in human teams (Neerincx et al., 2018; Harbers et al., 2014). We will deploy similar experiments as those used in experimental psychology by members of our consortium, but will now apply them not to human-only teams, as is customary, but instead to hybrid teams of human and artificial agents. By subjecting our HI systems to increasingly ambitious experimental designs, we can measure how well we are progressing along the dimensions of the HI design space.

These experiments are executed in well-controlled settings, such as a mixed human-machine HI system trying to solve a public-goods dilemma, where all individuals desire a public good that can only be obtained through combined effort from individual agents, but once attained, the public good is available to all agents, irrespective of the amount of their contribution, thereby encouraging free-riding by individual agents. Real life examples are taxation for roads, paying for TV licenses, etc.

In a related experimental setting, individuals have to balance their individual interest of getting as much of a shared resource as possible against the shared interest of not depleting the resource faster than its replenishment rate (this experiment is popularly known as the tragedy of the commons). A real-life example is the use of scarce water resources.

Communication capabilities experiment: Since the late 1950's it has been well known that in human groups communication enhances cooperation on resolving social dilemma (Balliet, 2010). This would predict that cooperative

performance of our HI systems will increase when we improve their communicative capabilities. Conversely, the cooperative performance of the system can serve as a metric for our progress on the communication dimensions in the design space.

Multi-modal communication experiment: Similarly, face-to-face communication is known to be 2-3 times more effective in social collaboration than written messages. So, ceteris paribus, we can use social collaboration experiments to measure our progress along the dimension from unimodal communication (written messages) to multimodal communication.

Trust experiment: Trust is a key aspect of cooperation and any HI system should be trusted and be built to facilitate trust among group members. Meta-analytic work (Balliet et al., 2014) uncovered that in-group favouritism is stronger with increased common knowledge of group membership (as opposed to only unilateral knowledge). This yields interesting experimental designs along the shared knowledge dimension of the HI design space.

We will execute an annual series of these controlled experiments with the HI systems designed and built by the consortium. This will not only yield insights in our progress along the various dimensions of the design space, but it will contribute to the theory of the "sociology of HI systems", by translating socio-psychological insights on human group performance to insights on the performance of Hybrid Intelligent systems.

Measuring progress through field experiments

Teams in the consortium are already actively testing and deploying AI systems in a variety of application areas. We will use our knowledge and our network of partners in these areas to field prototypes and applications based using HI techniques and principles:

Education: We have developed a RoboTutor using cognitive agent technology that is used as a teaching aid by teachers in the classroom. This RoboTutor teaching environment is deployed from the cloud by the TU Delft spin-off Interactive Robotics and in daily use by more than 15 primary and secondary schools (www.robotsindeklas.nl). The RoboTutor can assist children with key learning goals related to math and language skills. It also offers a platform to teach children computational thinking and programming skills. The interaction design developed using user-centered methodologies has been evaluated with respect to learning effects, the impact on stress of children related to the learning task (e.g. math), and effects of feedback styles and affective models (Xu et al., 2014).

Healthcare: In the European ALIZ-E and PAL-projects (www.pal4u.eu) on social robots for children with diabetes, we have developed and applied a situated Cognitive Engineering methodology to systematically assess robot's assistance for child's well-being (including the empowerment of child's social environment, i.e., the health-care professionals and parents) (Looije et al., 2017). Test metrics included health outcomes, goal achievements, knowledge (progress), cognitive and physical behaviours, mental and physical conditions (e.g. stress), trust and acceptance. Evaluations of specific components took place at schools and diabetes camps, whereas the overall system has been tested up to a period of use of 3 months in the hospital and at home (in successive design-test cycles) (Henkemans et al., 2017).

Disaster response: In the European NIFTi and TRADR-projects (www.tradr-project.eu) on robot-assisted disaster response, we have developed and applied a methodology to test the effects of the technological innovations on human-robot team performance in a systematic way (every year) (Kruijff-Korbayová et al., 2015). Challenging scenarios were developed with the domain experts and end-users from the participating German and Italian fire brigade organizations and, subsequently, simulated in real field settings and, partly, in virtual reality. To establish a sound and appropriate set of metrics, we conducted a value elicitation workshop (Harbers et al., 2017), and comprised a coherent set of sophisticated human factors measure instruments on effectiveness, efficiency, satisfaction, situation awareness and trust. Unit tasks were designed to evaluate functional components, whereas the overall scenario was used to evaluation the combined effects (Mioch et al., 2012; Horsch et al., 2013).

Negotiation: For the construction of negotiation support systems we combine insights on the strengths and weaknesses of human negotiators with those of automated negotiating agents to create synergy between humans and agents. Humans fail to find Pareto optimal outcomes, but have good contextual information, and vice versa agents are good in finding Pareto optimal outcomes, but lack context (Hindriks & Jonker, 2009; Jonker et al., 2017). In balanced field experiments (Bosse et al., 2008) we have measured the quality of the negotiation outcomes (individual utility, social welfare), the quality of the negotiation process (understanding of bids, reactions to bids, and strategy),

the understanding of the wishes of the negotiating parties (preference profile, underlying concerns, strategic considerations). Such a testbed for negotiation can also serve to test the quality of explanations (insight gained about outcomes, process and negotiation parties), as well as the adaptivity of the HI system in following the change in strategies, and preferences of the negotiating parties. The annual international Automated Negotiating Agent Competition (Jonker et al., 2017) will play an important role to measure our progress.

Testing with external parties

Whereas our yearly experimental cycle has the benefit of carefully controlled experimental conditions and clearly measurable outcomes, they lack the realism that comes from case studies in the external world. In order to test our understanding of HI systems in realistic circumstances, we will twice organize a challenge competition (in years 4 and 8), where external parties can submit proposals to deploy the results of our research in a practical setting. The winners of these challenge competitions will be awarded with 1 FTE of effort funded by the project, as well as hardware resources (e.g., for multimodal interaction, plus a certain amount of consulting time from members of the project). In return for this, the winners of the challenge are expected to field a prototype HI system in their specific societal domain. We expect contestants for these challenge calls to come from diverse domains, such as healthcare, policy making, business management, computer security and education.

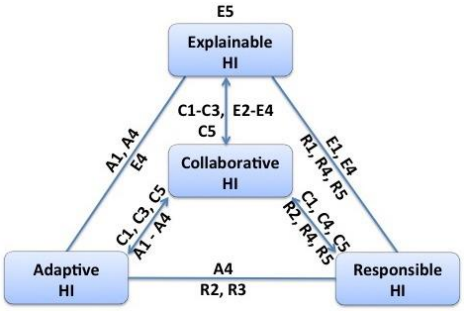


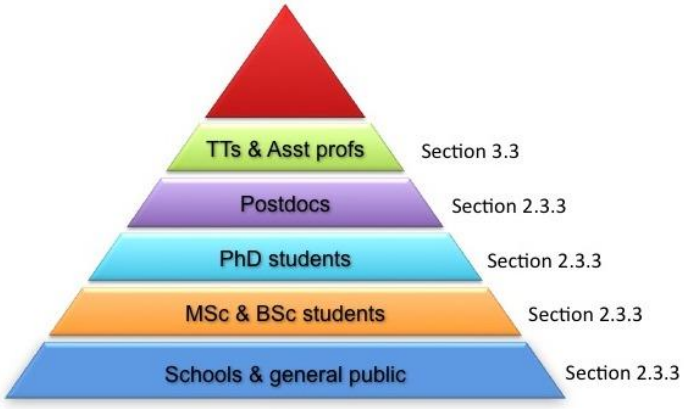
Figure 7: Interconnected research lines

2.3.3 Training and education policies

The HI Centre will ensure sustainability of the HI research agenda, and work towards the transfer of knowledge towards industry and society at large. It has policies in place to train and educate the future talent that will keep the Netherlands at the heart of the progress in HI research. The HI Centre will deploy training and education activities at all levels of the expertise pyramid (Figure 11). Here we describe education and training policies for PhDs and Postdocs. Section 3.3 describes special career development policies for tenure track positions and assistant professors.

Training and education are at the core of delivering the new generation of scientists in the new paradigm of Hybrid Intelligence. Special attention will be given to the development of mid-level careers which will be discussed in detail in Section 3.3. At PhD level and younger, our training, knowledge exchange, education and internal dissemination activities will comprise:

Figure 11: Training of scientists at various levels



An extensive PhD training and teaching program

The PhD training and teaching program of the HI Centre covers general skill development, scientific disciplinary knowledge and specialized topics, as well as ethics and responsible research practices. For the **courses from the broader discipline**, all HI Centre

PhD students will have access to the course portfolio of the national research schools SIKS (Netherlands Research School for Information and Knowledge Systems, www.siks.nl) and Kurt Lewin (social psychology, www.kurtlewininstituut.nl). All of universities participating in the HI Centre are members of the SIKS school, and five of the six have access to the Kurth Lewin Institute portfolio. SIKS organizes a wide range of courses in the broad discipline of Information and Knowledge Systems on a regular basis (24 events in 2018 alone), and covers, among others, AI, information systems, human computer interaction, multi-agent systems and other areas relevant to Hybrid Intelligence. The Kurt Lewin Institute provides postgraduate courses on theory, methodology and practice of social psychology (17 events in the academic year 2018/2019 alone). All these courses of both national research schools are available to HI Centre PhD students. Beyond these broad courses, **specialized courses** in HI topics will be organized by senior HI Centre researchers, with a target of 3 such courses each semester. To ensure impact beyond the HI Centre, these courses will be co-hosted with the national research schools, and will be open to other members of these research schools. Each of the participating Universities has a mature offering of **courses on general skills** such as scientific writing, presentation skills, grant writing, career development, etc. HI Centre PhD students will be able to benefit from these courses at their home University.

Each PhD student will be required to submit a **teaching and training plan** in the first months of their appointment to the satisfaction of the HI Centre Training and education director (a practice already in place in many of the participating universities). The HI Centre will comply with the formal requirements on PhD education of the home University of each PhD student. Each PhD student has a **personal travel and training budget** of 18k€ available for participation in international conferences, summer schools, doctoral symposia, research materials and computer time. Spending of this budget is under responsibility of the supervisors of each PhD student. The HI Centre Training and education director will oversee that a sufficient part of this budget is spent on training activities.

Training at a participating host institute: Each PhD student is expected to spend three months at one of the other participating universities, either with the 2nd supervisor or another institute. This immerses the PhD student in another research culture, as well as the consortium benefiting from stronger coherence.

Training in organizational tasks: The senior PhDs (3rd year) will take part in the organization of courses and teaching and will support the organization of the *HI Centre* Summer school and other activities (e.g. the HI Centre Olympiad, discussed later). In doing so, they will develop presentation, teaching and coaching skills resulting in their certified University Teaching Qualification (UTQ). Within these tasks, senior PhD's will furthermore be appointed the role as mentor for one or two junior PhDs.

Training in dissemination: All PhDs will present their work at international conferences and symposia, via which intensive participation and scientific exchange is achieved. Each PhD student has sufficient funds allocated to participate in the major meetings of their field.

Collaboration with industry: In cooperation with industry, traineeships for PhDs will be offered in order to provide a broader view of the HI field. Traineeships will typically last 3 months in order to create sufficient understanding of the industrial perspective of HI. The participating research groups have ample experience with such PhD traineeships at institutions such as TNO, ING, Google, IBM and others.

Supervision and quality assessment of PhD's: Each PhD student will be supervised by 2 faculty members: one from the home university and one from another HI Centre university. A personalized supervision plan will form the basis for the quality of the research project. Quality assessment of the individual PhD trajectories will be organized via an annual review per student, based on self-assessments, supervisor assessments, and research papers. Each PhD student in the HI Centre will comply with the quality assessment process of their home university.

Postdoc training and development

In addition to the extensive activities for TTs and UDs (Section 3.3), the training of Postdocs requires a number of additional activities that the institute will provide for:

1. **Structural *support and supervision*** by (associate) professors who act as mentors and advisory contact points;
2. **Regular *peer-to-peer intervision groups*** will be created for Postdocs from the different HI universities as a forum for the postdocs to share feedback and benefit directly from the unique network;

3. **Specific *advanced trainings***, including those provided in-house at the HI partners, and additionally externally from the SIKS and Kurt Lewin graduate schools, in particular course on grant writing and career choices.

Training at a participating host institute: Each Postdoc is expected to spend three months at another participating university, either with the 2nd supervisor or another institute. This immerses the Postdoc in another research culture, as well as the consortium benefiting from stronger coherence. Similar to PhD students, each Postdoc has a **personal travel and training budget** of 18k€ available for participation in international conferences, summer schools and doctoral symposia.

Dedicated exchange programme: Besides the regular exchange programme for all researchers in the HI Centre, we will set up a *dedicated exchange programme* for Postdocs with partners in the EU Flagship consortium on Human(e) AI. Such exchange activities with international FET Flagship partners will give our Postdocs the opportunity to gain international experience of working within other top research groups, an experience which is seen as crucial for making the step to a tenure track position.

Other training and education activities

Education materials for Bachelor and Master programs: We will develop course material for MSc and BSc level. All senior HI Centre staff are actively engaged in BSc and MSc teaching, and are committed to introducing HI techniques and insights into curricula such as Artificial Intelligence (UvA, VU, Utrecht, Groningen), Computer Science (all partners) and Information Science (UvA, VU, Utrecht). Many of the HI Centre faculty have experience with the production of online teaching material. This will attract awareness for HI and attract potential MSc and PhD students; 4 partners in the HI Centre (VU, UvA, Utrecht, Groningen) already run a master program in AI. Core topics include: Computer Vision, Evolutionary Computing, Information Retrieval, Knowledge Representation, Machine Learning, Multi-Agent Systems and Natural Language Processing.
Summer schools: Summer schools are organized bi-annually alongside the HI Centre conference. The HI Summer School attendees will be taught by PIs on cutting-edge research in their discipline. As the Summer School is held alongside the HI Centre conference, the PhD's and Postdocs also have the opportunity to present their latest results.
Education activities for secondary schools: Dissemination and training material for secondary school students will be prepared in order to increase awareness and interest. Also, talented secondary school students will be invited to events at the university, e.g. for a HI Olympiad. PhDs and Postdocs will play an important role in the organization of these activities. Education is a prime example for an application area of HI, as both machine and human intelligence is required to ideally support the learning at any age. Education activities in secondary schools (and maybe even primaries) will serve double purpose, help assess and further develop HI methods and tools, as well as educating young people in AI and HI. Previous examples of this synergy have been the usage of robots and other AI technology in schools or hospitals by members of the consortium.

2.3.4 Recruitment & diversity policies

*An **incentive of 20 %** of the funding for each university will become available when a 50/50 gender balance has been achieved on the first 80 % of the funding. If not yet achieved, that 20 % of the funding will have to be spent to **correct any gender imbalance.***

Our recruitment policies are aimed at attracting international talent at every career level. Key to this is a challenging and visible scientific program, the participation of top scientists in the field of HI, the well-known reputation of Dutch Universities as being open and international environments, excellent connectivity by air travel and public transport, and their location in cities with a high quality of life. Positions will be advertised internationally. We will utilize our broad network, both within the Netherlands and internationally, to actively scout for talents in a highly competitive market for talent, to ensure we attract top candidates. PhD positions will be allocated to new tenure track positions to increase attraction and to ensure we attract top talent. The HI Centre aims at a balanced composition of its staff regarding gender and diversity. Our participation of **33% female (co-)PIs**, is well above the average of our research fields, both nationally and internationally (10.8% full professors in natural science in The Netherlands).

Similarly, we have a **33% female participation among our mid-career scientists** (PhD date after 1998), rising to **55% female participation among the early career scientists**, showing our commitment to improving the gender balance in future generations. All HI Centre participants have agreed on a gender balanced hiring policy, targeting **a 50/50 gender balance**. An incentive scheme for this target has been agreed upon (see text-frame). Vice-director Jonker will act as diversity officer whose task it is to continuously monitor all activities in the consortium (hiring, invitations, roles in the consortium, etc) and to report to the management board on this. Vacancies in the HI Centre will be advertised internationally. It is expected that this will lead to a substantial non-western influx of scientists from e.g. China and India which will increase diversity. Two of our (co-)PI's have appointments at Chinese Universities which will facilitate this recruiting. Specific activities will be carried out within our network, e.g. Women in AI lunches or breakfasts and contributions to activities such as the Digivita summer camps.[3]

2.3.5 Ethics policies

For research on HI systems, ethical aspects are extremely relevant. All HI activities will be carried out in accordance with the legal regulations of the Netherlands and the ethical rules as formulated by the Royal Netherlands Academy for Arts and Science (KNAW) and as developed by the Association of Universities in the Netherlands (VSNU). Furthermore, an entire research line is devoted to Responsible AI. Research activities that involve ethical aspects (such as participation of human subjects, or use of sensitive data sources) will require approval from the ethics review boards of the participating universities. All of the participating Universities have such an ethics board in place. Two of the HI centre partners (the coordinator Vrije Universiteit Amsterdam and the University of Amsterdam) have a dedicated ethics board for research in Computer Science, AI and Information Systems <http://www.delaat.net/ecis/> which will advise on our experiments when required. Two of our consortium members (Dignum and Van Wynsberghe) are members of the High-Level Advisory Group on Artificial Intelligence of the European Commission, whose task it is among others to propose draft AI ethics guidelines to the Commission[4], ensuring that our ethics policies are in line with the European policies.

2.3.6 Progress monitoring and reporting

We will monitor progress for individual dimensions of HI systems, for integrated HI systems in lab settings, and for HI systems in field environments:

Per HI dimension through Capability Matrices. Each of our four research lines study and develop different aspects of HI systems. For each of the research lines we have defined a capabilities matrix: a number of "core capabilities" of HI systems, with three levels of increasing sophistication for each capability (see sections 2.2.1 - 2.2.4). These HI capabilities will be the basis for tracking progress of the four research lines. This will be monitored by the Executive board at **6-monthly intervals**, based on reports from the research line coordinators.
Integrated HI systems through lab experiments. We will subject the HI systems that we develop to a series of experiments that have been designed in experimental psychology the understanding of social collaboration in human teams, but will now apply them not to human-only teams, as is customary, but instead to hybrid teams of human and artificial agents. By subjecting our HI systems to increasingly ambitious experimental designs, we can measure how well we are progressing towards the goal of truly collaborative HI systems. The joint research line coordinators will report on the results of these experiments to the Executive Board **annually**.
Integrated HI systems through field experiments. Teams in the consortium will be using the results in field experiments that they are involved in through other projects in healthcare, education, disaster response and negotiation (see section 2.2.5). Furthermore, we will be conducting field experiments with external partners, to be selected through two rounds of competitions in years 4 and 8. Progress as measured through these field experiments will be reported to the Executive Board **annually**. The Executive Board will **report progress to NWO and to the Governing Board annually**, including scientific progress, a budget report, a diversity report, plus any major issues that occurred and how these were managed.

[3] https://www.amsterdamsciencepark.nl/news/digivita-zomerkamp-voor-meisjes/
[4] https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

2.4 Knowledge use and transfer

Together with scientific excellence, the usage and dissemination of the knowledge produced will be paramount to the success of the HI initiative. The HI Centre will serve a large and very diverse set of stakeholders, for which institutional embedding and a dedicated dissemination strategy is essential.

2.4.1 Foreseen results

The HI Centre results will include novel data sets, design methodologies, libraries, software packages, theories and tools for Hybrid Intelligence. These results will be enablers for new systems and applications in a wide range of disciplines. Among these, we will prioritize applications in education, healthcare and science:

Education*:* HI systems (both virtual agents and robots) to interact with children in an educational environment, e.g to help children with concentration deficits to study better, or to relieve teachers of repetitive teaching tasks. We will collaborate with the TU Delft spin-off Interactive Robotics, which already fields educational robots in 15 primary and secondary schools (www.robotsindeklas.nl).
Healthcare: first application*:* HI systems, again both virtual agents and robots that interact with children in hospitals, to provide emotional support for children with cancer to alleviate their stress and fear at hospital visits, and provide them with entertainment and information during prolonged hospital stays. We will build on existing collaborations of members of the consortium with paediatric oncology departments in academic hospitals. **Second application**: Virtual agents or robots interact with children with chronic diseases such as diabetes to monitor and assess their lifestyle and to provide them with appropriate information and feedback. We will build on existing work by members of the consortium in projects such as ALIZ-E and PAL (www.pal4u.eu).
Science: Virtual agents that interact with scientists on large scale analytics of the literature in a specific field, use these for formulating hypotheses and designing experiments to test such hypotheses. We will build on existing work by consortium members with pharmaceutical scientists in the OpenPhacts project (https://www.openphacts.org/) and on our work on a meta-analytic dataset that contains the entire history of laboratory research on human cooperation.

2.4.2 Target groups

Target groups for knowledge use and transfer are:

Industry: Industries of the future will require hybrid teams of robots, softbots and humans bringing their complementary skills to the task. This kind of collaboration is crucial for the economy and requires the hybrid intelligence that we are investigating. In complex process industries like (petro)chemical, transport and the retail sector, problem solving and system design will benefit from the support of and collaboration with HI systems.
Applied scientists and technology developers: As a scientific project the most direct target groups for knowledge use and transfer are the direct users at universities and (applied) research institutes (both public and private), as well as technology and system developers, for which the project deliverables will be immediate and valuable assets for further research in their value chains. These key users may carry out applied research or develop systems based on our results, or carry out research in which HI results are used. The evaluation and benchmark platforms delivered by HI will form a strong basis for future research as it defines the standard and state-of-the-art for system performance.
Societal stakeholders and the general public: Societal stakeholders and the general public will have to be informed about the developments within HI, to provide more nuance to the current debate about AI. As HI systems become more and more visible throughout society, it is important to involve stakeholders and to create awareness among the general public of the consequences and opportunities of HI systems for society. Consumer organizations and the general public have an interest in ensuring that important societal, ethical and cultural values such as transparency, trust and explainability are realized in future HI systems. Furthermore, communication of the potential

of science and innovation to the general public is essential for creating public support for investments in science and innovation and to attract students to the studies relevant to our program.

2.4.3 Dissemination strategies and valorisation activities

At the start of the program, we will develop a HI Centre dissemination strategy under direction of our Knowledge and Outreach officer with planned activities for the duration of the project, including sections of the HI Centre website and a communication plan both targeted at our knowledge-transfer stakeholders. During the program, we will adapt the strategy based on new developments and feedback from the advisory boards. We have allocated budget and staff (0,3 FTE) for dissemination and valorisation activities. Additionally, each PhD student and post-doctoral researcher will have a budget within their host university that can be utilised for dissemination of their research, training and development purposes. Our dissemination activities will be able to fully make use of the **Innovation Centre for AI (ICAI)**, an open collaborative initiative between knowledge institutes that is aimed at AI innovation through public-private partnerships. The centre is located at Amsterdam Science Park and is initiated by the UvA and VU together with the business sector and government, with three of our (co-)PI's (De Rijke, Welling and Van Harmelen as founding members). The organization is built around industry labs – multi-year partnerships between academic and industrial partners aimed at technological and talent development. ICAI is an excellent channel to transfer the scientific results of the program to more applied research and system development.

General dissemination and valorisation activities

During the program regular dissemination activities will be carried out, including articles in peer-reviewed academic journals, presentations at and participation in scientific conferences and summer schools, preparation of articles in popular magazines and newspapers, presentations and interviews on radio and television and publication of messages and articles on the HI Centre website, social media tools (LinkedIn, Twitter) and in our quarterly newsletters. We will also utilize existing active communities from the HI Centre Partners, for example our Meetup community for events which currently has 5000+ members. Infrastructure and cooperation will play an important role in disseminating HI. Current plans include building an exemplary high-standard research platform for collaboration. This will encompass costs for computer servers and robots and three engineers full time throughout the project (together €2.25M). The data and software licencing policy (based on Apache 2.0 licences) of the HI will help others to use and disseminate HI technology and encourage joint collaboration of HI partners with potential stakeholders. A specific aspect of our dissemination and valorisation activities is that we will use the results of our research to carry out dissemination activities, e.g. interactive and adaptive chatbots and robots that explain our results in plain language to school children (Hindriks, Vossen and Neerincx) and to potential end users like farmers (Van der Gaag), engineers (Monz) and the general public (Eiben and Vossen).

Targeted dissemination and valorisation activities

We will also implement a number of activities for particular target groups:

Universities, (applied) research institutes, technology and system developers: Specific activities include:

- Meetings, courses and symposia in conjunction with the HI summer schools.
- Innovative public research sessions (e.g. serious gaming) using the HI Centre Lab.
- Competitions for the use of knowledge with an allocated budget for the top proposition deploying the results of our research.
- Hackathons where external organizations participate intensively over a short period of time (e.g. 2 days) to work hands on with results from the consortium.
- Collaboration with the e-Science Centre for testing and deploying our results with early adopters in specific scientific fields.
- Industry outreach days: These will be organized bi-annually in order to present results to industry, to discuss potential areas of application and to get feedback on the research program.

Governments, education, healthcare and industry: Specific activities include:

- Knowledge sessions and workshops between researchers and relevant stakeholders.
- Joint implementations of research results, investor finding, research symposia. These could, e.g., take place within ICAI, creating multi-year partnerships between academic and industrial partners aimed at technological and talent development.
- Participation in the National Science Agenda, top sector and NWO topic related discussions.
- Facilitate HI-lab-usage by start-up companies to collaborate with scientists and engineers.

Societal Stakeholders and the general public: Specific activities include: meetings, open forums and information sessions with stakeholders, opinion articles in major new magazines; contributions to popular opinion and discussion programs on radio and TV; dissemination of news and results on the HI Centre website and the participants' websites. Members of the consortium are well known public educators: Vossen and Eiben in Paradiso lectures, Vossen with a keynote at the Infosecurity.nl, Data & Cloud Expo, van Harmelen on VPRO television and Universiteit van Nederland, Welling has a column in FD, de Rijke was speaker at the NRC day, events at NEMO, among others.

2.5 Connection to the National Science Agenda

Our program is well linked to the National Science Agenda (NSA). The NSA question "how can human insights be combined with AI to come to better decisions?" is exactly the main question of our programme. Our research contributes directly to the Agenda's route 25 *'Creating value through responsible access to and use of big data'*. Within this route, especially research question 112 *'Can we use big data and big data collection to define values, generate insights, and get answers?'* and the detailed questions related to this topic will be addressed with a multidisciplinary focus. A number of questions from the NSA is central to our Ethical HI research line. A detailed overview of the questions related to cluster question 112 and other related questions are presented below. For each of the individual questions we indicate to which research line or research aspect it is related.

Dutch National Science Agenda research questions	**research line**
How can human insight be combined with AI to make better decisions?	Central research question
Which items can better be decided on by computers and which items can better be decided on by humans?	Collaborative HI
Can a computer write a relevant article/book?	Collaborative Science scenario
How can a computer decide which information is relevant and which information is not?	Adaptive HI
How can robots that participate in our society learn to communicate with us?	Explainable HI
What are the fundamentals of the disciplines of artificial intelligence (AI) and knowledge technology?	Central Research question
We make our computers more and more intelligent and allow them to make decisions more independently steeds. How smart can we make and do we want to make computers? And how can we guarantee that they will keep doing what humans want?	Responsible HI, Collaborative HI
How can data provide an answer to questions that have never been posed?	Adaptive HI, Explainable HI
How can a user of sensor generated data in the Internet of Things (IoT) know whether this data is to be trusted?	Collaborative HI, Responsible HI
Is the ethical decision process computational?	Responsible HI
Why do we not we succeed in integrating automated reasoning in natural language?	Explainable HI Adaptive HI

Questions related to other topics	
113 Can we develop human language technology (HLT) that allows us to communicate with our computers (smartphones, tablets)?	Collaborative HI, Explainable HI
123 How can we manage the unpredictability of complex networks and chaotic systems?	Adaptive HI
107 How can we anticipate the impact of new technologies on humans and society, and understand and evaluate the influence of existing technologies?	Responsible HI
070 Understanding our behaviour: why do we do what we do, are we who we are and which factors influence our behaviour?	Collaborative HI Explainable HI
108 Which social changes caused by technological changes will emerge and influence our wealth?	Responsible HI

Next to answering the above-mentioned questions, the program will also deliver methods, models and software components that will be enablers for other researchers, helping them to carry out their research or enabling them to solve complex problems. Potentially this will help solve many other questions in a wide range of areas (e.g. climate, water and ecosystems, environment, industry, medicine, urban planning, transport and traffic management).

2.6 Literature references

Angwin, J., J. Larson, S. Mattu and L. Kirchner (2016). "Machine bias." ProPublica.

Arkin, R. (2009). "Governing lethal behavior in autonomous robots." CRC Press.

Arslan, B., N. A. Taatgen and **R. Verbrugge** (2017). "Five-Year-Olds' Systematic Errors in Second-Order False Belief Tasks Are Due to First-Order Theory of Mind Strategy Selection." Front Psychol 8: 275.

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.

Aydogan, R., I. Marsa-Maestre, M. Klein and **C. M. Jonker** (2018). "A Machine Learning Approach for Mechanism Selection in Complex Negotiations." J. of Systems Science and Systems Engineering 27: 134-155.

Baarslag, T., M. Kaisers, E. H. Gerding, **C. M. Jonker** and J. Gratch (2017). "When will negotiation agents be able to represent us? The challenges and opportunities for autonomous negotiators.", IJCAI.

Balliet, D. (2010). "Communication and cooperation in social dilemmas: a meta-analytic review." J. of conflict resolution 54(1): 39-57.

Balliet, D., J. M. Tybur and P. A. M. Van Lange (2017). "Functional Interdependence Theory: An Evolutionary Account of Social Situations." Pers Soc Psychol Rev 21(4): 361-388.

Balliet, D., J. Wu and D. Dreu (2014). "Ingroup favoritism in cooperation: a meta-analysis." Psychological Bulletin 140(6): 1556.

Balliet, D. P., L. Mulder and P. A. M. van Lange (2011). "Reward, punishment, and cooperation: A meta-analysis." Psychological Bulletin 137: 594-615.

Balliet, D. P. and P. A. M. van Lange (2013). "Trust, conflict and cooperation: A meta-analysis." Psychological Bulletin 139(5): 1090-1112.

Barbaros, V., A. Abdolmaleki, **H. van Hoof** and D. Meger (2018). "Eager and memory-based non-parametric stochastic search methods for learning control." Int. Conf. Robotics and Automation.

Berendt, B. and S. Preibusch (2014). "Better decision support through exploratory discrimination-aware data mining: Foundations and empirical evidence." Artif Intell Law 22.

Bex, F. J., **H. Prakken**, T. Van Engers and **B. Verheij** (2017). "Introduction to the special issue on Artificial Intelligence for Justice." Artificial Intelligence and Law 25(1): 1-3.

Bilgic, M. and R. J. Mooney (2005). Explaining recommendations: Satisfaction vs. promotion. Workshop on the Next Stage of Recommender Systems Research, San Diego, CA.

Bonnemains, V., C. Saurel and C. Tessier (2016). How Ethical Frameworks Answer to Ethical Dilemmas: Towards a Formal Model. 1st Workshop on Ethics in the Design of Intelligent Agents at Eur. Conf. on AI.

Bosse, T. and **C. M. Jonker** (2005). "Human vs. computer behavior in multi-issue negotiation." Rational, Robust, and Secure Negotiation Mechanisms in Multi-Agent Systems: 11-24.

Bosse, T., **C. M. Jonker**, L. van der Meij and J. Treur (2008). "Automated Formal Analysis of Human Multi-Issue Negotiation Processes." Multiagent and Grid Systems 4(2): 213-233.

Boumans, R., F. Fokke van Meulen, **K. Hindriks**, **M. Neerincx** and M. O. Rikkert (2018). Do You Have Pain?: A Robot who Cares. ACM/IEEE Int. Conf. on Human-Robot Interaction (HRI '18).

Bradshaw, J. M., **V. Dignum**, **C. Jonker** and M. Sierhuis (2012). "Human-agent-robot teamwork." IEEE Intelligent Systems 27: 8-13.

Branting, L. K. (2017). "Data-centric and logic-based models for automated legal problem solving." Artificial Intelligence and Law 25(1): 5-27.

Burger, F., J. Broekens and **M. A. Neerincx** (2016). A Disclosure Intimacy Rating Scale for Child-Agent Interaction Book, Springer.

Burger, F., J. Broekens and **M. A. Neerincx** (2016). Fostering relatedness between children and virtual agents through reciprocal self-disclosure. In Benelux Conf. on Artificial Intelligence, Springer, Cham.

Burgoon, J. K. et al. (2017). Social signal processing, Cambridge University Press.

Card, S. K. (2017). "The psychology of human-computer interaction." CRC Press.

Caruana, R. (1997). "Multitask learning." Machine Learning 28(1): 41-75.

Coelho, H., P. Trigo and A. C. da Rocha Costa (2010). On the operationality of moral-sense decision making. 2nd Brazilian Workshop on Social Simulation.
Cointe, N., G. Bonnet and O. Boissier (2016). Ethical Judgment of Agents' Behaviors in Multi-Agent Systems. Proc. of the 2016 Int. Conf. on Autonomous Agents & Multiagent Systems, Singapore.
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.
Dastani, M., **D. Grossi**, J. J. C. Meyer and N. Tinnemeier (2009). Normative multi-agent programs and their logics. In Knowledge Representation for Agents and Multi-Agent Systems. Knowledge Representation for Agents and Multi-Agent Systems, Springer: 16-31.
Daume III, H. and D. Marcu (2006). "Domain adaptation for statistical classifiers." J. of Artificial Intelligence Research 26: 101-126.
de Filippi, P. and S. Hassan (2016). "Blockchain Technology as a Regulatory Technology: From Code is Law to Law is Code." First Monday 21(12).
de Greeff, J., T. Mioch, W. van Vught, **K. Hindriks, M. Neerincx** and I. Kruijff-Korbayová (2018). Persistent Robot-Assisted Disaster Response. ACM/IEEE Int. Conf. on Human-Robot Interaction.
de Martino, B., D. Kumaran, B. Seymour and R. J. Dolan (2006). "Frames, biases, and rational decision-making in the human brain." Science 313(5787): 684-687.
de Weerd, H., D. Diepgrond and **R. Verbrugge** (2018). "Estimating the use of higher-order theory of mind using computational agents." The BE J. of Theoretical Economics 18(2): 12.
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.
de Weerd, H., **R. Verbrugge** and **B. Verheij** (2017). "Negotiating with other minds: The role of recursive theory of mind in negotiation with incomplete information." AAMAS 31(2): 250–287.
Dennis, L. A., M. Fisher, M. Slavkovik and M. Webster (2016). "Formal verification of ethical choices in autonomous systems." Robotics and Autonomous Systems 77: 1-14.
Dietvorst, B. J., J. P. Simmons and C. Massey (2015). "Algorithm aversion: People erroneously avoid algorithms after seeing them err." J. of Experimental Psychology 144: 114-126.
Dignum, F., B. Dunin-Keplicz and **R. Verbrugge** (2001). Agent Theory for Team Formation by Dialogue. Intelligent Agents VII, Agent Theories Architectures and Languages, Springer: 150-166.
Dudzik, B., **H. Hung**, **M. A. Neerincx** and J. Broekens (2018). Artificial Empathic Memory: Enabling Media Technologies to Better Understand Subjective User Experience. In Proc. of ACM MultiMedia.
Dunin-Kęplicz, B., A. Strachocka and **R. Verbrugge** (2011). Deliberation dialogues during multi-agent planning. Int. Symp. on Methodologies for Intelligent Systems. Berlin, Heidelberg, Springer: 170-181.
Dunin-Keplicz, B. and **R. Verbrugge** (2003). "Evolution of collective commitment during teamwork." Fundamental Informatics 56(4): 329-371.
Dunin-Keplicz, B. and **R. Verbrugge** (2010). Teamwork in Multi-Agent Systems: A Formal Approach. Chichester, Wiley.
Dunin-Kęplicz, B. and **R. Verbrugge** (2006). Awareness as a vital ingredient of teamwork. AAMAS.
Efferson, C., R. Lalive and E. Fehr (2008). "The coevolution of cultural groups and ingroup favoritism." Science 321: 1844-1849.
Eggensperger, K., M. T. Lindauer, **H. H. Hoos**, F. Hutter and K. Leyton-Brown (2018). "Efficient Benchmarking of Algorithm Configuration Procedures via Model-Based Surrogates." Machine Learning 107: 15–41.
Eiben, A. E. and J. Smith (2015). "From evolutionary computation to the evolution of things." Nature 521(476).
Engelbart, D. C. (1962). Augmenting Human Intellect: A Conceptual Framework.
European Commission (2018). "Responsible Research and Innovationworkstream."https://ec.europa.eu/programmes/horizon2020/en/h2020-section/responsible-research-innovation
Flores, A. W., K. Bechtel and C. T. Lowenkamp (2016). "False Positives, False Negatives, and False Analyses: A Rejoinder to Machine Bias" Fed. Probation 80.
Floridi et al., (2004). "On the Morality of Artificial Agents." Minds and Machines 14: 349-379.
Fokkens, A. S., P. Vossen, M. Rospocher, R. Hoekstra and W. van Hage (2017). Grasp: grounded representation and source perspective. Proc. of *Knowledge Resources for the Socio-Economic Sciences and Humanities.*
Ganascia, J. G. (2007). "Modelling ethical rules of lying with Answer Set Programming." Ethics Inf Technol 9: 39.
Garcia, J. and F. A. Fernandez (2015). "A Comprehensive Survey on Safe Reinforcement Learning." J. of Machine Learning Research 16: 1437-1480.
Garmash, E. and **C. Monz** (2016). Ensemble Learning for Multi-Source Neural Machine Translation. Proc. of the 26th Int. Conf. on Computational Linguistics.
Gerpott, F. H., **D. Balliet**, S. Columbus, C. Molho and R. E. de Vries (2018). "How do people think about interdependence? A multidimensional model of subjective outcome interdependence." J Pers Soc Psychol 115(4): 716-742.
Ghader, H. and **C. Monz** (2017). What does Attention in Neural Machine Translation Pay Attention to? Proc. of the Int. Joint Conf. on Natural Language Processing.
Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu and D. Warde-Farley (2014). "Generative adversarial nets." Advances in neural information processing systems 2672-2680.
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.
Grosz, B. J. and L. Hunsberger (2006). "The dynamics of intention in collaborative activity." Cognitive Systems Research 7(2-3): 259-272.
Gunes, H. and **H. Hung** (2016). "Is automatic facial expression recognition of emotions coming to a dead end? The rise of the new kids on the block." Image and Vision Computing 55: 6-8.
Gunes, H., B. Schuller, M. Pantic and R. Cowie (2011). Emotion representation, analysis and synthesis in continuous space: A survey. Automatic Face & Gesture Recognition and Workshops. IEEE Int. Conf..
Guszcza, J. (2018). "Smarter together: Why artificial intelligence needs human-centered design." Deloitte Review(22).
Guttmann, C., **F. Dignum** and M. Georgeff (2011). Collaborative Agents – Research and Development, Springer.
Haasdijk, E., **A. E. Eiben** and A. F. T. Winfield (2013). Individual, Social and Evolutionary Adaptation in Collective Systems. Handbook of Collective Robotics: Fundamentals

and Challenges. S. Kernbach, Pan Stanford Publishers: 413-471.
Han, T. A., L. M. Pereira and F. C. Santos (2011). "Intention recognition promotes the emergence of cooperation." Adaptive Behavior 19(4): 264-279.
Harbers, M., R. Aydogan, **C. M. Jonker** and **M. A. Neerincx** (2014). Sharing Information in Teams: Giving Up Privacy or Compromising on Team Performance? Proc. of the 13th AAMAS, Paris.
Harbers et al. (2012). Explanation in human-agent teamwork. Coordination, Organizations, Institutions, and Norms in Agent System, Springer: 21-37.
Harbers, M., J. de Greeff, I. Kruijff-Korbayová, **M. A. Neerincx** and **K. V. Hindriks** (2017). Exploring the ethical landscape of robot-assisted Search and Rescue. World with Robots, Springer: 93-107.
Harbers, M. and **M. A. Neerincx** (2017). "Value sensitive design of a virtual assistant for workload harmonization in teams." Cognition, Technology and Work 19(2-3): 329-343.
Hayes, B. and J. A. Shah (2017). Improving robot controller transparency through autonomous policy explanation. Proc. of the 12th ACM/IEEE Int. Conf. on Human-Robot Interaction.
Heider, F. (1958). The Psychology of Interpersonal Relations. New York, Wiley.
Heinerman, J., E. Haasdijk and **A. E. Eiben** (2017). "Unsupervised identification and recognition of situations for high-dimensional sensori-motor streams." Neurocomputing 262: 90-107.
E.M. Helsper, **L.C. van der Gaag** (2007). Ontologies for probabilistic networks: a case study in the oesophageal-cancer domain. The Knowledge Engineering Review 22: 67 - - 86.
Hendricks et al., (2016). Generating visual explanations. Proc. of the European Conf. of Computer Vision.
Hendricks, L. A., R. Hu, T. Darrell and **Z. Akata** (2018). Grounding visual explanations. Proc. of the European Conf. of Computer Vision.
Henkemans, O. A. B., B. P. Bierman, J. Janssen, R. Looije, **M. A. Neerincx**, M. M. van Dooren and S. D. Huisman (2017). "Design and evaluation of a personal robot playing a self-management education game with children with diabetes type 1." Int. J. of Human-Computer Studies 106: 63-76.
Herlocker, J. L., J. A. Konstan and J. Riedl (2000). Explaining collaborative filtering recommendations. Proc. of the 2000 ACM Conf. on Computer supported cooperative work.
Hindriks, K., C. Jonker and D. Tykhonov (2008). "Towards an Open Negotiation Architecture for Heterogeneous Agents." Cooperative Information Agents XII 5180.
Hindriks, K. V. and **C. M. Jonker** (2009). Creating human-machine synergy in negotiation support systems: towards the pocket negotiator. Proc. of the 1st Int. Working Conf. on Human Factors and Computational Models in Negotiation. New York: 47-54.
Hofmann, K., L. Li and F. Radlinski (2016). "Online evaluation for information retrieval." Foundations and Trends in Information Retrieval 10(1): 1-117.
Hofstede, G. J., **C. M. Jonker** and T. Verwaart (2012). "Cultural differentiation of negotiating agents." Group Decision and Negotiation 21(1): 79-98.
Hong, L. and S. E. Page (2001). "Problem Solving by Heterogeneous Agents." J. of Economic Theory 97(1): 123-163.
Horsch, C. H. G., N. J. J. M. Smets, **M. A. Neerincx** and R. H. Cuijpers (2013). Comparing performance and situation awareness in USAR unit tasks in a virtual and real environment. 10th Int. Conf. on Information Systems for Crisis Response and Management, Baden Baden, Germany.
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.
Idrissou, A., R. Hoekstra, **F. V. Harmelen,** A. Khalili and P. V. D. Besselaar (2017). "Is my:sameAs the same as your:sameAs? Lenticular Lenses for Context Specific Identity." Int. Conf. Knowledge Capture.
Indurkhya, B., J. Hage and B. Broszek (2017). "Special issue on Machine Law." AI and Law 25(3).
Jennings, N. R., L. Moreau, D. Nicholson, S. Ramchurn, S. Roberts, T. Rodden and A. Rogers (2014). Human-agent collectives. Communications of the ACM 57: 80-88.
Johnson, M., J. M. Bradshaw, P. J. Feltovich, **C. M. Jonker**, **M. B. Van Riemsdijk** and M. Sierhuis (2014). "Coactive Design: Designing Support for Interdependence in Joint Activity." J. of Human-Robot Interaction 3(1): 43-69.
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.
Jonker, C. M., R. Aydoğan, T. Baarslag, J. Broekens, C. A. Detweiler, **K. V. Hindriks**, A. Huldtgren and W. Pasman (2017). An Introduction to the Pocket Negotiator: A General Purpose Negotiation Support System. Multi-Agent Systems and Agreement Technologies. EUMAS 2016, AT 2016. LNCS vol 10207.
Jonker, C. M., R. Aydogan, T. Baarslag, K. Fujita, T. Ito et al. (2017). Automated Negotiating Agents Competition. Proc. of the Thirty-First AAAI.
Jonker, C. M., **M. B. van Riemsdijk** and B. Vermeulen (2011). Shared mental models. Coordination, organizations, institutions, and norms in agent systems. Springer 6541: 132-151.
Jonker, C. M., J. L. Snoep, J. Treur, H. V. Westerhoff and W. C. A. Wijngaards (2010). "The Living Cell as a Multi-agent Organisation: A Compositional Organisation Model of Intracellular Dynamics." Trans. on Computational Collective Intelligence 6220: 160-206.
Jonker, C. M. and J. Treur (2002). "Compositional verification of multi-agent systems: a formal analysis of pro-activeness and reactiveness." Int. J. of Cooperative Information Systems 11: 51-91.
Jonker, C. M. and Treur, J (1997). Modelling an Agent's Mind and Matter. Multi-Agent Rationality and 8th European Workshop on Modelling Autonomous Agents in a Multi- Agent World, Springer.
Kambhampati, S. (2018). Challenges of Human-Aware AI Systems. AAAI 2018 Presidential Address.
Kaptein, F., J. Broekens, **K. Hindriks** and **M. Neerincx** (2017). The role of emotion in self-explanations by cognitive agents. In Affective Computing and Intelligent Interaction Workshops and Demos: 88-93.
Karpathy, et al. (2015). "Visualizing and understanding recurrent networks." preprint arXiv:1506.02078.
Kasparov, G. (2010). The Chess Master and the Computer. Chess Metaphors: Artificial Intelligence and the Human Mind. D. Rasskin-Gutman, MIT Press.
Kayhan, V. O. (2013). "Seeking Health Information on the Web: Positive Hypothesis Testing." Int. J. of Medical Informatics 82(4): 268-275.
Kelly, D. (2009). "Methods for evaluating interactive information retrieval systems with users." Foundations and Trends in Information Retrieval 3: 1-224.
Koeman, V. J., **K. V. Hindriks** and **C. M. Jonker** (2017a). "Designing a source-level debugger for cognitive agent

programs." Autonomous Agents and Multi-Agent Systems 31: 941-970.
Koeman, V. J., **K. V. Hindriks** and **C. M. Jonker** (2017b). Omniscient Debugging for Cognitive Agent Programs. Proc. of the 26th Int. Joint Conf. on Artificial Intelligence.
Koh, P. W. and P. Liang (2017). "Understanding black-box predictions via influence functions." arXiv preprint arXiv:1703.04730.
Koops, B. (2010). "Ten Dimensions of Technology Regulation-Finding Your Bearings in the Research Space of an Emerging Discipline." Dimensions of technology regulation: 309-324.
Kotthoff, L., C. Thornton, **H. H. Hoos** and et al. (2017). "Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA." J. Machine Learning Research 18: 1-25.
Kruijff-Korbayová, I., F. Colas, M. Gianni, F. Pirri, J. de Greeff, **K. Hindriks, M. A. Neerincx**, P. Ögren, T. Svoboda and R. Worst (2015). "TRADR project: Long-term human-robot teaming for robot assisted disaster response." Artificial Intelligence 29(2): 193-201.
Leenes, R. and F. Lucivero (2014). "Laws on robots, laws by robots, laws in robots: regulating robot behaviour by design." Law, Innovation and Technology 6: 193-220.
Looije, R., **M. A. Neerincx** and **K. V. Hindriks** (2017). "Specifying and testing the design rationale of social robots for behavior change in children." Cognitive Systems Research 43: 250-265.
Lorini, E. (2012). On the logical foundations of moral agency. Proc. of the Eleventh Int. Conf. on Deontic Logic in Computer Science Berlin, Springer: 108-122.
Mercado, J. E., M. A. Rupp, J. Y. Chen et al (2016). "Intelligent agent transparency in human–agent teaming for Multi-UxV management." Human Factors 58(3): 401-415.
Mioch, T., M. Peeters and **M. A. Neerincx** (2018). Improving Adaptive Human-Robot Cooperation through Work Agreements. 27th IEEE Int. Symp. on Robot and Human Interactive Communication.
Mioch, T., N. Smets and **M. A. Neerincx** (2012). Predicting performance and situation awareness of robot operators in complex situations by unit task tests. 5th Int. Conf. on Advances in Computer-Human Interactions.
Mitra, S. et al. (2007). "Gesture recognition: A survey." IEEE Trans. on Systems 37(3): 311-324.
Monitor Vrouwelijke Hoogleraren (2017), Landelijk Netwerk Vrouwelijke Hoogleraren, LNVH.
Neerincx, M. A., J. Van der Waa, F. Kaptein and J. Van Diggelen (2018). "Using perceptual and cognitive explanations for enhanced human-agent team performance." EPCE LNAI 10906: 204–214.
Ng, A. Y., D. Harada and S. Russell (1999). "Policy invariance under reward transformations: Theory and application to reward shaping." ICML 99: 278-287.
Ng, A. Y. and S. J. Russell (2000). "Algorithms for inverse reinforcement learning." ICML: 663-670.
Pan, S. J. and Q. Yang (2010). "A survey on transfer learning." IEEE Trans. on knowledge and data engineering 22(10): 1345-1359.
Park, D. H., L. A. Hendricks, **Z. Akata**, A. Rohrbach, B. Schiele, T. Darrell and M. Rohrbach (2018). Multimodal Explanations: Justifying Decisions and Pointing to the Evidence. Proc. of IEEE Conf. on Computer Vision and Pattern Recognition.
Partan, S. R. and P. Marler (2005). "Issues in the classification of multisensorycommunication signals." Am Nat 166: 231-245.
Plous, S. (1993). The psychology of judgment and decision making, Mcgraw-Hill Book Company.
Polevoy, G., M. de Weerdt and **C. M. Jonker** (2016). The Convergence of Reciprocation. Proc. of AAMAS.
Poria, S., E. Cambria, R. Bajpai and A. Hussain (2017). "A review of affective computing: From unimodal analysis to multimodal fusion." Information Fusion 37: 98-125.
Prakken, H. and G. Sartor (2015). "Law and logic: a review from an argumentation perspective." Artificial Intelligence 227: 214-245.
Ranjbar-Sahraei, B., H. B. Ammar, D. Bloembergen, K. Tuyls and G. Weiss (2014). Theory of Cooperation in Complex Social Networks. Proc. of the 25th AAAI Conf. on Artificial Intelligence (AAAI).
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.
Ren, Z., S. Liang, P. Li, S. Wang and **M. de Rijke** (2017). Social Collaborative Viewpoint Regression with Explainable Recommendations. Proc. of the Tenth ACM Int. Conf. on Web Search and Data Mining.
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.
Roijers, D. M., Vamplew, S. Whiteson and R. Richard Dazeley (2013). "Survey of Multi-Objective Sequential Decision-Making." J. of Artificial Intelligence Research 48: 67-113.
Roijers, D. M., L. M. Zintgraf and A. Nowé (2017). Interactive Thompson Sampling for Multi-Objective Multi-Armed Bandits. Proc. of the 5th Int. Conf. on Algorithmic Decision Theory.
Romano, A. and **D. Balliet** (2017). "Reciprocity Outperforms Conformity to Promote Cooperation." Psychological Science 28(10): 1490-1502.
Romano, A., **D. Balliet** and J. Wu (2017). "Unbounded indirect reciprocity: Is reputation-based cooperation bounded by group membership?" J. of Experiemental Social Psychology 71: 59-67.
Romano, A., **D. Balliet**, T. Yamagishi and J. H. Liu (2017). "Parochial trust and cooperation across 17 societies." Proc Natl Acad Sci U S A 114(48): 12702-12707.
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.
Rothmund, T., A. Mojzisch and S. Schulz-Hardt (2011). "Effects of Consensus Information and Task Demonstrability on Preference- Consistent Information Evaluation and Decision Quality in Group Decision Making." Basic and Applied Social Psychology 33(4): 382-390.
Sanderson, M. (2010). "Test Collection Based Evaluation of Information Retrieval Systems." Foundations and Trends in Information Retrieval 4(4): 247-375.
Schouten, D. G., F. Venneker, T. Bosse, **M. A. Neerincx** and A. H. Cremers (2018). "A digital coach that provides affective and social learning support to low-literate learners." IEEE Trans. on Learning Technologies 11(1): 67-80.
Segers, R., T. Caselli and **P. Vossen** (2018). The Circumstantial Event Ontology (CEO) and ECB+/CEO; an Ontology and Corpus for Implicit Causal Relations between Events. Proc. of the Language Resources and Evaluation Conf..

Serban, I. V., A. Sordoni, Y. Bengio et al (2016). Building end-to-end dialogue systems using generative hierarchical neural network models. AAAI.
Shamekhi et al. (2017). Augmenting group medical visits with conversational agents for stress management behavior change. Int. Conf. on Persuasive Technology: 55-67.
Sharchilev, B., Y. Ustinovsky, P. Serdyukov and **M. de Rijke** (2018). Finding influential training samples for gradient boosted decision trees. Int. Conf. on Machine Learning
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.
Sinha et al. (2018). Stackelberg Security Games: Looking Beyond a Decade of Success. IJCAI.
Sinha, R. and K. Swearingen (2002). The role of transparency in recommender systems. CHI '02 Extended Abstracts on Human Factors in Computing Systems.
Sycara, K., T. J. Norman, J. A. Giampapa, M. J. Kollingbaum, C. Burnett and D. Masato (2010). "Agent support for policy-driven collaborative mission planning." The Computer J. 53(5): 528-540.
ter Hoeve, M., M. Heruer, D. Odijk, A. Schuth, M. Spitters and **M. de Rijke** (2017). Do News Consumers Want Explanations for Personalized News Rankings? FATREC Worksh. on Responsible Recommendation.
ter Hoeve, M., A. Schuth, D. Odijk, **M. de Rijke**(2018). "Faithfully explaining rankings in a news recommender system." arXiv preprint arXiv:1805.05447.
Thornton, C., F. Hutter, **H. H. Hoos** and K. Leyton-Brown (2013). "Auto-WEKA: combined selection and hyperparameter optimization of classification algorithms." Proc. KDD: 847-855.
Timmer, S., J.-J. C. Meyer, **H. Prakken**, S. Renooij and **B. Verheij** (2017). "A two-phase method for extracting explanatory arguments from Bayesian networks." Int. J. of Approx. Reasoning 80: 475-494.
Ustun, B. and C. Rudin (2016). "Supersparse linear integer models for optimized medical scoring systems." Machine Learning 102(3): 349-391.
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.
van der Waa, J., M. Robeer, J. van Diggelen, M. Brinkhuis and **M. Neerincx** "Contrastive Explanations with Local Foil Trees." arXiv preprint arXiv:1806.07470.
van Doorn, J., D. Odijk, **D. M. Roijers** and **M. de Rijke** (2016). Balancing Relevance Criteria through Multi-Objective Optimization. Int. ACM SIGIR Conf. on Information Retrieval.
van Wissen, A., Y. A. Gal, B. A. Kamphorst and **M. V. Dignum** (2012). "Human–agent teamwork in dynamic environments." Computers in Human Behavior 28(1): 23-33.
Verbrugge, R. and L. Mol (2008). "Learning to apply theory of mind." J. of Logic, Language and Information 17(4): 489-511.
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).
Verheij, B. (2016). "Formalizing Value-Guided Argumentation for Ethical Systems Design." Artificial Intelligence and Law 24(4): 387-407.
Verheij, B. (2017). "Proof With and Without Probabilities. Correct Evidential Reasoning with Presumptive Arguments, Coherent Hypotheses and Degrees of Uncertainty. ." Artificial Intelligence and Law 25(1): 127-154.
Verheij, B., F. Bex, S. Timmer, Vlek, Meyer, Renooij and **H. Prakken** (2016). "Arguments, scenarios and probabilities: connections between three normative frameworks for evidential reasoning." Law, Probability and Risk 15: 35-70.
Vilalta, R. et al. (2002). "A perspective view and survey of meta-learning." AI Review 18(2): 77-95.
Vlek, C. S., **H. Prakken** and S. Renooij (2014). "Building Bayesian networks for legal evidence with narratives: a case study evaluation." Artif Intell Law 22: 375.
Vlek, C. S., **H. Prakken** and S. Renooij (2016). "A method for explaining Bayesian networks for legal evidence with scenarios." Artif Intell Law 24: 285.
Vossen, P., S. Baez, L. Bajčetić and B. Kraaijeveld (2018). "Leolani: a reference machine with a theory of mind for social communication." Preprint
Vossen, P., M. Postma and F. Ilievski (2018). Don't Annotate, but Validate: A Data-to-Text Method for Capturing Event Data. Proc. of the Language Resources and Evaluation Conf.
Wahlster, W. (2006). Dialogue Systems Go Multimodal: The SmartKom Experience. SmartKom: Foundations of Multimodal Dialogue Systems. Cognitive Technologies. . In: Wahlster W. (eds) SmartKom: Foundations of Multimodal Dialogue Systems. Cognitive Technologies. Springer..
Wallach, W. and C. Allen (2010). Moral machines : teaching robots right from wrong. Oxford Univ. Press.
Wilcke, X., P. Bloem and V. de Boer (2017). "The Knowledge Graph as the Default Data Model for Machine Learning." Data Science 1: 1-19.
Wu, J., **D. P. Balliet** and P. A. M. van Lange (2016). "Reputation management: Why and how gossip enhances generosity." Evolution and Human Behavior 37: 193-201.
Xu, J., J. Broekens, **K. Hindriks** and **M. A. Neerincx** (2014). Effects of bodily mood expression of a robotic teacher on students. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems: 2614-2620.
Xu, J., J. Broekens, **K. Hindriks** and **M. A. Neerincx** (2015). Effects of a robotic storyteller's moody gestures on storytelling perception. Int. Conf. on Affective Computing and Intelligent Interaction.
Xu, J., J. Broekens, **K. Hindriks** and **M. A. Neerincx** (2015). "Mood contagion of robot body language in human robot interaction." Autonomous Agents and Multi-Agent Systems 29(6): 1216–1248.
Zintgraf, L. M., **D. M. Roijers**, S. Linders and et al. (2018). Ordered Preference Elicitation Strategies for Supporting Multi-Objective Decision Making. Proc. of AAMAS.

2.7 Data management

The HI Centre will follow the below principles for data storage and management:

1. **Ownership of data**: Data is owned by the host university. Data is stored on the computer infrastructure of the host university. The host university policies for data (generation, storage and usage) will be adhered to with respect to ethical, legal and professional obligations and regulations. The policies of all participating universities ensure *open access* for scientific data, and a *commitment to long term archiving* for purposes of data reuse, experimental replication and data integrity. All universities comply with guidelines from VSNU and KNAW.
2. **One unified HI Centre data portal:** To facilitate access to data resources from the HI Centre, we create a unified HI Centre data portal that provides access to the different repositories of the individual universities. The coordinating partner VU has extensive experience in setting up such portals (www.openphacts.org).
3. **FAIR**: We implement the FAIR principles to ensure that data is Findable, Accessible, Interoperable and Reusable; metadata standards and archiving procedures will be developed in cooperation with data stewards at each university. Data will be open unless a strong case can be made otherwise, for example on the basis of privacy or prohibitive cost. Each case for non-open data will have to be brought to the attention of the project management.
4. **Acknowledgement and accessibility of data:** Data and code will be accessible to other parties based on the terms and conditions of the license policy of the *Apache Software Foundation licence version 2*. This license dictates that reuse of code and data is allowed, including for commercial purposes, under the condition that users and distributors acknowledge the contribution of the licensor.

Answers to specific questions requested by NWO:

1. **Will data be collected or generated that are suitable for reuse?** Yes. Experimental data will be collected both involving human subjects (collaboration experiments) as well as data involving only computing equipment. We expect many of these to be of interest for re-use by other scientists. For (re)using data the Research Data Management policies and the General Data Protection Regulation (EU regulation 2016/679) will be followed. Based on a Privacy Impact Assessment measures will be determined and included in the Data Management Plan (DMP). The DMP will be implemented at the start of the program. For the development and implementation cooperation will be sought with data stewards and the ICT support departments of the participating universities. Furthermore, the executive board will ensure that all participating scientists are aware of the principles set out above. This includes training for younger scientists in the relevant aspects of data management.
2. **Where will the data be stored during the research?** The data generated is stored on the servers of the participating universities. All have storage facilities for scientific data, facilities for daily and automatic backup, and policies and technology to protect sensitive data. In cooperation with data stewards and ICT support a dedicated research data infrastructure will be created which optimally facilitates the research.
3. **After the project has been completed, how will the data be stored for the long-term and made available for the use by third parties? To whom will the data be accessible?** All data, code and other results will be stored for at least 10 years using the FAIR principle on the servers of the participating universities. All of the universities have long term data archiving infrastructure for this purpose in place. We foresee the HI Centre to be a sustainable research centre after the end of the NWO funding period, but in any case, the HI Centre data portal will remain active for 10 years after the funding period to ensure findability, accessibility and reusability.
4. **Which facilities (ICT, (secure) archive, refrigerators or legal expertise) do you expect will be needed for the storage of data during the research and after the research? Are these available?** Research data will be stored on servers. ICT facilities are needed for access. During the project, we have budgeted for sufficient ICT infrastructure and human resources. For long term storage, infrastructure of the participating universities is available.

If the HI Centre is granted funding, a data management plan will be created before month 3 in which all aspects of data management will be described in more detail.

3. Researchers

3.1 Information on principal investigators

Name	m/f	Univ.	Expertise
Prof. Frank van Harmelen	M	VU	Knowledge Representation
Prof. Catholijn Jonker	F	TUD & UL	Interactive Intelligence
Prof. Rineke Verbrugge	F	UG	Logic and Cognition
Prof. Maarten de Rijke	M	UvA	Information Retrieval
Prof. Piek Vossen	M	VU	Natural Language Processing
Prof. Max Welling	M	UvA	Machine Learning

1.

3.2 Information on other participants

advanced	mid-career	early career

Name	m/f	Univ.	expertise
dr. Frank Dignum	M	UU	Social Multi-agent Systems
Prof. Linda van der Gaag	F	UU	Probabilistic Graphical Models
Prof. Guszti Eiben	M	VU	Adaptive Collective Systems
Prof. Henry Prakken	M	UU	Argumentation Models
Prof. Mark Neerincx	M	TUD	Human Centred Computing
Prof. Bart Verheij	M	UG	Argumentation Models
Prof. Holger Hoos	M	UL	Machine Learning
Prof. Koen Hindriks	M	VU	Social Robotics
dr. Stefan Schlobach	M	VU	Knowledge Representation
dr. Christof Monz	M	UvA	Machine Translation
dr. Virginia Dignum	F	TUD	Ethics of AI Systems
dr. Birna van Riemsdijk	F	TUD	Socially Adaptive Computing
dr. Dan Balliet	M	VU	Human Cooperation
dr. Davide Grossi	M	UG	Multi-agent Decision Making
dr. Hayley Hung	F	TUD	Multimodal Interaction
dr. Frans Oliehoek	M	TUD	Interactive Machine Learning

dr. Aimee van Wynsberghe	F	TUD	Ethics and Robotics
dr. Antske Fokkens	F	VU	Computational Linguistics
dr. Zeynep Akata	F	UvA	Explainable AI
dr. Herke van Hoof	M	UvA	Machine Learning for Robotics
dr. Diederik M. Roijers	M	VU	Multi-objective Reinforcement Learning

3.3 Plan for the career development of talented researchers in the middle level

With a span of 10 years, the career development of the future talent will be paramount to the long-term success of the HI Centre. In 2030, with some of the current (co-)PI's close to retirement, the current mid-career researchers will have taken leading roles in the team, and the first generation of Postdocs will be in positions to push the new Dutch and European research agenda in Hybrid and Artificial Intelligence. The HI Centre therefore has dedicated policies for supporting the career development of talented researchers at the various stages of their career from Postdoc, TT, Assistant Professor to Full Professor. These policies are in addition to the extensive education and training facilities for PhDs and Postdocs described in section 2.3.3.

3.3.1 Middle level career training and development

To ensure that future generations of Dutch researchers will maintain the leading role of the Netherlands in HI and AI research, it is essential that the younger generation of AI researchers in the Netherlands will be prepared to take over the roles of the current leaders toward the end of the project. We will invest substantial effort and resources in the development of talented researchers in the middle level (associate professors and assistant professors). The following actions will be carried out:

1. **Creation of tenure track positions:** Each participating university will create up to two tenure track positions during the project. There will be two timepoints for such tenure track appointment: at the program start (in years 1 and 2) and after the midterm evaluation (in years 5 and 6). This will be a significant injection towards the next generation of HI researchers. The tenure track positions may be a growth model for some of our own Postdocs, and of course the posts will be open to those outside the program, with the aim to attract diverse top talent to the HI program. Each new tenure track position will be allocated a PhD position to bootstrap their research in the HI program.

   Each university will make available up to two tenure track positions.

2. **Coaching and training facilities:** All facilities that the participating universities have for training assistant and associate professors are available to the HI Centre. This concerns regular coaching programmes and specific training activities to develop the skills relevant to the management of a research group. Courses include leadership, supervision, organization, individual growth and acquisition of funding.

3. **Intervision groups:** An intervision group will be developed in order to stimulate discussion and feedback between talented researchers. A workshop on communication and feedback skills will be provided by an external trainer and funding will be provided for meetings. Such intervision groups have proven to be useful development tools in the academic environment.

4. **Mentoring and sponsoring program:** A mentoring and sponsoring program will be created by assigning to each tenure tracker in the HI Centre a senior mentor at (co-)PI level of another university. Next to their mentoring and coaching activities they will also serve as a 'sponsor' to help further the career of the tenure tracker, using their international networks.

5. **Exchange program:** Funding has been allocated for an exchange program of talented researchers. Exchange activities will concern assignments of 6 months to universities participating in the FET Flagship and universities in the international network of the (co)PIs. (see also section 3.3.2)

6. **Practical training as vice-coordinator of a research line:** Within the HI Centre, mid-career researchers will take the role as vice-coordinator, offering them an excellent opportunity to train in management, research leadership and coordination. In the 2nd stage of the HI Centre, after 4-6 years, current early-career scientists will have progressed sufficiently in order to take on such vice-coordinator positions.

The financial consequences of the above-mentioned activities have been included in the budget. Funding is being made available for coaching, training and the exchange activities. Also, the costs for meetings with peers, coaches, mentors and sponsors have been foreseen in the budget.

3.3.2 Creating extra career opportunities

Bi-annual HI Centre conference: We will organize a bi-annual HI Centre conference. Here young and mid-level scientists present their work and meet international keynote speakers. At our conferences, we will ensure that diversity of the keynote speakers is always balanced to promote diverse role models to the next generation of HI researchers. Our scientific and societal advisory boards provide a rich source of such invited speakers. Budget (€50k) has been allocated for the organisation of these conferences.

Exchange program: We will establish exchanges with related programs, in particular the Human(e) AI FET Flagship program, and we will leverage existing international exchange agreements through Amsterdam Data Science with Columbia University in New York, Tsinghua University in Beijing and Ireland's National Data Science Research Centre. These exchange programs will promote and stimulate exchange and interaction of HI Centre researchers with international groups. PI's, co-applicants, Postdocs and PhD students can apply for funding to visit international centres of excellence and or laboratories with infrastructure, equipment or expertise, that are not available within the consortium. Additionally, counter-visits from other institutions will be utilised to scout potential new candidates for the PhD, Postdoc and tenure track positions

4. Budget

Total program costs

The HI Centre has set highly ambitious scientific goals and these can only be met if we allocate the majority of our funding towards outstanding, highly-innovative research. We envisage total costs of **€41,9M** of which **€19M** (45%) is provided by the NWO Zwaartekracht programme, and **€22.9M** (55%) by the participating universities. The co-finance provided by the participating universities covers overhead costs **(€6.9M),** in kind costs for personnel and facilities **(€ 9.1M)** and in-cash contributions **(€ 6.8M)** in the form of newly appointed staff dedicated to research in Hybrid Intelligence (Figure 12).

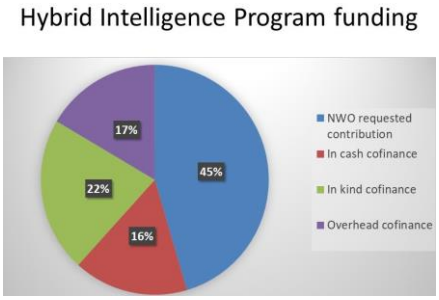


Figure 12: Hybrid Intelligence Program funding

Requested NWO budget and justification

The requested NWO contribution will cover the following cost categories (Figure 13):

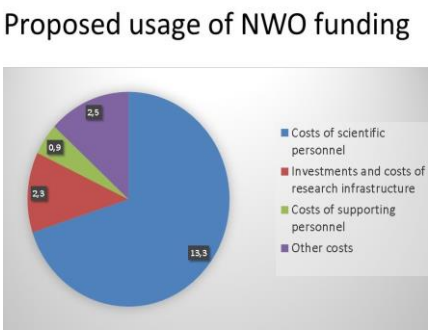


Figure 13: Proposed usage of NWO funding

- **Personnel costs for scientific personnel:** for the appointment of excellent young scientists (PhDs and Postdocs) and to allow for rapid scientific progress, and for effective training of the next generation of HI scientists in a stimulating environment most of the NWO budget will be used. In total 39 PhDs and 22 Postdoc positions will be created amounting to a budget of €13,3M. Within this budget each PhD and Postdoc student will have €18K to utilize towards attending conferences, training courses, workshops and subsistence. Personnel costs have been calculated based on NWO rates. The PhDs and Postdocs have been distributed across all four HI research lines with between 7,5 and 12,5 PhDs and 4-7 Postdocs allocated to each research line over the 10-year period.
- **Personnel costs for supporting personnel:** have been allocated at €0,89M, a modest 4% of the total budget. This will be utilized towards support capacity within and across the four research lines of the program in the form of a support team manager (0.3fte), a knowledge transfer and outreach officer (0.3fte), an education and training director (0.2fte), and a financial officer (in kind from the VU).
- **Joint infrastructure investments and costs:** include engineer costs (average 2.2fte for 10 years, both programming capacity and human interface experts; € 1.95M), computer servers and robots (€ 0.3M) amounting to a total €2,25M (12% of the NWO funding). Of the NWO contribution, €300K will be used for investments. These will be allocated starting from the first year in order to allow for maximal usage in the program:
  - Access to SURF-SARA cloud computing and storage facilities, as well as local computer facilities for processing and storage (€150K over 10-year period, taking into account a 4-year depreciation period). This will include software that we need from external parties (either commercial or academic).
  - Robot platforms for experimentation with embodied interaction, we will invest in acquiring state of the art humanoid robots. At the time of writing, this would be the Pepper Robot platform from Softbank, with which participating researchers at VU and Delft have ample experience. However, this choice may change during the course of the research programme in this highly dynamic market where new platforms are appearing rapidly. We will need at least 2 robot platforms simultaneously during the entire course of the research programme. With a depreciation period of 4 years, these are budgeted at a total of 150k€.

- **Other costs** include budgets for communication and dissemination, organization of trainings and symposia, the exchange program, experimentation fees for human subjects, advisory board costs, and open access publication fees. In total, other costs of €2.49M are estimated and divided among the participating organizations taking into account their contribution to the program and the relative weight of experiments. Specific cost items include:
    - A personal budget of 18k€ for each PhD or Postdoc position to cover conference participation, training and education, personal computing requirements and consumables;
    - Communication and dissemination costs (website and other material); total € 300k;
    - Open access publication fees: total €300k;
    - Costs of organising trainings, courses and symposia including costs of a bi-annual HI Centre conference; total: € 335k;
    - Costs of the Scientific and Societal Advisory Boards and the Professional User Panel: total €100k;
    - Fees for experimentations: € 155k;
    - €335K will be made available for applied projects to be executed by external parties based on the results of the HI Centre research, as selected in a national competition.

    These support costs will be supplemented by in-kind matching from all of the HI partners, for example for educational support.

Distribution of NWO funding (€19M) across the four research lines

The requested NWO contribution to the HI Centre has been allocated to the different partners according to their participation in the research program. We have taken into account the number of scientists involved in the HI Centre research program and the roles they have in the program. Individual partners have allocated these resources in view of their role in the various research lines and on the distribution between PhDs and Postdocs. The research budget covers the four research lines, with the research on Collaborative HI receiving slightly more than the other 3 lines because of its central role in the study of Hybrid Intelligence (30 % versus 23 %).