Project Strategy

Hybrid Intelligence Centre

1 Summary Hybrid Intelligence programme

Hybrid Intelligence (HI) is the combination of human and machine intelligence, expanding human intellect instead of replacing it. HI takes human expertise and intentionality into account when making meaningful decisions and performing appropriate actions, together with ethical, legal and societal values. Our goal is to design Hybrid Intelligent systems, an approach to Artificial Intelligence (AI) that puts humans at the centre, changing the course of the ongoing AI revolution.

Over the past decade, researchers in AI have made ground-breaking progress on long-standing problems. Now that AI is increasingly becoming part of our daily lives, we need to avoid being ruled by machines and their decisions. By providing intelligent artificial collaborators that interact with people, we strengthen our human capacity for learning, reasoning, decision making and problem solving. This interaction has the potential to amplify both human and machine intelligence by combining their complementary strengths. HI requires meaningful interaction between artificial intelligent agents and humans to negotiate and align goals, intentions and implications of actions.

The Hybrid Intelligence Centre is a collaboration of top AI researchers from the VU Amsterdam, the TU Delft, and the Universities of Amsterdam, Groningen, Leiden, Utrecht and Twente, in areas such as machine learning, knowledge representation, natural language understanding and generation, information retrieval, multi-agent systems, psychology, multi-modal interaction, social robotics, AI and law, and ethics of technology. The HI centre is creating a national and international focus point for research on all aspects of Hybrid Intelligent systems.

Developing HI requires fundamentally new solutions to core research problems in AI: current AI technology surpasses humans in many pattern recognition and machine learning tasks, however it falls short on general world knowledge, common sense reasoning, and human capabilities such as collaboration, adaptivity, responsibility and explainability (CARE).

2 HI Consortium Composition

The HI Consortium comprises the following partners:

- Vrije Universiteit Amsterdam
- Technische Universiteit Delft
- Universiteit Utrecht
- Universiteit Leiden
- Rijksuniversiteit Groningen
- Universiteit van Amsterdam
- Universiteit Twente

3 HI Vision

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 to optimize decision making and problem solving. This interaction has the potential to amplify both human and machine intelligence by combining their complementary strengths. HI focuses on the assistive and collaborative role of AI, emphasizing its potential to enhance human intelligence instead of replacing it.

4 HI research agenda

A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence IEEE Computer, Aug. 2020, vol. 53, pp. 18-28. Zeynep Akata, Dan Balliet, Maarten de Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, Hayley Hung, Catholijn Jonker, Christof Monz, Mark Neerincx, Frans Oliehoek, Henry Prakken, Stefan Schlobach, Linda van der Gaag, Frank van Harmelen, Herke van Hoof, Birna van Riemsdijk, Aimee van Wynsberghe, Rineke Verbrugge, Bart Verheij, Piek Vossen, Max Welling.

5 HI Impact Ambition

The HI Centre has the ambition to make both scientific impact and societal impact. The impact ambition has been set in the context that contemporary societies face problems that have a complexity, weight and scale novel to humanity: Maintaining democratic institutions, resource scarcity, environmental conservation, and climate change, to name a few. To solve these problems, humans need 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 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). On the other hand, people can add values, deeper contextual interpretation to data for machines for which this is more difficult. Instead, machines and humans need to work together through a collaborative conversation, where machines engage with us, explain their reasoning, and learn from their mistakes. 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.

We need to enhance the power of machine learning with the strength of human reasoning and the precision of automated reasoning. Hybrid Intelligence will allow organizations to innovate faster and more creatively, using understandable and trustworthy systems.

6 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.

7 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. 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.

8 HI objectives

The HI paper operationalised the overall HI research challenge ("How to build adaptive intelligent systems that augment rather than replace human intelligence, that leverage our strengths and compensate for our weaknesses?") into four objectives: (i) systems that collaborate, (ii) systems that adapt to changes in their environment, (iii) systems that can explain themselves, and (iv) systems behave responsibly according to awareness of norms and values. In our original proposal for each of the four objectives a Research Line was defined (Collaborative, Adaptive, Responsible, Explainable; CARE) and four "core capability tables".

Table 1: Collaborative HI Core capabilities

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

Table 2: Adaptive HI Core capabilities

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

Table 3: Responsible HI Core capabilities

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	Monologically assessing 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

Table 4: Explainable HI Core capabilities

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	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)

9 Levels of Ambition

We define how to evaluate ourselves at the end of the project by defining six levels of achievement (On Target (+1) vs. Off-target (-1), Exceeds expectations (+2) vs. Failing (-2), WOW! (+3) vs. Nightmare (-3)). These levels help define the overall ambition and design the related plans to achieve this ambition. Obviously, defining the plans and evaluating the project status is a continuous process, consequently, the Project Strategy and its related plans need to be revised regularly.

The annual Scientific Advisory Board (SAB) meetings, our annual reports and Mid-Term Self-evaluation reports are natural checkpoints to discuss and revise the Project Strategy. Currently, Spring 2023, we evaluate our programme to be On Target. We have satisfied the expectations for the type of grant we received by satisfing the following criteria:

- **Relevant and high quality scientific publications**: We have delivered scientific publications in high impact journals and conferences. These publications together show evidence of our success in filling the capability tables we defined in our project proposal.
- **HI community**: The high consortium consists of a well-connected participants and alumni which is characterized by their ongoing collaborations based on shared interests, problems and data.
- **HI Research field**: We have established the HI research field with seminal papers that define the field and we have established an internationa conference series. The Netherlands in the form of the HI Centre and its researchers are seen as leading this field.
- **Clear HI example**: We have provided as set of clear examples of Hybrid Intelligent Systems.
- **Impactful alumni**: Many of the first and second cohort of our PhD candidates and Postdocs have moved into positions of impact in academia and industry. Think of positions as assistant- or associate professors in academia and high-level consultants in industry. Similarly, those that counted as mid-career researchers when joining the consortium have moved up to senior positions.

In comparison, we would feel we would be Off Target (level -1), if:

- Our output is of scientifically speaking mediocre quality and perceived to be of mediocre relevance, e.g., we fail to satisfy even the lower levels of the capability tables.
- The HI Centre participants do not form ongoing collaborations.
- HI is not a recognized research field, neither are there clear and convincing examples of HI systems.
- The HI Centre alumni are not seen as bringing the expertise needed by society.

For now, we also believe that the HI consortium could strive for the highest level. To achieve the WOW factor we need to define what would be the criteria we believe we need to satisfy for that:

- **Sustainable HI community**: We have ensured the sustainability of the HI community after the duration of the grant according to one of the scenario's specified in Section 13.
- **HI textbook**: We have written a textbook on HI methodologies, HI measurements, HI technologies, and HI case studies that is referred to as a standard work, and used in education of BSc and/or MSc programmes on artificial intelligence.

- **HI Practical**: We have developed practical assignments to accompany the HI textbook and provide students with a clear experience of building their first HI system.
- **HI Materials**: We have a rich open-access repository of code and corpora that is used by both academia and industry to develop HI systems.
- **HI Workshops**: We have set up a series of annual or bi-annual workshops in major AI conferences on HI topics and we have an annual or bi-annual international HI conference.

We might not be able to achieve the WOW factor due to all sorts of circumstances. Yet, we believe that we would have exceeded expectations for the type of grant we received if we would have successfully managed to satisfy the following criteria:

- **Outside academia**: Our definition of Hybrid Intelligence is picked up not only by academia, but also by private- and public organisations. This is evident from examples, preferably both national and international.
- **Breakthrough show-case**: The HI Centre has become well-known internationally due to at least one show-case HI system.
- **Explainable to all stakeholders**: We have developed the notion of Hybrid Intelligence so well that its philosophy, definition and measurements can be explained to all stakeholders and we have demonstrably done so for many public and private organisations and on all levels of public administration.

We feel that we would have failed the expectations (level -2) for the type of grant we received if we would not be able to connect and the HI participants would remain fragmented delivering individual non-related projects. Moreover, we would also have failed if these projects consist of only computational experiments and we fail to communicate about these projects to the external stakeholders. Obviously, we are well aware of our nightmare scenario (level -3) for which we conducted a complete Quality Strategy to prevent ethical issues, data infringement and poor science reflected in a set of policies including the Ethics Policy, Communication Policy and Data Management Policy. Additionally, we make sure to take good care of our participants through an Education Strategy, our HI Diversity Statement and Supervision Policy.

10 Strategy for the next phase of the HI Centre

10.1 Management summary

1. **Resources**: We will roll out two further cohorts over the remaining 7 years of the HI Centre:
   - a cohort of 21 positions with a starting date early 2023, running through years 4-8 of the project.
   - a cohort of 39 positions with a starting date mid 2025, running through years 6-10 of the project.
2. **Commitment**: To improve involvement of our mid-career staff, they will be in the lead for writing the proposals for the second cohort.
3. **Continuity**: Projects in the second cohort do not by default continue work from projects in the first cohort, and collaborations in the second cohort are not by default the same as those in the first cohort, but are expected to explain how they relate to projects from the first cohort (even if not a simple continuation).

4. **Relevance**: We will require each project to commit to contributing to specific entries of the "hybrid intelligence capability tables", which specify different capabilities of hybrid intelligent systems at increasing levels of competence. We will position ourselves with respect to HI aspects of foundation models such as ChatGPT.

5. **Coherence**: We will improve coherence by organising the projects in the second cohort in clusters:

   - either around a combination of projects that can be empirically tested in newly installed HI lab
   - or in one of a small number of application scenarios, to be developed with external partners.

10.2 Status

The Hybrid Intelligence Centre is currently at the 2.5 year point out of a 10 year journey. All our PhD candidates and postdoc projects have now been running for at least 1.5-2 years. With this first cohort of projects, the Centre has spent just under half of its available budget (counting both NWO funding and matching positions "in cash" from the participating universities). Now is the right time to start thinking about how to move the project forward with the next cohort of projects.

This strategy has been prepared by the Centre's Management Team as a discussion document for the Executive Board, the Scientific Advisory Board and the Governing Board. We first list the success criteria we want to optimise in our second cohort, we then evaluate how we are currently doing on these criteria, we sketch the resources that are available, and based on all these, we sketch the strategy.

10.3 Success criteria

Our next round of internal funding should be aimed at maximising the following three criteria:

Relevance: Every new position that gets funded should be obviously relevant to the research question of the Hybrid Intelligence Centre: "What are hybrid intelligent systems?", "how do we build them?", and "how do we build them responsibly?". A concrete metric for relevance is the degree to which each project contributes to the tables of "core capabilities" for Hybrid Intelligence that we listed in our original proposal.

Coherence: The HI Centre aims to be more than a collection of individual projects, no matter how successful such individual projects might be.This ambition requires coherence among the individual projects, but the degree to which this coherence can or should be achieved is open for discussion. We list a number of options later in this document.

Continuity: We are proud of the successes of the individual project from our first cohort, and we certainly want to build on that in the second cohort. We want to ensure continuity on both content and collaborations, but this should not go at the cost of innovation. On content, we value continuation of the broad themes of the projects from the first cohort, but new projects should not simply be a continuation of existing projects. Similarly, it will be possible to continue very successful collaborations from the first cohort, but we don't expect that the new projects will simply be executed by the same teams as before.

10.4 How are we doing

Relevance: All of our projects are clearly relevant to the overall mission of the Centre. This is based on inspection of all the project kick-off documents and from their first year reports. However, although each of the individual projects are very well aware of, and relevant to, the overall goals of

the Centre, it is difficult to relate them to specific core HI capabilities from the tables at the end of this document, much less to specific levels of those capabilities. We aim to improve this with the next cohort of projects. We will position ourselves with respect to HI aspects of foundation models such as ChatGPT.

Coherence Through the bi-weekly meetings of the PhD candidates, through the 4-6-weekly meetings in the Research Lines and through our 6 consortium meetings, the PhD candidates and postdocs are well aware of each other's projects. However, this has not yet led to concrete collaborations between these individual projects, for example with projects using each other's results, or working towards a joint result. An exception to this is the collaboration between projects 4 and 18, who have embarked on a large-scale experiment on partner choice in collaborative settings. At the level of supervisors of the projects (both senior and mid- career researchers), the joint supervision of PhD candidates creates good coherence inside a project, but awareness of work in other projects is limited to those in their research line. In particular, coherence between the mid-career staff members of the Centre is lower than we want it to be. By encouraging mid-career people to co-supervise projects in the second cohort, we aim to improve the coherence. Again, we aim to improve this coherence with the next cohort of projects.

10.5 What resources do we have available

We still have funding available for 60 positions for the remaining 7 years of the project, 40 from NWO funding and 20 from "cash" matching from the participating universities. We propose to roll out these positions in two further "cohorts" over the remaining 7 years:

- Our first cohort was for 27 positions, with a starting date late 2020, so these projects are running through years 1-4 of the project.
- A second cohort of 21 positions will have a starting date early 2023, which means it will run through years 4-8 of the project.
- A final third cohort of 39 positions will have a starting date mid 2025, which means it will run through years 6- 10 of the project.

This timeline means that the final projects are projected to end in month 6 of our final year. The remaining 6 months will be necessary to absorb delays in recruiting, and for projects that get extended because of an internship, allowing us to wrap up at or soon after the end of year 10.

We have chosen to make the second cohort smaller than the first and the third so as not to overstretch the capacity for supervision and for overall project management. The distribution between PhD and postdocs positions originally planned in the proposal was divided across the three cohorts. However, since a 4-year PhD or a 3-year postdoc position have the same financial budget, we can adjust this distribution based on the interests of the Centre, of the individual project, and of the research group hosting the project. The same figure also shows that there is a large divergence in size between partners, with available positions ranging from 20 to 4. This can to some extent be evened out by involvement as co-supervisors, but it might also mean that some of the smaller partners are perhaps better off by not participating in the second cohort.

There is a technical issue with the NWO funding from 2024 onwards being conditional on a positive midterm review. We have approval from our governing board to plan for a positive outcome of the midterm review.

We will postpone any detailed choices for the third cohort until closer to the time. We will discuss here important choices for the 2nd cohort, since recruiting for this cohort will take place in 2023.

We considered three possible models for **coherence**, of which we adopt an intermediate model:

- A low coherence model would take the approach of "many blossoming flowers". Individual projects will no doubt achieve good results, and spontaneous collaborations may form, but no coherence is planned for. However, our ambitions for the HI Centre go beyond a large collection of good PhD candidate projects.
- A strict coherence model would aim to build a single coherent platform to which every individual project contributes through software, theories, or experiments. We think the scale of the HI Centre, both in terms of number of researchers, running time and diversity of research disciplines, is too large to make this a feasible option.
- We therefore propose an intermediate coherence model in the form of clusters.

  Clusters can be formed around either a shared experiment, or an application:

  1. We will install an HI lab, where coherent combinations of project results can be tested empirically, following the entries in the capabilities tables.
  2. We will initiate a small number of application scenarios, to be developed with external partners where a number of different projects contribute to the construction of a demonstrator. Two concrete candidates for such applications that are currently on the table are prevention in public healthcare and robotic surgery.

To increase the **commitment** of our mid- career staff to the Centre, that they will take the lead in writing the proposals for the clusters and projects of the second cohort. This will increase the "ownership" that they feel for these projects, and for the Centre as a whole.

For **relevance**, each of the projects in the second cohort commits to contributing to a specific row (or even a specific cell) in one of the **capability tables** listed in this document. Of course, the specific content of these tables will change as our understanding progresses, but the instrument of the capabilities and levels of competence for these capabilities will become more central than it was for the first cohort.

11 National Strategy

We are extending our alliance with TNO into a signed partnership and expect to extend the current collaboration on Diabetes II management to other case studies with third parties who collaborate with TNO. We initiated a collaboration with Dr. Dalibor Vasilic, a plastic surgeon specialising in microsurgery at the Erasmus medical centre in Rotterdam. He uses a robot assistant for training and operational purposes, and we will work with him and his team to increase the collaborative capabilities of that robot assistant in a pre-operation training setting. We will initiate a collaboration with the Kurt Lewin Institute, a graduate school for Social Psychology, to expand the training opportunities for our PhD candidates and to reach out to the young generation of social psychology researchers. We will establish collaborations with industrial and governmental organisations through ICAI labs, LTP projects, and case studies through our Taskforce. This will strengthen our potential for securing second and third money stream projects and provide an ecosystem for hybrid intelligence that is open for BSc, MSc and PhD candidates for internships. Strategic collaborations with other consortia will be established to strengthen our scientific basis in disciplines such as the social sciences and life sciences. A particularly relevant candidate is the Gravitation-funded consortium ALGOSOC, of which three of our scientists are members.

12 International Strategy

We have established the Hybrid Human-Artificial Intelligence (HHAI) conferences as an annual international conference series. The second edition (2023) is co-chaired by dr. Tiddi from the HI consortium and is hosted by DFKI in Munich. The HHAI 2024 edition has already been secured. All this shows that the HHAI conference series is a fruitful new addition to the international HI research community. Profs. Virginia and Frank Dignum, co-founders of the Hybrid Intelligence Center, have moved to Umeå University, where they are leading a large Swedish initiative funded by the Wallenberg foundation, which is closely related to Hybrid Intelligence. We will set up an exchange program with this initiative for PhD candidates, PDs and staff members. They have agreed to organise HHAI2024. We aim to engage with new and closely related initiatives in Europe (in Oxford, in Aarhus, in St. Gallen, in Bielefeld, in Oulu) and the US (Microsoft, Stanford). A longer term option that we are considering is to establish a pan-European HI network of academic and industrial partners. Possible partners for this are the Humane AI Net programme, its requested successor proposal in which we play a role, as well as the aforementioned Swedish programme.

13 HI Sustainability Plan

The HI sustainability plan has been determined by the Executive Board. The EB foresees a number of possible scenario's to sustain the HI Centre after the end of the subsidy period.

- Scenario 1: HI is the way, everybody researches AI as HI, HI Centre dissolves, Universities take up HI output
- Scenario 2: HI training centre - paid courses, grants
- Scenario 3: HI internationally renowned institute funded by public money, co-funded by industry with EU values
- Scenario 4: HI Centre is one of the leading European institutes on Human-Centred AI and is (co-)founded with EU money.
- Scenario 5: The HI Centre is transformed into a network of scientists and institutions, comparable with the ELLIS network.