Engineering Hybrid Intelligence Systems: Synthesising a Decade of Research

ILARIA TIDDI✉ and CHIARA NATALI✉, Vrije Universiteit Amsterdam, The Netherlands
HERKE VAN HOOF, Universiteit van Amsterdam, The Netherlands
SHENGHUI WANG, University of Twente, The Netherlands
MAAIKE DE BOER, TNO, The Netherlands
VICTOR DE BOER, Vrije Universiteit Amsterdam, The Netherlands
ANTSKE FOKKENS, Vrije Universiteit Amsterdam, The Netherlands
CATHOLIJN JONKER, Technische Universiteit Delft, The Netherlands
MICHIEL VAN DER MEER, Universiteit Leiden, The Netherlands
PRADEEP MURUKANNAIAH, Technische Universiteit Delft, The Netherlands
NANNE VAN NOORD, Universiteit van Amsterdam, The Netherlands
FRANS OLIEHOEK, Technische Universiteit Delft, The Netherlands
ROXANA RADULESCU, Utrecht University, The Netherlands
MERLE REIMANN, Vrije Universiteit Amsterdam, The Netherlands
COR STEGING, University of Groeningen, The Netherlands
SHIHAN WANG, Utrecht University, The Netherlands

Hybrid Intelligence (HI) is a rapidly growing discipline that envisions the creation of scenarios where humans and agents operate in mixed teams collaboratively, synergistically, and proactively to achieve a given goal. This ambition was formulated in 2019 and followed up by a comprehensive research agenda, proposing that a more effective human-machine collaboration could be promoted based on four foundational requirements: Collaborativeness, Adaptiveness, Responsibility, and Explainability (the CARE framework). Approaching the decade since the appearance of the first research outputs, what is missing is a comprehensive overview of how the HI discipline has evolved. This paper attempt to fill this gap by providing a reflection and overview of the HI field. We analyse approximately 350 papers published by the Hybrid Intelligence Consortium in which the seminal HI research agenda originated, and its closest collaboration circle. Using the CARE principles as an analytical framework, we investigate how the original HI principles

Authors' Contact Information: Ilaria Tiddi, i.tiddi@vu.nl; Chiara Natali, c.natali@vu.nl, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands; Herke van Hoof, Universiteit van Amsterdam, Amsterdam, The Netherlands, h.c.vanhoof@uva.nl; Shenghui Wang, University of Twente, Entschede, The Netherlands, shenghui.wang@utwente.nl; Maaike de Boer, TNO, Den Hague, The Netherlands, maaike.deboer@tno.nl; Victor de Boer, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands, v.de.boer@vu.nl; Antske Fokkens, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands, antske.fokkens@vu.nl; Catholijn Jonker, Technische Universiteit Delft, Delft, The Netherlands, c.m.jonker@tudelft.nl; Michiel van der Meer, Universiteit Leiden, Leiden, The Netherlands, m.t.van.der.meer@liacs.leidenuniv.nl; Pradeep Murukannaiah, Technische Universiteit Delft, Delft, The Netherlands, p.k.Murukannaiah@tudelft.nl; Nanne van Noord, Universiteit van Amsterdam, Amsterdam, The Netherlands, n.j.e.vannoord@uva.nl; Frans Oliehoek, Technische Universiteit Delft, Delft, The Netherlands, f.a.oliehoek@tudelft.nl; Roxana Radulescu, Utrecht University, Utrecht, The Netherlands, r.t.radulescu@uu.nl; Merle Reimann, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands, m.m.reimann@vu.nl; Cor Steging, University of Groeningen, Groeningen, The Netherlands, c.c.steging@rug.nl; Shihan Wang, Utrecht University, Utrecht, The Netherlands, s.wang2@uu.nl.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.
Manuscript submitted to ACM

turned into practice, and which building blocks are needed when creating Hybrid Intelligence systems. Our contribution is a conceptual and practical view of how the HI field has evolved, targeting especially researchers who start to investigate Hybrid Intelligence.

CCS Concepts: • **Human-centered computing** → **Collaborative and social computing**; • **Computing methodologies** → **Distributed artificial intelligence**.

ACM Reference Format:

Ilaria Tiddi, Chiara Natali, Herke van Hoof, Shenghui Wang, Maaike de Boer, Victor de Boer, Antske Fokkens, Catholijn Jonker, Michiel van der Meer, Pradeep Murukannaiah, Nanne van Noord, Frans Oliehoek, Roxana Radulescu, Merle Reimann, Cor Steging, and Shihan Wang. 2018. Engineering Hybrid Intelligence Systems: Synthesising a Decade of Research. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 37 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Humans increasingly work with Artificial Intelligence (AI) in contexts such as healthcare, education, and public decision-making, where judgment, contextual awareness, and coordinated decision-making is required [175]. Yet, the extent to which these systems can participate meaningfully in real-world activities characterised by uncertainty, conflicting goals, and social complexity is limited [9]. As a result, systems are increasingly designed as hybrid configurations in which human and artificial agents combine strengths and complement their weaknesses [30]. This vision is at the core of Hybrid Intelligence (HI), a rapidly growing research discipline at the crossing between AI, Behavioural and Organisational Psychology, and Human-Computer Interaction.

At its origin, Hybrid Intelligence envisions scenarios where humans and agents operate in mixed teams collaboratively, synergistically, and proactively to achieve a given goal (the so-called "HI-system"). In this ideal view, human limitations such as stereotypes, error proneness, in-group favouritism or short memory would be mitigated by machine-driven decision-making, while the human feedback would ensure that the artificial agent exhibits fairness, sample efficiency and task generalisability [168]. To achieve this, a comprehensive research agenda for HI was formulated in 2019 [5, 30], proposing four foundational requirements, or challenges, which together provide the conditions to promote human augmentation rather than substitution: Collaborativeness, Adaptiveness, Responsibility, and Explainability (later referred to as the CARE framework). A number of solutions were also promoted as a starting point including a.o. psychological theories of cooperation and theory of mind to improve synergetic teamwork, deep informed reinforcement learning to enable contextual sensitivity in machines that would adapt to humans and their environment, societal and personal value-aware methodologies to design socio-technical systems, and neuro-symbolic combinations where ontologies and knowledge graphs would provide structured background knowledge to neural methods to better share and explain awareness, goals, strategies and decisions.

Since the publication of its agenda, HI has grown immensely, seeing the establishment of several research centres of excellence (e.g. centres for Hybrid Intelligence in the Netherlands[1], Denmark[2], Finland[3], Italy and UK[4], the European Network for Human-Centred Artificial Intelligence[5], the Human-Centred AI research group[6]) and events such as the nowadays well-established Hybrid Human-Artificial Intelligence Conference (HHAI[7]). While we approach the decade

[1]https://www.hybrid-intelligence-centre.nl/
[2]https://mgmt.au.dk/center-for-hybrid-intelligence
[3]https://www.oulu.fi/en/research/human-ai-co-evolution-and-strengthening-human-capabilities
[4]https://tango-horizon.eu/
[5]https://www.humane-ai.eu/
[6]https://hcai.site/
[7]hhai-conference.org

since the appearance of the first research outputs, a comprehensive overview of how the HI discipline has progressed and evolved with respect to its original vision is still missing. Additionally, there are varying opinions on how and to which extent AI should be a team member in human-AI collaboration [155]. This lack of harmonised view makes it hard to design, build, compare and reuse HI methods and solutions, ultimately hampering the deployment of HI systems that meet user needs in the real-world. In other words, we miss a standardised way to create, or *engineer*, HI systems.

This paper is an attempt to fill this gap, and to see where this cycle from vision to implementation is all going. In order to do so, we need to understand (a) how the HI challenges from the original agenda have been tackled in terms of methods and solutions, and (b) what the recurrent characteristics and elements are that can be used to build HI systems. We therefore investigate the following research questions: (RQ1/*Evolution*) How have the original HI design principles been operationalised since the original HI research agenda? (RQ2*Anatomy*) What are the building blocks that we need to create an HI system? To answer these questions, we present an analysis of the evolution of the Hybrid Intelligence discipline since its early conceptualisation. Our goal is not to capture the broader AI, HCI or Human-Centred AI landscape, but rather to provide a snapshot of where the HI research stands while approaching the decade, in terms of main research questions, proposed solutions, exemplary applications, and future outlook. We do this by building on the case-study of the Dutch Hybrid Intelligence Centre associated with the seminal HI research agenda. Using the CARE framework as an epistemological analytical framework, we analysed approx. 350 publications produced by the research consortium and its closest collaboration circle, documenting their research questions, the methods they developed and the results obtained when answering these questions[8]. These findings are summarised to answer how the original HI principles turned into practice (RQ1/Evolution, Section 3), and further contribute to identifying the core building blocks to use when creating HI systems (RQ2/Anatomy, Section 4). Finally, we identify and discuss the key fundamental challenges still open for future research to tackle (Section 5).

Our contribution is both a conceptual and practical view of how the HI field evolved. From the conceptual perspective, we discuss and provide input to the key question that has engaged the research community since its start: what are the characteristics of Hybrid Intelligent systems compared to other types of (AI) systems? From a practical perspective, we provide the design blocks to build HI systems, including pointers to the approaches and technologies that might be used to implement them, and concrete examples of HI applications. Our aim is that this material will support researchers starting their research on HI to shape the future of the field and the community.

2 Terminology

We introduce here the main terminology of our work, based on the key concepts envisioned by seminal HI papers [5, 30].

2.1 Conceptual Background

Intelligence. Hybrid Intelligence emerged as early as in cybernetics [31, 104], whose account of intelligent behaviour moved away from metaphysical accounts toward a functional understanding that could apply equally to biological and artificial systems. Rosentblueth e.a. [150] described intelligence as purposive, goal-directed behaviour maintained through feedback mechanisms. Intelligence therefore resides not solely within individual components but emerges through their interactions and capacity for self-regulation. In [30], intelligence is characterised as "the ability accomplish complex goals, learn, reason, and adaptively perform effective actions within an environment", implicitly linking it back to the purposive behaviour tradition of cybernetics. Moving beyond the necessarily reductive cybernetic perspective,

[8] cfr. Appendix for the methodology details.

a distinction between human and artificial intelligence is proposed in terms of complementary strengths: human intelligence is inherently intuitive, contributing flexibility, empathy, creativity, common sense; machine intelligence is analytic, excelling in pattern recognition, probabilistic reasoning, consistency, speed, efficiency.

System. Understanding HI also requires adopting a systems perspective. The term *system*, derived from the Greek for "putting together", refers to an organised whole whose elements are integrated into a cohesive structure [19]. This related to the General Systems Theory [191], where systems are complex wholes whose properties emerge from their organisation as a unified entity. The idea of socio-technical system [172] complements this by representing the interplay between social and technical components. In the HI context, this means designing AI agents not as standalone entities, but as integral parts of a larger, interconnected system, where humans and machines collaborate with a purpose and based on shared values, beyond the capabilities of either alone.

Augmentation. The idea of augmenting humans traces back to the *man–computer symbiosis* envisioned as a tightly coupled partnership [94] where humans would provide goal-setting, judgement, contextual interpretation, and flexibility, while machines would contribute speed, memory, and computational precision. In the 1960s, research in Human-Computer Interaction explicitly positioned computing as a means of "increasing the capability of a man to approach a complex problem situation, to gain comprehension to suit his particular needs, and to derive solutions to problems" [41]. This vision was epistemic and cognitive: computational systems should amplify the human "feel for a situation" – the capacity to navigate complexity, ambiguity, and uncertainty (ibid.). This augmentationist tradition remains today in concepts such as Augmented Intelligence, Cognitive Augmentation, and Machine-Augmented Intelligence, that similarly emphasise the use of AI to improve human decision-making, learning, and problem-solving rather than substitute for them [7]. Similarly, Human-Centred AI (HCAI) [155] prioritises systems that preserve meaningful human control while amplifying human capabilities.

Hybrid. Hybrid Intelligence can be seen as a specific form of collective intelligence, where groups of individuals act collectively "in ways that seem intelligent" [110]. This typically refers to homogeneous groups (e.g. humans or animals), while HI introduces the concept of heterogeneous agents [5, 30], namely humans and AI agents constituting together a HI system. This perspective was formalised in 2019 as the HI research agenda [5], bringing forward the idea that human and machine intelligence could be combined to achieve synergistic, proactive collaborations in mixed team, with the purpose to achieve shared goals that were unreachable by either humans or machines alone.

What distinguishes HI is its attempt to consolidate these traditions toward designing human-AI systems in which intelligence emerges through complementarity, coordination, and mutual adaptation. In contrast to automation-centred paradigms, HI is fundamentally concerned with the conditions under which human and artificial agents can enable augmentation in Human-AI systems in which the whole becomes more capable than the sum of its parts.

2.2 The CARE framework

The CARE framework formulates the main research questions that should be faced by Hybrid Intelligence research [5].

Collaborativeness. Within the broader HI agenda, the line of research tackling collaborativeness explores how humans and artificial agents can work together synergistically, combining their complementary strengths to achieve outcomes beyond the capabilities of either alone. This involves investigating how collaborative processes should be modelled, how agents should reason about humans, and how hybrid teams can coordinate actions and decisions. By

developing intelligent agents that work in synergy with humans, i.e. aware of each other's strengths and limitations, HI aims to amplify both human and machine intelligence through meaningful interaction, aligning goals, intentions, and the implications of actions. The central challenge, then, is not just to design systems that can collaborate, but to ensure that collaboration is meaningful, measurable, and enacted in practice.

Adaptiveness. Adaptive HI investigates how HI systems learn from and adapt to humans and their environment. First of all, agents in a HI team might need to learn to properly interact, through e.g. dialogue or non-verbal means. Interaction can also be leveraged to learn more efficiently. Changes in the composition of a team where an artificial agent operates might require rapid adaptation to shifting team preferences and needs in real time. This might be done by actually adjusting the learning mechanism itself. Lastly, users might express their knowledge, needs or (ethical, legal, safety) constraints most naturally in symbolic expressions, so learning systems might need to be conceived to use such expressions. As such, the adaptive research line aims to bolster the capacity of intelligent agents to work in mixed teams by making them more versatile and responsive.

Responsibility. Responsible HI investigates how ethical, legal, and societal (ELS) considerations can be integrated into systems where humans and AI collaborate. This involves examining how ELS principles and trustworthiness can be embedded into HI systems, ensured throughout their development and lifecycle, and evaluated against relevant requirements. Addressing the perceived risks of AI technologies requires ethical and legal concerns to be central to the design and operation of HI systems. Values such as transparency, accountability, trust, privacy, and fairness can no longer be treated as afterthoughts or relegated to post-deployment regulations. HI research aims to develop methods for incorporating ELS considerations into both the design process (ethics in design) and the performance (ethics by design) of HI systems, ensuring that responsibility is a foundational pillar of their development and deployment.

Explainability. Explainability in HI concerns the development of transparent and interpretable artificial agents to improve shared understanding and trust within the collaborating team, which are well-studied both in social sciences and AI. Shared representations may lead to a stable model to be leveraged for effective explanations, prediction and control. The challenge lies in finding representations that can faithfully capture what caused an agent to provide a given output that is accessible to people. Only in this way can humans gain a good understanding of the workings of the agent. An important open challenge in designing such explanations is that we need a better understanding of how people interpret and interact with explanations coming from machines, notably w.r.t. people's ability to evaluate their trustworthiness. Inversely, there is the challenge of facilitating ways in which humans can provide machine interpretable explanations as to steer and improve the agent's outcome.

3 Evolution of HI Systems

Here we analyse how the HI principles have been operationalised in terms of typical methods and enabling technologies in the HI literature. For a better organisation of this section, each subsection below discusses one of the four CARE lines, introducing first the research question it addresses, and then an organised summary of the works that contributed to answering the question.

3.1 Collaborative HI Systems

Collaborative HI systems focus on the synergistic collaboration of different actors that complement each other. Collaborating agents vary in number and form, e.g. human-agent(s), human-human, or agent(s)-agent(s) actors, and their

communication ranges from asynchronous unimodal communication all the way to fully synchronous multimodal exchanges. The main research questions addressed by this area of research are (RQ1) How to characterise collaboration and synergy in HI systems?,(RQ2) How to design HI systems that support effective collaboration?, and (RQ3) How can HI systems interpret, represent, and generate multimodal collaborative interaction?

3.1.1 *Characteristics of collaboration and synergy in Hybrid Intelligence.*

Partner Modelling. To characterise collaboration and synergy in HI systems, studies of human social interaction provide essential grounding [116]. [209] explores Bayesian reasoning as a solution to model a partner's beliefs, while modelling affinity alongside beliefs further eases selections and negotiation [44]. [115] combines social science theory with social signal processing to understand, model and predict partner selection, with the aim of informing HI systems that support initiating and facilitating collaborations. Partner selection is not fixed but sensitive to the task: across four studies, people generally prefer warm partners, yet shift towards competence-related traits when the task requires them [63]. The ConfLab dataset [142] encodes these findings into a rich multimodal corpus of partner selection strategies.

Wisdom of the Crowd. At a theoretical level, work on social choice and multi-agent rationality [101, 102] asks what it means for artificial agents to be socially rational. [101] shows that learning to cooperate with adaptive partners is fundamentally hard without additional assumptions, underscoring that collaboration is not simply a matter of optimising individual reward, but requires a shared understanding of goals and mutual expectations. A practical example here is the Condorcet Markets framework [4], that demonstrates how aggregating individual intelligence can produce collectively rational outcomes when establishing equivalences between information markets and collective decision rules. Similarly, jury theorems with costly information show that effort and incentives can dissolve crowd wisdom, though network structure may mitigate this [125], while weighted delegation in liquid democracy can be arranged to optimise the group's truth-tracking ability [202]. The structure of social networks itself can distort perception of the majority opinion, producing "majority illusions", that can be prevented by limiting voting rules [187]. The AI4GCC climate negotiation challenge [199] shows how a negotiation protocol with shared mitigation targets promotes cooperation among self-interested parties.

Embracing the Hybrid Collaboration. Collaboration also requires acknowledging the diversity and complementarity of human and artificial participants [148, 165]. [49] envisions a human-embodied agent synergy where collaborative competence and trustworthiness are incrementally accrued through mutually beneficial interactions. HyEnA (Hybrid Extraction of Arguments) [119, 181] operationalises this idea as a hybrid human-AI pipeline for argument mining that outperforms both fully automated and fully manual approaches, illustrating the complementarity of human judgement and machine speed. Similar complementarity is found in complex claim verification, with teams of experts and LLMs [126], and in hybrid NLP approaches where human intelligence and LLMs is combined to capture the diversity of perspectives in large-scale online discussions [117]. Finally, [158] compares default moral reasoning architectures in humans and LLMs and concludes that LLMs do not exhibit the threshold-deontological pattern characteristic of human moral reasoning, and people's prior beliefs about LLMs strongly shape moral agreement. This suggests a further precondition for synergy, namely accounting not only for behavioural complementarity, but also for asymmetries in the reasoning architectures of human and artificial partners.

3.1.2 *Design Principles for an Effective HI collaboration.*

Computational Theory of Mind. A common theme is the role of Theory of Mind (ToM) in collaboration, i.e. the ability of the agents in an HI system to model the beliefs, goals, and intentions of the others. [45] use higher-level concepts, like social roles, values and norms, as heuristics to make collaborative decisions, with a computational ToM framework based on epistemic abstractions, whose consistency is further evaluated in a medical domain scenario [43]. These ToM-equipped agents can improve decision quality, reduce negotiation rounds, and better preserve the interests of all parties especially when managing privacy in group conflicts [11, 42]. However, [170] shows how ToM can also cause deceptive cooperation, where humans – that rarely use more than second-order ToM– cannot distinguish artificial agents from other humans.

Team-aware Task Allocation. Multi-party interaction involves multiple humans and agents allocating tasks in a team. Effective delegation and task assignment requires human-agent teams to track their competencies and prior information, and adapt strategies in a trust-aware manner [1, 47, 73]. This can help an agent to efficiently approximate the full diversity of human judgements, valuing minority and majority perspectives equally [118]. [185] studies how trust is shaped by player behaviour in an online multiplayer games. If the composition of the team changes, graph-based policy learning support agents to adapt to the new team policies [140]. A deliberation process based on deliberative coalition formation is proposed by [40] to help teams finding a proposal with the most support.

Learning from Human Teachers. More intuitive interactions can be learnt from human teachers, e.g. [138, 154] draw on experience from dog training, highlighting shared understanding, feedback legibility, and the distribution of initiative between humans and agents. In human-robot teaching, early evaluative feedback can be used to allow non-experts to take over and increase the robots efficiency and safety [72]. Co-creation can also help, e.g. children shaping the interaction with a robot improves their sense of agency and co-regulation, producing higher enjoyment of the interaction and social acceptance of the robot [169]. When robots are employed in the wild, a novelty effect can often be observed [144]. Large Language Models offer then an opportunity to safely co-create interaction content in child-robot interaction [109, 183], which can also improve the child's humour [108]. Long-term sustained interest can be maintained with memory-based personalisation [95].

Human-AI Co-Creation for Problem Solving. An effective collaborative co-creation in an HI system requires a common language between humans and agents [112]. To avoid reluctance in adopting AI tools, the beliefs of the agent have to be aligned with the human's [209]. This lesson is applied to HI systems in healthcare [38, 130], where collaboration between patients, caretakers, and AI requires careful consideration of the roles each interaction partner takes, responsible actionable explanations, and engagement in reflective interactions. Subjective choices can be used as probability constraints and model a Bayesian network [134]. Effective co-creation also requires knowledge of how humans and machines differ, e.g. generalisation in [65]. Others study how to calibrate reliance in the team at runtime, using "hypervision" and bidirectional interaction to refocus operators on task completion in high-stakes operational settings [113].

3.1.3 *Enabling Multimodal Collaborative Interaction.*

Multimodal Perception of Social Signals. Understanding what humans are communicating during collaboration requires going beyond understanding speech. For example, facial cues identify spontaneous recollections [37], audiovisual cues recognise perceived interdependence in negotiations [35], and prosody as an implicit teaching signal [80] can be used. A partner perception study shows that perceived warmth can be predicted from facial and acoustic

cues during the interaction, whereas this is not the case for perceived competence [116]. The work of [145] identifies physical situatedness, multimodality, and variability as core challenges in human-robot dialogue. Robust perception for collaboration also requires reasoning over what is perceived: the ORBIT benchmark shows that current vision-language models still fall well short of human-like abstraction and common-sense when reasoning about object properties [82]. Episodic knowledge graphs can help encoding this multimodal aspects, as proposed by the modular architecture for multimodal agents of [14].

Dialogue Systems for Long-Term Grounded Interaction. Building dialogue systems capable of sustaining grounded, long-term interaction is a core challenge for HI. [78] shows that LLMs still struggle with classifying whether an agent's and human's memories of an event overlap, contradict or complement each other. Memory-based personalisation can be used to sustain long-term interest [95]. To maintain conversational continuity, conversational patterns from human-human interaction can be leveraged [89]. Structured knowledge in the form of knowledge graphs can also ground conversational agents in the domain long-term, e.g. in cultural heritage [100, 194]. Yet, extracting triples from conversations is still challenging when compared to text [192]. Knowledge graphs can also be used to model the different memory systems (short-term, episodic, semantic) of humans and robots [76]. The information in the knowledge graph can then be verbalised by leveraging pragmatic principles [86]. Other approaches for dialogue creation include hybrid frameworks combining rule-based scaffolding and selective LLM generation to produce pedagogically safe dialogues at scale [107, 109]. A different challenge is resolving user intent efficiently: this can be done by turning a model's uncertainty into a clarification question guaranteed to contain the true intent at a chosen confidence level [53], by designing evidence-based models that address barriers such as fatigue, time management, and habit formation [166] or by choosing the right conversational strategy, to avoid affecting users' self-conscious emotions [22]. To enable smooth conversational interactions, the human has to be aware of the agent's capabilities [145]. The human trust and their willingness to collaborate with an agent influences how they perceive their capabilities [143] and the frequency of interventions during a task [195]. Establishing a shared vocabulary between humans and LLMs can help creating more human-like interactions [85].

Common Ground and Shared Mental Models. Achieving genuine collaboration requires humans and agents to build and maintain a shared representation of their joint task and each other's perspectives. The Finding Common Ground framework [183] compares associative learning and ToM models in a collaborative game, finding that both can achieve effective alignment once partners are well-acquainted, suggesting that explicit perspective modelling may be most valuable in early or unfamiliar collaborations. Agents that establish common ground improve their joint performance, while reducing their reliance on active Theory-of-Mind modelling, showing that shared alignment can act as a mental shortcut [182]. The IGLU competition [77] offers a benchmark that also shows the challenges when building an agent that can learn to construct structures from natural language instructions. Personal values also play a role in collaborative dynamics, for example in online discussions [120], where the dissimilarities between values correlate with disagreement of users in online discussions.

Value alignment impacts the trust humans show towards the agent [123]. When humans and agents work together, they have to establish an appropriate level of trust for each other, without over- or undertrusting each other. Trust in the context of collaboration is formalised as mutual trust [70] or actual trust [6]. [23] shows how alignment dialogues help understand which dimensions of dialogue matter to model and update the partner's trust effectively. In interactions with uncertainty, communication becomes even more important to achieve cooperation [135, 136]. This benefit is conditional,

as e.g. in asymmetric uncertainty, agents that face no uncertainty can learn to exploit emergent communication to deceive their uncertain peers [137].

In Collaborative HI, several cross-cutting themes emerge. First, genuine HI collaboration requires **agents that model their partners**: their beliefs, goals, values, and communicative signals. Computational Theory of Mind can be here used both a theoretical framework and a practical design tool to enable agents and humans to make assumptions about the interaction partner(s), and enable successful collaboration. Second, **multimodal communication** – the integration of verbal, non-verbal, and gestural channels – is both a prerequisite for naturalistic collaboration and a major source of technical difficulty. Embodiment and situatedness add additional challenges to the interaction. Third, **temporal continuity matters**: systems that can sustain personalised, grounded relationships over time consistently outperform those optimised for single interactions. Structured knowledge in the form of Knowledge Graphs offer a technical solution to model and store both long-term and multimodal interactions.

3.2 Adaptive HI Systems

Adaptive HI systems are systems where the agents learn from and adapt to humans and their environments. Through communication, the artificial agents improve their learning, making dialogue processes computationally tractable. This can be achieved by interacting with humans, by dynamically adapting to on-task changes within the team, and by using symbolic constraints and representations that specify (human) knowledge. We distinguish here the following groups of research questions: (RQ1) How can interaction be used for learning (learning through interaction)?, (RQ2) How can learning algorithms be used for dialogue (learning to interact)?, (RQ3) How can learning methods interact with symbolic methods or constraints?, (RQ4) How can AI agents adapt on-task to changing circumstances?

3.2.1 Learning through Interaction.

Multi-agent Communication. Several studies on multi-agent reinforcement learning (MARL) investigate how artificial agents can learn to communicate effectively to improve collaboration. Work by [51] introduces a methodology that explicitly balances coordination performance with communication overhead, enabling multiple RL agents to achieve strong collaborative performance while minimizing unnecessary communication. Complementing this, [207] provided a comprehensive survey and a nine-dimensional taxonomy of communication strategies in MARL, offering a structured overview of existing approaches and identifying common design dimensions for communication-aware learning systems. Together, these works contribute to a broader understanding of how communication protocols can emerge and be optimised in collaborative learning environments.

Human-Agent Interaction. Another important research direction explored how human interaction can guide and shape agent learning. [189] demonstrates that human evaluative feedback and demonstrations can stimulate the emergence of collaborative behaviour between artificial agents. Expanding on how to process such input, [61] introduces an imitation learning framework that maps subjective human evaluations to objective algorithmic refinements, while [62] optimises this data collection via an active learning algorithm (EARLY) that queries the most effective expert demonstrations to accelerate convergence. Similarly, [111] investigates the use of human proxemics as corrective feedback during robot navigation, showing that physical interaction cues can serve as intuitive teaching signals. Extending these natural cues to non-expert human teachers, [154] adapts behavioural training principles such as luring and rewarding into an interactive programming method for robots.

Leveraging Affective Dialogue. Several works address the role of affective interaction in dialogue and learning systems. [81] explores whether machine learning techniques can predict how situations influence personal values based on psychological characteristics. The results show that personalised interactions substantially improve predictive accuracy and that active learning techniques can achieve strong performance using only a fraction of the training data. Meanwhile, [201] investigates whether RL-based appraisal models can predict and explain emotions during interactive tasks. The proposed model successfully captured temporal emotional dynamics, highlighting the importance of modelling interaction as an evolving process rather than a static state. Other progress also focuses on improving interaction strategies of dialogues through reinforcement learning. [203] addresses the problem of dead-end dialogue states in task-oriented dialogue systems, where certain conversational paths prevent users from successfully completing their tasks. Based on this, [204] proposes an RL-based dialogue strategy to support users in complex tasks by decomposing large goals into manageable subgoals. This hierarchical interaction strategy improves both the efficiency of dialogue and the completion of user tasks. [206] further addresses the issue of massive action spaces by leveraging the power of language models to assess action feasibility based on dialogue context, allowing an agent to perform fine-grained exploration over the pruned action subset.

3.2.2 Learning to Interact. We find here work focused on developing learning algorithms that can support conversational interaction in teams of humans and machines effectively. Across these studies, an important shared direction is the recognition that intelligent systems must not only optimise individual behaviour, but also coordinate communication, collaboration, and adaptation within multi-agent or human-agent environments.

Interaction among Multiple Agents. These works address how reinforcement learning can support collaboration among multiple agents. [51] proposed a methodology for balancing collaborative performance with communication overhead in multi-agent reinforcement learning (MARL). Complementing this, [207] provides a comprehensive taxonomy of communication in MARL systems, which categorised existing approaches and clarified the major dimensions involved in communication-aware multi-agent learning. This offers a conceptual framework for future development of collaborative learning algorithms. Another work investigates how conversational and deliberative interaction structures influence collaboration among agents [40]. The authors explored how the complexity of proposal spaces and coalition formation operators affect deliberative coalition formation processes. The findings suggest that group interaction in intelligent systems cannot rely solely on simple communication protocols, but instead requires adaptive reasoning and negotiation mechanisms that support collaborative decision-making in complex environments.

Interaction between Human and Agent(s). Human-centred interaction is also an important component. [20] investigates how reinforcement learning can support personalised interventions for reminding eating and drinking behaviour in people with early dementia. The study in simulations demonstrate that learning algorithms can adapt to the unique behavioural patterns and needs of individual users. Similarly, [209] explore a 'centaur' interaction and show that Bayesian reasoning over key variables in the model of the partner can lead to quick adaptation. In contrast, in ad-hoc teamwork, where there is no prior knowledge about the strategy of the partner(s) in a team, there are inherent limits to what we can expect to be able to adapt to [101], which motivates restricting to certain classes of socially intelligent agents [102]. Another study emphasises the importance of richer communication signals in collaborative learning systems. Researchers examine how prosody in human speech can function as a teaching signal for learning agents [79]. The study show that prosodic information carries task-relevant signals and that combining prosody with explicit feedback can improve learning performance compared to speech-only baselines. This work highlights the importance

of multimodal interaction in conversational learning environments. Beyond passive feedback signals, agents must also actively engage with users to resolve ambiguities. In this context, [53] addresses how dialogue agents can effectively ask clarification questions to quickly identify user intent, while [205] leverages a synergistic combination of an evolutionary algorithm for efficient dialogue interactions. Finally, [75] calibrates systems to model diversity of crowd annotations, so that systems request human input in ambiguous scenarios.

3.2.3 Integrating Symbolic Constraints. In the last decade, much progress in machine learning has focused on systems that process raw perception data, e.g. in the form of images (pixels) captured by sensors. Such representations contrast with the symbolic representations that humans typically use to reason and communicate in. As such, the question of how to unify subsymbolic learning with the symbolic reasoning in a HI system has become an urgent one.

Specifying and Incorporating Constraints. Constraints are often naturally specified by humans, for instance, using rules [184], geometric properties [48], guidelines [54], and causal models [152]. To respect such constraints, an agent needs to understand how 'percepts' (in terms of sensor readings) map to these symbols. Incorporating such mappings has received wide-spread attention in neuro-symbolic AI community [57]. The mappings themselves can be manually specified, learned, or one can try to get them from foundation models [24]. Given symbol-mappings together with rules or constraints an agent can filter out actions that would violate any hard constraints [54], also referred to as 'shielding' in the reinforcement learning literature.

Ill-defined Constraints and Preferences. It is not always the case that it is easy to specify constraints. E.g. how much distance should a robot keep from a human? In such cases, imitation learning could directly try to learn the behaviour and inverse reinforcement can be applied to try and infer a reward function that actually captures the preferences of an individual, potentially in a symbolic form [83]. Similarly, understanding how to balance preferences of different users (each of which might be poorly understood) is challenging. For instance, [11] proposes to tackle the issue of balancing privacy preferences, e.g. in social networks, using a combination of auction mechanisms and multiagent reinforcement learning. To this end, the authors identify symbolic knowledge in terms of context dependent factors, time dependent factors, social dependent factors.

Verifying Constraints after Learning. An alternative is off-line reinforcement learning [93], which trains on a previously collected data set. Since the resulting policy might take very different actions than the one used to collect the data, determining if it adheres to particular quantitative constraints may be very difficult. [60] proposes to use data augmentation to solve this issue. Instead, [54] uses off-policy evaluation (OPE) to evaluate the performance of an agent in a medical domain. OPE typically only assumes statistical guarantees, i.e. the probability that a constraint is satisfied. As such, safety is increasingly approached using a combination of statistical and symbolic methods. In other words, constraints are one form of trying to guarantee properties such as safety, but more generally learning systems need (statistical) verification pipelines since deductive reasoning on induced models might fail.

Other Neuro-symbolic Learning Combinations. Constraints are an important form of symbolic knowledge, and critical for safe deployment. However, the integration of symbolic and learning systems might present many more benefits: symbolic representations may make learned solutions interpretable [57], and symbolic knowledge can accelerate learning as well as lead to better generalisation [48, 209]. As an example, [52] show that by merely labelling states with a symbol that indicates which (if any) subtask was just completed, the reward structure in complex tasks, can be exploited in hierarchical learning for significantly more efficient learning in tasks like "construct a bed" in Minecraft.

With differentiable fuzzy logics [184], the agent uses symbolic background knowledge (e.g. a part of a chair can be an armrest, but not the other way around) to reason about unlabelled, or poorly labelled data. This can be translated into a differentiable loss functions thus informing the overall optimisation. Notably, this work analyses different fuzzy operators and shows that the best-performing operator combinations for learning violate classical logical laws, thus exposing a fundamental tension between logical correctness and optimisation behaviour.

3.2.4 Adapting On-task to Changing Circumstances. In HI systems, the goals, constraints, and environments of artificial agents can constantly shift, e.g. due to changing priorities or team changes within the team. Artificial agents should be able to react to such changes rapidly—for example, by avoiding learning from scratch or by using rapid learning methods rather than gradient-based learning. We distinguish several approaches:

- *Accommodating Changes by Conditioning.* To accommodate changes without gradient-based learning, one can include a description of the relevant context in the agent's input. In [59], this context is the set of capabilities of the agent. These helped to adjust trajectory prediction and decision making in cross-embodiment scenarios. In [68], a virtual museum guide adapts its behaviour based on gaze information included in the agent prompt. Such context can consist of prior interactions such as in [209], where the agent acts optimally with respect to a posterior belief about the interaction partner.
- *Accommodating changes by rapid learning.* If re-learning is unavoidable, this need to be done efficiently. [20] uses contextual bandits to select actions that smartly balance exploration (information gathering actions) and exploitation (of the knowledge about the user). In [81], a baseline model is personalised into individual models, using active learning strategies to select the unlabelled data points with the largest uncertainty to query.
- *Accommodating changes using reasoning and logic.* Systems that can perform logical reasoning are not dependent on gradient-based learning for adaptation. For example, in [3], an agent reasons about the capabilities of sensors and processing algorithms to select the most suitable perception components. [197] describes a framework where the agent's knowledge about the world and the user can be updated using a proposed logic model.
- *Accommodating changes with game-theoretic strategies.* Modelling HI interaction through a game-theoretic lens gives rise to unique strategies to propose changes. [103]'s strategy allows a 'leader' to learn without access to 'follower' gradients by using a commitment strategy. Instead, [15] starts with a 'safe' imitation strategy to learn the partner's behaviour, then switches to a 'coercive' strategy that leads to superior collaboration.
- *Adaptive Learning Mechanisms.* Lastly, some approaches learn the adaptation mechanism itself. In [149], models a (instance-independent) relationship between instance features and parameters to configure negotiation algorithms for different instances and negotiation partners — allowing more quick adaptation in future negotiations. [152] targets discovering causal mechanisms by learning a (instance-independent) relation of the state of the problem to the choice of interventions and construction of causal diagrams.

The Adaptive HI literature reflects a gradual shift away from viewing learning as a purely data-driven optimisation problem toward understanding adaptation as a continuous process shaped by **interaction, context, and structured knowledge**. **Communication** is no longer treated merely as an output channel, but as an essential mechanism for coordination, feedback exchange, personalisation, and efficient learning in both mixed human–machine or machine-machine teams. Placing the interaction at the centre of learning, **soft-constraints** are preferred to hard constraints to favour optimisation. Additionally, to enable agents to remain responsive under

dynamic circumstances while maintaining computational efficiency, **symbolic representations** in the form of rules, constraints, preferences, and logical abstractions complement statistical learning and allow to incorporate human knowledge, guide behaviour, and verify system properties that are difficult to guarantee through data alone. The resulting **neuro-symbolic perspective** shifts from "learning from data" towards "learning under structural constraints", where the structure corresponds to safety critical symbols or codifies domain knowledge. These developments position adaptation as the product of an ongoing interplay between learning, reasoning, and interaction within hybrid human–AI systems. Overall this sketches a route forward where "data + rules + evaluation = safe/interpretable AI system".

3.3 Responsible HI Systems

Responsible HI systems combine socio-technical governance, explicit representations of norms and values, monitoring mechanisms, and empirical evaluation strategies to assess ethical, legal, and societal (ELS) considerations. They are organised according to two research questions: (RQ1) How can ELS considerations be effectively integrated into the design of HI systems? and (RQ2) How to ensure ELS alignment throughout the development and lifecycle of HI systems?

3.3.1 In-design Integration of ELS considerations. A recurring challenge in designing HI systems that take ELS considerations into account is the issue of clearly identifying what a safe system entails within this context. Research in HI has aimed to address this issue by designing general frameworks, as well as approaches and frameworks relating to specific ELS considerations. The Hard Choices in AI (HCAI) framework [33] aims to empower AI developers to address these issues and inform design decisions, by highlighting the attention needed on societal impacts, whilst framing HI systems as socio-technical systems [34]. Notably, HI work also recognises how users may have differing preferences across values, which may be modelled as part of an agent's interactions [157]. We also observe critique of broadly used mechanisms such as Reinforcement Learning from Human Feedback (RLHF) [97], for its lack of nuance and reflexivity. In addition to general frameworks, we find that across various ethical, legal an societal considerations, we recognise work that has focused on the following questions:

- *Privacy.* An individual user needs to frequently decide whether they would like to share particular pieces of personal information or whether to keep it to themselves. When interacting with many different agents on a daily basis, it becomes impractical to do this alone. As such, [90] proposes a privacy assistant to automate part of this process whilst deferring to the user for ambiguous cases. However, in cases where personal information refers to multiple individuals (e.g. a group picture) this becomes challenging, due to potentially conflicting preferences [178]. The proposed PANOLA privacy assistant [179] addresses this complexity through an auction-based privacy mechanism, thereby learning user preferences over time.
- *Ethics.* Work surrounding ethics in HI primarily focuses on either moral values [98, 127] or notions of fairness [56, 151], where the former is situated in specifics contents, and the latter reflects top-down values instilled into a system. In [127], the view of ethics in HI as a socio-technical problem is detailed, identifying the challenge of clearly identifying the stakeholders' values. The Axies system [99] is specifically proposed to address this challenge, combining human annotation and active learning with NLP techniques. However, it remains a challenge to balance between context-specified and general values.
- *Legal.* Unique characteristics of the law make it challenging to apply existing systems to take the considerations from this domain into account [164]. As of yet, HI systems have therefore largely focused on a case-based reasoning approach to better align with legal practice [186, 196].

- *Environmental Justice.* The increasing societal concerns surrounding the environmental impact of AI and data centres have so far received limited attention the HI context. A notable exception here is the Social-Ecological-Technological Systems (SETS) approach [141], which defines HI systems as SETS, which allows for connecting research on the environmental impacts of HI to existing work on environmental justice.

3.3.2 Alignment throughout the System Lifecycle. Whether a HI system is alignment with ELS values and considered trustworthy is not a static notion, and whilst methodologies for gathering and understanding design considerations provide a foundation, it is crucial to ensure continued alignment during both the development and the lifecycle of the system. Responsible HI systems attempt to achieve this through a combination of algorithmic interventions and methodological frameworks that enable new ways of approaching the challenges of ensuring continued alignment.

During Development. A challenge for developing well-aligned HI systems is operationalising design priorities. For AI agents, it is increasingly common to pursue alignment through methods that use reinforcement learning from (human) feedback, although [97] critically evaluates this paradigm's theoretical underpinnings and practical implementations and finds significant limitations in their ability to capture the complexities of human ethics. Beyond methods which study system alignment, a cluster of work also relates to methodological frameworks to adopt during development. For instance, [105] argues that the neo-republican conception of freedom as non-domination provides a suitable framework for informing how to develop an AI agent. Translating this framework to development informs how and which accountability mechanisms should be developed. Similarly, the Hard Choices in Artificial Intelligence framework [33] seeks to empower developers how to navigate decisions surrounding design goals and major socio-technical challenges, thereby steering away from vagueness through deliberation.

During Deployment. A number of papers in this line investigate how to ensure continued alignment of deployed HI systems. To this end, one approach for this is proposed by [133], through normative monitoring to determine whether input-output combinations of HI systems are acceptable in specific contexts. Another study introduces a strategy that uses challenge sets to carefully test whether a model produces the output that is intended by its design [74]. These methods are specifically designed for black-box systems, making it broadly applicable for understanding the behaviour of a system once deployed. Such an algorithmic approach to monitoring can subsequently be combined with an auditing approach as proposed in [141], which proposes a structured qualitative audit guide for stakeholders around ontological reframing, qualitative question framework and policy recommendations. These works highlight how both actively monitoring and auditing are necessary to ensure continued alignment. Other works relate to trust, e.g. [17] argues that trust in the HI system should be calibrated throughout the entire process of system deployment, offering multiple trust calibration points depending on the lifecycle step (pre-, during-, post-deployment and use). Acknowledging honesty in decisions also allows to gain more subjective trust from users[124]. A complementary approach to monitoring for alignment during deployment is found in systems which actively optimise their behaviour to align with values that the system is designed around. For example, [56] proposes a method for recommendating policies optimised for both fairness and user utility. More closely addressing the issue of personalising recommendations, [10] proposes a personal assistant which learns over time which privacy labels to recommended to users for shared content, which is tailored to the user's own understanding of privacy. An important point highlighted by these works is that considerations around ethical, legal and societal norms are highly situated, and can depend on both the user and the content.

Post-Deployment Considerations. Due to the application of HI systems across domains, including life-critical domains such as law or medicine, it is also crucial to understand their potentials and limitations. One such task is the

identification of moral values in textual discourse, and whether a state-of-the-art system for this task can generalise across domains [98]. The findings in [98] shows that knowledge can be transferred to new domains, but that there is a risk of catastrophic forgetting. In the legal domain, an interesting approach presented in [186] is to use explainable AI methods to evaluate whether a given dataset indeed measures what it is intended to measure. To confirm the potential of such an approach, their analysis finds such strong inconsistencies in the COMPAS dataset that this even brings into doubt prior work using this dataset. In [160], a method is proposed to evaluate and improve the decision-making of black-box machine learning systems using manually crafted rationale evaluation test sets based on domain knowledge. The approach demonstrates that machine learning systems can arrive at correct decisions for unsound reasons. Moreover, [159] showed that contemporary explainable AI methods are often unable to reliably uncover such flawed reasoning.

The Responsible HI literature provides frameworks and methods for all phases for creating HI systems. Designing for ethical, legal and societal (ELS) considerations is mainly approached through **socio-technical design frameworks** that identify stakeholders, values, and institutional contexts within specific application domains or for specific ELS considerations. There remains a lack of broadly applicable methodologies across multiple ELS considerations. For assessing the alignment of an HI system to ELS considerations, a number of methodologies have been proposed, that focus on both **auditing and monitoring**, in addition to frameworks to guide the development process itself. Crucially, ELS considerations are highly situated, requiring **adaptation to user and context**, which has been clearly argued theoretically, but practical implementations remain lacking.

3.4 Explainable HI Systems

Explainable HI systems extend beyond a one-way account of decisions or predictions made by black-box AI systems for human interpretation. In these systems, agents and humans leverage the environment to construct a mutual understanding of the decisions to be made, maintaining transparency in the process. The key is to foster understanding and (co-)learning in a mutually beneficial manner. The research in this area is organised according to the following key questions: (RQ1) How to build shared representations to be used as the basis for explanations, covering both the external world and the internal problem-solving process?, (RQ2) What are the different types of explanations that make the decision-making process more transparent and understandable?, and (RQ3) How can explanations be communicated to users such that the explanations improve the user's trust and leads to a successful agent-user collaboration?

3.4.1 Shared Representations for Explanations. Several HI methods reveal how decisions are made – the so-called *white-box* models. We find a core focus on argumentation [200], knowledge graphs [163, 192], and Bayesian Networks [147]. These techniques prevail because they yield a clear shared representation that can be leveraged by humans and artificial agents. Further, a key capability of successful representations seems to be able to support diverse, potentially disagreeing, viewpoints [18, 75, 117]. Lastly, connecting outputs back to input cases (i.e. training data) by means of case-based reasoning [139] and connecting definitions used to create datasets to carefully designed behavioural testing [75] enhances explainability of how data impacts decision-making ability. Techniques are often applied to scenarios that are common in their respective fields (e.g. legal setting in the case of argumentation, or hate speech classification for pluralistic NLP). Besides knowing what to put in shared representations, the prevailing questions include when to explain [87] and how to construct accurate representations [2].

Several works focus on explaining neural *black box* systems [46, 84, 88, 129], varying in the extent to which they are able to provide faithful analyses and their accessibility to non-experts. Feature attribution methods, highlighting input

contribution to a decision, are intuitive to humans [26], but different methods may provide conflicting outputs [129]. More fine-grained mechanistic interpretability methods can be tested for faithfulness at a more detailed level [46], but they may not be readily accessible to non-expert users.

3.4.2 Types of Explanations. Making decision-making processes transparent means choosing *when*, *what*, and *how to explain*, under the constraint that explanations be tailored to the user and faithful to the model. We observe that faithfulness, more than personalisation, is what forces the move from approximate attribution to traceable symbolic reasoning. This is evidenced by formal proofs that interpretable models are sound explainers, for Bayesian Networks [147] and argumentation [50, 58]. Knowing when to explain, or what type of features to reveal can establish transparency, understandability and trust [67, 74, 87, 106]. Given that symbolic methods are promising, existing research focuses on novel methods to inspect black boxes, either from a human [26] or machine [55] perspective. Examples of use cases include finding user preferences in museum exhibitions based on Virtual Reality [67], legal reasoning [161] or decision-making in the medical domain [171].

3.4.3 Communicating Explanations. Several studies investigate how to construct systems that aim to align users' trust to the agent's capability, because successful agent–user collaborations do not come from maximising trust, but from calibrating *appropriate trust* [121]. We find that the works operate on two different levels: (i) direct communication, and (ii) establishing context. In the first case, the agent actively manages the expectation from the user, driving it towards the (known) capability of the system through interaction [177]. In the second, agents are built to adapt to the situation, as an explanation may arrive differently depending on personal factors [69] and cultural factors [156]. Hence, situating the explanation and understanding its relevant context is required to make systems practical and usable [122].

Explainable HI systems strongly focus on **faithful white-box models** that are verifiable. Argumentation and case-based reasoning allow to go beyond simple verbalisation of rationales, and reflect accurately the decision-making process by handling inconsistent data and provide explanations. Contrastive explanations seem to be relatively understudied, as feature-/importance-based techniques (black box techniques) are prevalent. Additionally, knowing when, what, and how to explain is crucial to create **faithful models beyond personalisation**. This helps establish transparency, understandability and trust. Finally, there is a strong emphasis on **not over-aligning with the agents' capabilities**. Either through explicit communication or contextual cues, agents are expected to reveal their shortcomings, leading to a more trustworthy and adequate use of these tools.

4 Anatomy of Hybrid Intelligence Systems

Once we know how HI has evolved (RQ1/*Evolution*), we can extract the building blocks that can be used to create HI systems (RQ2, *Anatomy*). Following the common system development life cycle methodologies [64], this section presents how an HI system need to be *designed* using the right building blocks (Section 4.1), tested and *evaluated* with the right evaluation methods (Section 4.2, and *deployed* and maintained in applications in the real-world (Section 4.3).

4.1 Design of HI Systems

Here we focus on the structural foundations of Hybrid Intelligence, identifying the core methods characteristics to model an effective synergistic human-AI collaboration. Inspired by Software and Knowledge Engineering methodologies, we identified and combined methods into a unified language, in the form of shared vocabularies, taxonomies, ontologies and pattern-based design methodologies. These can be then used to identify recurrent patterns and characteristics of HI

systems. Ultimately, these recurrent design solutions can help reduce development time, avoid repeated work, support dissemination, and allow HI systems to be compared and evaluated.

4.1.1 *Unified Languages.*

Shared Taxonomies and Vocabularies. Taxonomies, vocabularies and ontologies allow to characterise and formalise concepts of a domain. These frameworks guide system designers in conceptualising and creating of HI systems, particularly in terms of commonalities, differences, and recurrent reusable components that can used to compare across different applications. Some papers analyse actors and processes in HI systems, e.g. the taxonomy of HI actors, HI processes and HI and input/output data types in [16], and the HI Knowledge Model [168] that provides a high-level ontology of HI common components across HI applications, describing HI actors, their capabilities, tasks, goals, and context. Other works focus on standardising the design of hybrid groups and teams, e.g. the Quality Model of HI teams including properties, attributes and measures to design cooperating human-AI teams [29], or models and methods to design trust calibration and appropriate trust in human-AI teams [71, 121].

Boxologies and Compositional Design Patterns. Various methodologies to design HI systems are inspired by Software and Knowledge Engineering. Team Design patterns have been used to describe recurring elements and guide the design of HI systems, e.g. AI Advisor and Human Performer, AI Performer and Human Assistant, AI Performer and Human Validator [208]. Boxologies have also been used, particularly to represent and standardise process workflows in the form of labelled nodes and paths between them. For instance [16] presents a boxology to categorise neuro-symbolic architectures used in HI applications, while a more general boxology based on Knowledge Engineering is used help engineers in designing high-level workflows in HI applications [66].

Specific Design Frameworks. Other design frameworks are proposed with more specific angles. SHARPIE [12] allows engineers to design and experiment with HI systems involving humans and Reinforcement Learning-based agents. Some studies focus on how to foster appropriate mutual trust when designing HI systems, proposing factors such as e.g. development ethicality, communication of trust and interactivity [28], personal values [123], point interventions and goal alignment [195]. Other studies explore the integration of system design and stakeholder requirements. For instance, "Hard Choices in AI" framework [33] helps designers in finding trade-offs between the design of their system and the socio-technical values and norms at stake, [25] illustrates how Digital Humanities applications should be designed according to the HI CARE principles since they require a combination of human expertise, critical judgment, and the computational power of AI agents, while [96, 138] present how to integrate HI principles at design-time in human(or child)-robot interaction settings.

4.1.2 *Recurring Patterns in HI systems.*

Capabilities of an HI system. We identify six recurring capabilities in HI Systems:

(1) *Partner Modelling* refers to the ability to represent beliefs, goals, intentions, values, and competences of human collaborators in order to support coordination and anticipation [45, 183, 197].
(2) *Context Adaptivity* refers the ability to adjust actions, outputs, or policies based on context, user feedback, or physical interaction, including learning from explicit or implicit signals [20, 68, 109, 138, 197].
(3) *Grounded Communication* concerns the interpretation and generation of language and multimodal signals in context, including common ground and conversational continuity [77, 78, 100, 145, 183].

(4) *Normative Alignment* refers to reasoning about and acting in accordance with societal values, norms, legal constraints, and fairness criteria relevant to the human–AI context [8, 42, 120, 133].
(5) *Transparent Reasoning* concerns explanations, uncertainty, provenance, and interpretable representations that support oversight and appropriate trust [28, 139, 162, 188, 197].
(6) *Collaborative Action* captures the coordination of initiative, task allocation, joint problem-solving, structured deliberation, and co-creation [1, 40, 138, 183, 185].

Roles in an HI team. We distinguish a set of recurring human and artificial agent roles in an HI system. A single person or agent may hold multiple roles simultaneously (e.g. a clinician who is both domain expert and decision-maker), and the analytically interesting unit is often the *dyad*: which human and agent roles are actively paired in a given system. Human roles can be distinguished in the following categories, with a growing degree of agency and expertise:

H1 primary recipient or *beneficiary*, receiving support from the system or participating in the interaction without necessarily providing domain expertise, making consequential decisions;
H2 domain expert or *trainer*, validating outputs, curating knowledge, teaching agents through feedback/demonstration;
H3 decision-maker, controller or *operator*, who uses agent decisions to decide in a consequential context;
H4 auditor, or *governor*, who monitors system behaviour and compliance, bearing institutional accountability.

Artificial agents take the following roles, with a growing degree of support and interaction abilities (dyads to groups):

A1 information or reasoning *assistant*, supporting retrieval, synthesis, argumentation, or decision support;
A2 adaptive or *personalising agent*, adjusting behaviour based on individuals, context, feedback, or physical actions;
A3 interaction *facilitator*, coordinating dialogue, collaboration, structured deliberation among human participants;
A4 monitor or *explainer*, exposing reasoning, provenance, values to support oversight and appropriate trust.

4.2 Evaluation of HI systems

Here we provide an overview of the categories of evaluation methods (Section 4.2.1) and metrics (Section 4.2.2) that allow to undertake and measure HI research, ranging from formal methods and metrics, to qualitative socio-technical evaluation. These methods and metrics are meant as a non-exhaustive overview, but rather serving as a base toward building methodologies to assess Hybrid Intelligence systems as a whole – i.e. an analytical lens through which the quality of "HI-ness" [66] can be made visible– that we discuss in Section 4.2.2.

4.2.1 Types of Evaluation Methods. Evaluation methods in HI research reflect its inherently interdisciplinary nature, spanning a spectrum from highly formalised and abstract approaches to increasingly qualitative and situated methodologies. This diversity also highlights how HI is still maturing methodologically.

(1) **Formal and theoretical evaluations** can involve logical proofs of the different semantics that agents can use in their interactions [39], checking of symbolic models representing the team's beliefs and behaviours [198] and mathematical analysis of e.g. social networks dynamics [187] or fairness [56].
(2) **Experimental system validation** involves empirical testing through experiments and simulations to assess HI systems' performance, reliability, and safety. This involves the use of expert-defined benchmarks [27], controlled experiments [63], simulations [20] and games [42], among others.
(3) **Empirical studies with users** involve direct engagement with human participants to gather rich data on HI system effectiveness, user satisfaction, and interaction quality. Qualitatively, interviews and focus groups [180, 193] capture self-reported, subjective experiences. Quantitative studies involve behavioural observations of

user performance, e.g. frequency of human intervention or reliance on AI [91, 153, 195], and physiological observations, eg. biosignals [177], eye-tracking [151], or indicators of collaborative dynamics such as gaze patterns, body movement, and speaker activity [142, 176].

(4) Addressing the challenges of socio-technical systems evaluation calls for holistic **socio-technical methods**. These involve real-world organisational studies [180, 193], but also specific analytical frameworks, dedicated for example to the detection of normative issues in the lifecycle of HI systems [17, 34, 141], including system safety frameworks [32] and power analysis [141].

It is worth noticing that these methods are often explicitly combined, e.g. benchmarks and user-studies [109, 115, 203, 204], or re-contextualised, e.g. the agent's accuracy in aligning with others in the team [42] to evaluate how well the agent aligns with a human partner's decision logic. This indicates a trend toward creating novel, "multidimensional evaluations", where the agent performance is validated alongside the outcomes of its collaboration with the human(s) within the HI team. In other words, the HI literature shifts away from pure technical, objective performance metrics, toward a multi-stakeholder, socio-technical evaluation approach where the "HI-ness" and success of the HI system is defined by the quality of the human-agent interaction, rather than just the efficiency of the single AI model.

4.2.2 Evaluation Metrics. Recurring evaluation metrics in the corpus of HI research can be grouped into two families: more classical, *technology-centric* metrics that evaluate various aspects of AI performance, e.g. standard ML/statistical benchmarks, and HI-specific, *interaction-centric* metrics, that are more human-centred, qualitative, or collaboration-focused and are aimed at assessing the quality of collaboration emerging in the HI team.

Evaluation of the AI agent. Understanding how suited certain AI components that are used in HI systems are for particular tasks requires careful evaluation.

Performance metrics in the HI literature primarily assess technical performance of the AI agent, including typical task accuracy metrics [173], precision and recall [192], calibration [190], and computational efficiency indicators including throughput, latency, and retrieval speed [131].

Fairness, alignment, and normative considerations are also evaluated, e.g. measuring the systems' performance in terms of normative compliance [132, 133], value alignment and consistency [92, 97], and fairness metrics such as exposure fairness [151] and ranking fairness [56]. Some work examines bias detection and mitigation through measures of lexical diversity, computational flow diversity, and representation equity [174].

Other HI work range from providing clear definitions and criteria for evaluation [133] to empirically evaluating how well systems perform on ethical, legal and societal dimensions [27]. [132] proposes a framework for normative monitoring of HI systems, which is extended with a Bayesian approach for setting a dynamic threshold to judge whether a system is behaving as intended [133]. This dynamic approach is motivated by the recognition that a given environment may have a specific context that determines how appropriate behaviour is defined.

Evaluation of the Team Interaction. HI-specific metrics to measure the team interaction go beyond the traditional AI performance metrics and treat the human as an active, subjective, and indispensable part of the HI system. Their novelty lies in the ability to operationalise and quantify typical psychology or sociology constructs. Example of such metrics include human perception, e.g. usability, satisfaction, engagement, perceived usefulness [99] confidence in AI-assisted decisions [91] and cognitive workload of users [167], trust in automation and technology acceptance [193], subjective

trust [122] and trust calibration, which is tied to the concept of appropriate reliance [122, 153]. Appropriate reliance is measured on the basis of users' adherence to correct AI advice and override of incorrect recommendations [153].

Some metrics focus on evaluating the conversation as part of an interaction, since this presents a distinct measurement challenge. The level-based scoring system for conversational agents [146] decomposes interaction quality into utterance, dialogue-move, and genre levels. At team level, the CHATboard framework [1] – designed for task allocation in ad-hoc human-agent teams – introduces metrics for team performance, miscalibration, and improvement over time. The Conversation metrics in [13] also evaluates enjoyability (Interesting, Engaging, Specific, Relevant), semantic correctness (Correct, Semantically Appropriate, Understandable), and fluency, while [146] focuses on perceived conversation quality and likeability questionnaires. In social robotics, the robots' social presence and friendship can be assessed [96].

Evaluating HI Systems as a Whole. Our analysis underscores the need to establish a canonical set of evaluation methodologies to provide policymakers, researchers, and developers with concrete insights into the assessment of HI solutions. Meta-reflection in the field has already led to the first instances of HI-specific frameworks [29, 66, 112, 168]. These frameworks focus on *meta-level constructs*: generalisable concepts not tied to a specific HI system or project, but applicable across a variety of contexts. For example, the HIKE toolkit [66] uses Knowledge Engineering to evaluate the level of *HI-ness* of an HI system when performing high-level tasks specific of HI, i.e. involving creativity, explainability, team awareness and task negotiation, multimodal communication. This toolkit is then applied to virtual heritage applications, showing both its diagnostic value and its modelling limitations. As as team-oriented approach to evaluate HI-ness, the Quality Model [29] frames HI systems as HI teams and proposes related quality attributes, representing properties relevant for HI teams, and quality measures, quantifying quality attributes. The quality attributes include Boundedness, Interdependence, Competency, Purposefulness, Initiative, Normativity, and Effectiveness. These attributes are evaluated in terms of the HI quality measures of Users satisfaction, Strategy appropriateness, Knowledge and skills, Quality of group processes and Quality of interactions, Satisfaction with relations, Growth opportunities, General satisfaction. Recognising the limitations of pre-existing models of technology acceptance in capturing the "dynamic mutual learning process of HI systems", [112] proposes to extend the Technology Acceptance Model into a novel Hybrid Intelligence Technology Acceptance Model (HI-TAM) that includes HI aspects of user control, AI output transparency, perceived partnership, willingness to train, willingness to co-develop and willingness to adopt a proposed solution.

4.3 Applications

HI systems have been deployed in a number of different settings. We list them below, also emphasising the core aspects of the HI systems should be further investigated especially when working in these application contexts.

(1) *Law and legal reasoning* make explanation, argumentation, justification, and accountability central to the design of hybrid systems [139, 162, 188].
(2) *Healthcare* work emphasises personalised behaviour support, lifestyle change, understandable user modelling, trust, and responsible support systems [21, 38, 130, 197].
(3) *Governance and deliberation* foreground values, disagreement, pluralism, and collective decision-making [40, 99, 120].
(4) *Social Robotics* expose embodiment, timing, safety, shared control, and situated dialogue as central design issues [2, 111, 138, 145].
(5) *Education* highlights learning, co-regulation, personalisation, and the distribution of initiative between learners, teachers, and agents [96, 109].

(6) *Simulation and game-based testbeds* provide controlled settings for studying cooperation, coordination, trust, common ground, and adaptation [140, 183, 185].
(7) *Assistance to Scientific Discovery* foregrounds knowledge access, dialogue, retrieval, question answering, and epistemic support as central problems for HI systems, often enabled by NLP, knowledge graphs, and dialogue technologies [77, 86].
(8) *Cultural heritage* settings foreground multimodal engagement, visitor modelling, polyvocal interpretation, and adaptive interaction [66–69].
(9) *Privacy and online platforms* work examines how agents can support users in preference-sensitive and multi-party decision contexts [179].

Taken together, these domains show that HI in practice is not limited to case studies or small deployed demonstrators. Instead, HI is operationalised across a spectrum of settings: from formal and simulation-based environments, to broader application domains, to case-study settings with concrete stakeholders and societal constraints.

Roles in HI Applications. Table 1 categorises the above HI application domains according to the recurring HI capabilities and HI roles identified in Section 4.1.2. Several structural patterns appear: H1–A2 (beneficiary paired with an adaptive agent) is the most common configuration, appearing across health [20, 130], education [109], cultural heritage [68], and robotics settings [138, 145]. This suggests that adaptation or personalisation is one of the more developed forms of human–AI teaming in the current HI literature. By contrast, H4 (auditor or governor) appears most clearly when the systems shows normative alignment and transparent reasoning. In the first case, the evaluator (H4) is associated with adaptive agents, monitors, and explainers (A2-A4 roles) that support oversight of values, norms, consent, fairness, and socio-technical risk. Examples include socio-technical vulnerability mapping [34], consent mechanisms for responsible autonomy [8], value-conflict mitigation through computational Theory of Mind [42], contextual critique of affective-data collection [36], and appropriate trust in human-centred AI design [28]. In the case of transparent reasoning, H4 is associated with A1/A4 roles, where argumentation, case-based reasoning, provenance, and explanation mechanisms support inspection, justification, and accountability [124, 139, 162, 188].

Finally, Table 1 also suggests that human roles oriented to oversight are still less systematically integrated into partner modelling, adaptive behaviour, grounded communication, and collaborative action. This points to a concrete gap in the current HI literature: evaluation and governance are most visible where responsibility and explanation are already foregrounded, but are not yet routinely embedded across the full range of HI capabilities.

Table 1. Recurring HI capabilities, role dyads, and representative application domains.

Capability	**Role dyads**	**Representative Domain Applicaitons**
Partner modelling	H1-H3–A1/A2	Privacy and Online Platforms; Governance and Deliberation; Social Robotics; Education ; Healthcare
Context Adaptivity	H1/H2–A2	Healthcare; Education; Cultural Heritage; Social Robotics; Assistance to Scientific Discovery
Grounded Communication	H1-H3–A2/A3	Assistance to Scientific Discovery; Cultural Heritage; Social Robotics ; Education
Normative Alignment	H2-H4–A2/A4	Law and legal reasoning; Governance and Deliberation; Privacy and Online Platforms; Healthcare
Transparent Reasoning	H1-H4–A1/A4	Law and legal reasoning; Assistance to Scientific Discovery ; Healthcare ; Cultural Heritage
Collaborative Action	H1-H3–A1/A3	Governance and Deliberation; Education; Social Robotics ; Simulation / Game-based Testbeds; Cultural Heritage

HI Maturity. Ultimately, we characterise the maturity of the HI applications identified by distinguishing three levels of HI behaviour. At the *reactive* level (level L1), systems respond to explicit user input, predefined rules, or immediate task demands. At the *proactive* level (level L2), systems initiate actions, adapt to implicit feedback, or use contextual cues to support the human partner. At the *social* level (level L3), systems sustain, repair, or improve longer-term collaborative relations, shared understanding, norms, or role distributions over time.

Most progress on HI capabilities is concentrated at L1-L2: systems increasingly move beyond reactive behaviour toward proactive partner modelling, grounding, explanation, and coordination [43, 78, 183]. Mature L3 social behaviour remains rare: while some work addresses shared episodic memory, common ground, capability communication, social personalisation, and interaction in the wild, these remain partial mechanisms rather than sustained systems that maintain, repair, or transform relationships, and shared understanding over time [78, 108, 144, 183].

Cross-application findings. First, we remark that what counts as success differs across HI applications. For example, legal reasoning foregrounds justification and argumentative explanations [139, 163]; the education domain foregrounds learning, co-regulation, and personalisation [109]; health foregrounds safe, trustworthy, and personalised support [38, 130]; cultural heritage foregrounds engagement, interpretation, and adaptive visitor interaction [66, 68]; social robotics foregrounds situated interaction, timing, feedback legibility, and shared control [138, 145]; and deliberation foregrounds disagreement, values, and collective decision-making [40, 120].

Second, CARE principles rarely appear in isolation. In health settings, personalisation has to be balanced with safety, trust, and responsible support [38]. In cultural heritage settings, adaptive interaction is tied to multimodal grounding, explanation, and human engagement [68]. In education and scientific discovery, knowledge access and computational support have to remain accountable to human expertise, interpretation, and critical judgement [25].

Third, application domains act as stress tests for HI assumptions by exposing practical tensions. Trust and appropriate reliance become visible in human-centred design practice [28]; long-term use of social robots exposes novelty effects and sustainability challenges [144]; value-sensitive applications foreground tensions around values and disagreement [120]; and human–robot teaching highlights timing, feedback legibility, and shared control [138].

Fourth, methods travel across domains. Dialogue methods recur in human–robot interaction and conversational agents [145]; knowledge graphs support grounded dialogue and cultural heritage interaction [66]; reinforcement learning supports adaptive and personalised interventions [20]; argumentation provides reusable structures for explanation and justification [139]; human modelling supports personalisation and understandable behaviour change support [197]; and gaze or multimodal sensing supports adaptive interaction in VR settings [68]. These recurring methods suggest that HI practice is beginning to generate reusable design and evaluation knowledge rather than isolated prototypes.

Finally, we suggest that HI applications should be understood not merely as deployed systems in the work, but as socio-technical configurations in which human and artificial agents take complementary roles, coordinate around shared goals, adapt over time, and remain accountable and understandable to stakeholders.

5 Challenges and Open Questions

To open up new pathways for future HI research generations, we finally discuss problems that have not been solved.

Disentagling Explainability, Responsibility and Trustworthiness. The research HI agenda remains deeply ingrained in the explainable AI (XAI) literature. For instance, the focus on the construction of sound explainers that move beyond correlation toward *interpretability* through validation with (causal) methods for identifying the internal mechanisms that play a role in a decision. Crucially, we wish to be certain that a model genuinely behaves in a faithful way, which is

essential from a compliance standpoint, e.g. by frameworks such as the AI Act. Transparency and right-to-explanation obligations for high-risk systems demand verifiable accounts of system behaviour. These are increasingly treated as technical requirements, and are often decoupled from studying effective methods for explaining agent decisions to users directly. We also observe the development of HI-specific challenges. For instance, calibrating for mutual trust as agents become more autonomous, is a central motivator at the moment. However, the downstream effects of explanations on human critical thinking and long-term learning remain understudied. Further, XAI literature remains largely one-directional (e.g. explaining machine decisions to humans) and devoid of specific context. However, the broader goal of creating shared understanding in a HI system requires us to calibrate bidirectionally, and adjust communication to specific contexts. How to practically do this remains an open question: XAI has from the outset been interested in revealing reasoning in agents to humans, rather than the reverse. In this process, the human is given little control over teaching the model what it should care about, or which explanations are effective, since the shared representations are typically set beforehand. Intriguingly, the fact that LLMs are themselves trained on human-provided reasoning traces shows that this channel, too, can run both ways, a mutually beneficial, co-adaptive opportunity.

Post-GenAI Considerations. While the HI agenda envisioned collaborative agents capable of reasoning about human goals, adapting to users, and participating in mixed human–machine teams, the emergence of GenerativeAI (GenAI) has introduced challenges that were largely absent from the original formulation. In particular, the widespread deployment of Large Language Models (LLMs) in activities traditionally associated with human expertise, including writing, tutoring, and creative work has intensified concerns regarding AI anthropomorphisation and over-reliance. This blurs the boundaries between assistance and delegation, raising important questions about accountability, responsibility, and human autonomy. Hybrid Intelligence research must therefore increasingly concern itself not only with how AI systems support human performance, but also with how they shape human cognition, judgement, learning, and decision-making over time. This concern has been echoed by the HI community, with emerging strategies such as Frictional AI [128] setting a research agenda for HI emphasising cognitive engagement and meaningful human control. With its foundational commitment to augmentation rather than substitution, HI is uniquely positioned to preserve human agency, and foster appropriate reliance, ensuring that AI systems support human understanding and participation rather than encouraging passive acceptance of machine-generated outputs.

From HI systems to HI socio-technical systems. The HI literature remains focused on dyadic human–AI interaction. The challenge is to move toward a "holistic" approach [208], i.e. recognising that intelligence often emerges from the structured interaction of multiple human and artificial actors acting in a socio-technical system, rather than focusing on technical challenges alone. Future research should extend beyond individual users and isolated human–AI teams, to examine how hybrid systems operate within larger socio-technical arrangements, including organisations, institutions, and society. An important challenge is therefore to develop theories, methods, and evaluation frameworks capable of analysing HI across multiple levels of an organisation. This includes understanding how interaction protocols scale from individuals to teams, how organisational structures shape hybrid performance, and how emergent effects propagate across larger socio-technical systems. The HI Manifesto[9] follows this idea, arguing that HI requires moving beyond the design of individual interactions toward the engineering and governance of hybrid socio-technical systems in which humans, AI systems, institutional rules, and governance structures become jointly responsible for collective outcomes.

[9] https://hi-infinity.ai/

From Multimodality to Embodiment. While embodied agents are recurrent in HI systems, the limitation currently lies in their use mostly as social communicative companions. A main challenge remains in enabling embodied agents to collaborate in situations where activities may change on the fly (i.e. they are *open-ended*), and where achieving the task requires dealing with multiple levels of knowledge (i.e. they are *knowledge-intensive*). For this, embodied agents need to manage, represent and reason with large sources of multimodal perceived knowledge about themselves, the others, and the environment around them. Such knowledge cannot be dealt with only by existing sub-symbolic learning methods to deal with multi-modal data, but needs to be symbolically embedded to enable a transparent and accountable collaboration. Modern symbolic AI techniques to handle large-scale heterogeneous structured knowledge (from neuro-symbolic methods to semantic technologies, and Knowledge Engineering at scale) offer the unique opportunity to enhance HI methods to efficiently handle multi-modal data. This idea of an "embodied HI", integrating multimodal perception and high-level concept reasoning, is seen in recent activities [2, 114][10], and would ultimately allow to tackle long-standing robotics challenges – i.e. acquiring/identifying relevant knowledge for a task (including long-term ones), aligning tasks and goal representations, and collaborating with humans in a trustworthy way.

Toward a Science of HI. As Hybrid Intelligence remains a relatively young research field, only a limited number of contributions explicitly address the epistemological foundations of a science of HI. Existing efforts only focus on specific aspects, e.g. the HI Quality Model for assessing hybrid human–AI teams [29], ontological representations of HI tasks and roles [168], and reusable interaction design patterns [16]. While these contributions provide valuable building blocks, they stop short of establishing a unified scientific foundation for the field. A recurring challenge is the absence of the meta-level infrastructure required for cumulative knowledge building. It remains difficult to determine what constitutes Hybrid Intelligence, how different forms of human–AI collaboration relate to one another, and which design choices consistently lead to desirable outcomes. Establishing a science of HI requires the development of shared frameworks for describing, comparing, assessing, and ultimately theorising human–AI collaboration. This paper is a first attempt to provide the conceptual and methodological foundations needed to transform Hybrid Intelligence from a collection of promising approaches into a cumulative and generalisable body of knowledge.

6 Conclusions

We presented a reflection and overview of the evolution of almost a decade of Hybrid Intelligence (HI) research. Starting from the original Hybrid Intelligence research agenda and building on the analysis of approx. 350 papers originating from it, we investigated how the original HI principles have been turned into practice, and which building blocks are needed when creating HI systems. The result is a conceptual and practical view of how the HI field evolved for new research generations. Conceptually, we discussed and provided input to the question of "what are the characteristics of HI systems w.r.t. other types of (AI) systems?". In practice, we provided (i) the building blocks to create HI systems, ranging from shared vocabularies, ontologies and design patterns to recurrent human and AI roles, HI system capabilities and methods to evaluate them; (ii) pointers to approaches and technologies that might be used to implement them, including a.o. adaptive dialogue systems, knowledge graphs, socio-technical design, informed reinforcement learning, partner modelling strategies, value-aware NLP techniques, and multimodal knowledge acquisition methods; and (iii) examples of HI applications. Finally, we have discussed a few fundamental challenges still open for future research to tackle.

[10] Including e.g. https://www.dagstuhl.de/en/seminars/seminar-calendar/seminar-details/26153

References

[1] Sami Abuhaimed and Sandip Sen. 2022. Effective Task Allocation in Ad Hoc Human-Agent Teams. In *HHAI2022: Augmenting Human Intellect*. IOS Press, 171–183.

[2] Mark Adamik, Romana Pernisch, Ilaria Tiddi, and Stefan Schlobach. 2024. Advancing Robotic Perception with Perceived-Entity Linking. In *The Semantic Web - ISWC 2024 - 23rd International Semantic Web Conference, Baltimore, MD, USA, November 11-15, 2024, Proceedings, Part II* (2024-01-01) *(Lecture Notes in Computer Science, Vol. 15232)*, Gianluca Demartini, Katja Hose, Maribel Acosta, Matteo Palmonari, Gong Cheng, Hala Skaf-Molli, Nicolas Ferranti, Daniel Hernández, and Aidan Hogan (Eds.). Springer, 192–209. doi:10.1007/978-3-031-77850-6_11

[3] Mark Adamik, Romana Pernisch, Ilaria Tiddi, and Stefan Schlobach. 2024. ORKA: An Ontology for Robotic Knowledge Acquisition. In *Knowledge Engineering and Knowledge Management - 24th International Conference, EKAW 2024, Amsterdam, The Netherlands, November 26-28, 2024, Proceedings* (2024-01-01) *(Lecture Notes in Computer Science, Vol. 15370)*, Mehwish Alam, Marco Rospocher, Marieke van Erp, Laura Hollink, and Genet Asefa Gesese (Eds.). Springer, 309–327. doi:10.1007/978-3-031-77792-9_19

[4] Stéphane Airiau, Nicholas Kees Dupuis, and Davide Grossi. 2024. Condorcet Markets. In *Algorithmic Game Theory - 17th International Symposium, SAGT 2024, Amsterdam, The Netherlands, September 3-6, 2024, Proceedings* (2024-01-01) *(Lecture Notes in Computer Science, Vol. 15156)*, Guido Schäfer and Carmine Ventre (Eds.). Springer, 501–519. doi:10.1007/978-3-031-71033-9_28

[5] Zeynep Akata, Dan Balliet, Maarten de Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Anstke Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, Hayley Hung, Catholijn Jonker, Christof Monz, Mark Neerincx, Frans Oliehoek, Henri 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, and Max Welling. 2020. A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *IEEE Computer* 53, 08 (2020), 18–28. Place: Los Alamitos, CA, USA Publisher: IEEE Computer Society. doi:10.1109/MC.2020.2996587

[6] Michael Akintunde, Vahid Yazdanpanah, Asieh Salehi Fathabadi, Corina Cirstea, Mehdi Dastani, and Luc Moreau. 2024. Formal specification of actual trust in multiagent systems. In *HHAI 2024: Hybrid Human AI Systems for the Social Good: Proceedings of the Third International Conference on Hybrid Human-Artificial Intelligence*. SAGE Publications 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP, 22–35.

[7] Costanza Alfieri, Martina De Sanctis, Donatella Donati, and Paola Inverardi. 2024. In search of clarity: discerning between human replacement and augmentation. In *HHAI 2024: Hybrid Human AI Systems for the Social Good*. IOS Press, 36–45.

[8] Anastasia S. Apeiron, Davide Dell'Anna, Pradeep K. Murukannaiah, and Pınar Yolum. 2025. Model and Mechanisms of Consent for Responsible Autonomy. In *24th International Conference on Autonomous Agents and Multiagent Systems* (2025-01-01). https://ifaamas.csc.liv.ac.uk/Proceedings/aamas2025/pdfs/p133.pdf

[9] Angela Aristidou, Rajesh Jena, and Eric J Topol. 2022. Bridging the chasm between AI and clinical implementation. *The Lancet* 399, 10325 (2022), 620.

[10] Gönül Ayci, Murat Şensoy, Arzucan Özgür, and Pinar Yolum. 2022. Uncertainty-Aware Personal Assistant for Making Personalized Privacy Decisions. *ACM Transactions on Internet Technology* (2022). Publisher: Association for Computing Machinery. doi:10.1145/3561820

[11] Hüseyin Aydın. 2024. Assisting Users in Privacy Conflicts with Partially Observable Multi-Agent Reinforcement Learning. In *HHAI 2024: Hybrid Human AI Systems for the Social Good*. IOS Press, 55–63. https://ebooks.iospress.nl/volumearticle/67999

[12] Hüseyin Aydın, Kevin Godin-Dubois, Libio Goncalvez Braz, Floris den Hengst, Kim Baraka, Mustafa Mert Çelikok, Andreas Sauter, Shihan Wang, and Frans A. Oliehoek. 2025. SHARPIE: A Modular Framework for Reinforcement Learning and Human-AI Interaction Experiments. In *AAAI Bridge Program Workshop on Collaborative AI and Modeling of Humans (CAIHU)* (2025-02-01). Philadelphia, Pennsylvania, USA. doi:10.48550/arXiv.2501.19245

[13] Selene Baez Santamaria, Piek Vossen, and Thomas Baier. 2022. Evaluating Agent Interactions Through Episodic Knowledge Graphs. In *Proceedings of the 1st Workshop on Customized Chat Grounding Persona and Knowledge* (2022-10-01). Association for Computational Linguistics, Gyeongju, Republic of Korea, 15–28. https://aclanthology.org/2022.ccgpk-1.3

[14] Thomas Baier, Selene Baez Santamaria, and Piek Vossen. 2022. A modular architecture for creating multimodal agents. In *MMAI2022: Workshop on the representation, sharing and evaluation of multimodal agent interaction, Workshop at HHAI 2022* (2022-01-01). doi:10.48550/arXiv.2206.00636

[15] Saptarashmi Bandyopadhyay, Mustafa Mert Çelikok, and Robert Loftin. 2025. On the Complexity of Learning to Cooperate in Populations of Socially Rational Agents. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*. 233–241.

[16] Michael van Bekkum, Maaike de Boer, Frank van Harmelen, André Meyer-Vitali, and Annette ten Teije. 2021. Modular design patterns for hybrid learning and reasoning systems. *Appl. Intell.* 51, 9 (2021), 6528–6546. doi:10.1007/s10489-021-02394-3

[17] Jacob T Browne, Saskia Bakker, Bin Yu, Peter A Lloyd, and Somaya Ben Allouch. 2022. Trust in Clinical AI: Expanding the Unit of Analysis.. In *HHAI*. 96–113.

[18] Luana Bulla, Misael Mongiovì, and Aldo Gangemi. 2025. Underperformance or pluralism: A machine learning perspective on Inter-Annotator agreement. In *HHAI 2025: Proceedings of the 4th International Conference on Hybrid Human-Artificial Intelligence*. SAGE Publications 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP, 303–315.

[19] Federico Cabitza, Chiara Natali, Francesco Varanini, and David Gunkel. 2025. Beyond cyborgs: the cybork idea for the de-individuation of (artificial) intelligence and an emergence-oriented design. *AI & SOCIETY* 40, 5 (2025), 3333–3348.

[20] Wen-Tseng Chang, Shihan Wang, Stephanie Kramer, Michel Oey, and Somaya Ben Allouch. 2024. Human-Centered AI for Dementia Care: Using Reinforcement Learning for Personalized Interventions Support in Eating and Drinking Scenarios. In *HHAI 2024: Hybrid Human AI Systems for the Social Good - Proceedings of the Third International Conference on Hybrid Human-Artificial Intelligence, Malmö, Sweden, 10-14 June 2024* (2024-01-01) *(Frontiers in Artificial Intelligence and Applications, Vol. 386)*. IOS Press, 84–93. doi:10.3233/FAIA240185

[21] Pei-Yu Chen, Selene Baez Santamaria, Maaike HT de Boer, Floris den Hengst, Bart A Kamphorst, Quirine Smit, Shihan Wang, and Johanna Wolff. 2024. Intelligent Support Systems for Lifestyle Change: Integrating Dialogue, Information Extraction, and Reasoning. In *HHAI 2024: Hybrid Human AI Systems for the Social Good*. IOS Press, 457–459.

[22] Pei-Yu Chen, M. Birna van Riemsdijk, Dirk K. J. Heylen, Catholijn M. Jonker, and Myrthe L. Tielman. 2025. Why don't you do what you said you would? Conversational strategies for agents to understand users' reasons in supporting behavior. *Behaviour & Information Technology* 0, 0 (2025), 1–20. Publisher: Taylor & Francis. doi:10.1080/0144929X.2025.2483793

[23] Pei-Yu Chen, Myrthe L Tielman, Dirk KJ Heylen, Catholijn M Jonker, and M Birna Van Riemsdijk. 2023. Acquiring semantic knowledge for user model updates via human-agent alignment dialogues. In *HHAI 2023: Augmenting Human Intellect: Proceedings of the Second International Conference on Hybrid Human-Artificial Intelligence*. 93–107.

[24] Daniel Cunnington, Mark Law, Jorge Lobo, and Alessandra Russo. 2024. The role of foundation models in neuro-symbolic learning and reasoning. In *International Conference on Neural-Symbolic Learning and Reasoning*. Springer, 84–100.

[25] Victor de Boer and Lise Stork. 2024. Hybrid Intelligence for Digital Humanities. In *HHAI 2024: Hybrid Human AI Systems for the Social Good*. IOS Press, 94–104. doi:10.3233/FAIA240186

[26] Maria Luigia Natalia de Bonis, Giuseppe Fasano, Angela Lombardi, Aldo Testino, Eugenio di Sciascio, Tommaso di Noia, and Carmelo Ardito. 2025. Explainable AI for Brain Age Prediction: Design, Implementation, and Formative Evaluation of an Interactive Tool. In *HHAI 2025*. IOS Press, 248–261.

[27] Milena de Swart, Floris Den Hengst, and Jieying Chen. 2025. Detecting Linguistic Bias in Government Documents Using Large language Models. In *Proceedings of the ACM on Web Conference 2025* (2025-01-01) *(WWW '25)*. Association for Computing Machinery, Sydney NSW, Australia, 5034–5044. doi:10.1145/3696410.3714526

[28] Chadha Degachi, Siddharth Mehrotra, Mireia Yurrita, Evangelos Niforatos, and Myrthe Lotte Tielman. 2024. Practising Appropriate Trust in Human-Centred AI Design. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems* (2024-01-01) *(CHI EA '24)*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3613905.3650825

[29] Davide Dell'Anna, Pradeep K. Murukannaiah, Bernd Dudzik, Davide Grossi, Catholijn M. Jonker, Catharine Oertel, and Pinar Yolum. 2024. Toward a Quality Model for Hybrid Intelligence Teams. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2024, Auckland, New Zealand, May 6-10, 2024* (2024-01-01), Mehdi Dastani, Jaime Simão Sichman, Natasha Alechina, and Virginia Dignum (Eds.). International Foundation for Autonomous Agents and Multiagent Systems / ACM, 434–443. doi:10.5555/3635637.3662893

[30] Dominik Dellermann, Philipp Ebel, Matthias Söllner, and Jan Marco Leimeister. 2019. Hybrid intelligence. *Business & information systems engineering* 61, 5 (2019), 637–643.

[31] Davide Dell'Anna, Pradeep K Murukannaiah, Mireia Yurrita, Bernd Dudzik, Davide Grossi, Catholijn M Jonker, Catharine Oertel, and Pınar Yolum. 2026. From human teams to hybrid intelligence teams: identifying, characterizing, and evaluating foundational quality attributes. *Autonomous Agents and Multi-Agent Systems* 40, 1 (2026), 10.

[32] Roel Dobbe. 2022. System Safety and Artificial Intelligence. In *Oxford Handbook of AI Governance*. Vol. To Appear. Oxford University Press, Oxford. https://arxiv.org/abs/2202.09292

[33] Roel Dobbe, Thomas Krendl Gilbert, and Yonatan Mintz. 2021. Hard choices in artificial intelligence. *Artificial Intelligence* 300 (2021). doi:10.1016/j.artint.2021.103555

[34] Roel Dobbe and Anouk Wolters. 2024. Toward Sociotechnical AI: Mapping Vulnerabilities for Machine Learning in Context. *Minds and Machines* 34, 2 (May 2024), 12. doi:10.1007/s11023-024-09668-y

[35] Bernd Dudzik, Simon Columbus, Tiffany Matej Hrkalovic, Daniel Balliet, and Hayley Hung. 2021. Recognizing Perceived Interdependence in Face-to-Face Negotiations through Multimodal Analysis of Nonverbal Behavior. In *Proceedings of the 2021 International Conference on Multimodal Interaction*. Association for Computing Machinery, New York, NY, USA, 121–130. doi:10.1145/3462244.3479935

[36] Bernd Dudzik, Tiffany Matej Hrkalovic, Chenxu Hao, Chirag Raman, and Masha Tsfasman. 2025. Indeterminacy in Affective Computing: Considering Meaning and Context in Data Collection Practices. https://arxiv.org/abs/2502.09294

[37] Bernd Dudzik and Hayley Hung. 2022. Exploring the Detection of Spontaneous Recollections during Video-Viewing In-the-Wild Using Facial Behavior Analysis. In *Proceedings of the 2022 International Conference on Multimodal Interaction* (2022-01-01) *(ICMI '22)*. Association for Computing Machinery, New York, NY, USA, 236–246. event-place: Bengaluru, India. doi:10.1145/3536221.3556609

[38] Bernd Dudzik, Jasper van der Waa, Pei-Yu Chen, Roel Dobbe, Inago M. D. R. de Troya, Roos M. Bakker, Maaike H. T. de Boer, Quirine T. S. Smit, Davide Dell'Anna, Emre Erdogan, Pinar Yolum, Shihan Wang, Selene Baez Santamaría, Lea Krause, and Bart A. Kamphorst. 2024. Viewpoint: Hybrid Intelligence Supports Application Development for Diabetes Lifestyle Management. *J. Artif. Intell. Res.* 80 (2024), 919–929. doi:10.1613/JAIR.1.15916

[39] Wolfgang Dvorák, Atefeh Keshavarzi Zafarghandi, and Stefan Woltran. 2020. Expressiveness of SETAFs and Support-Free ADFs under 3-valued Semantics. *CoRR* abs/2007.03581 (2020). arXiv: 2007.03581. https://arxiv.org/abs/2007.03581

[40] Edith Elkind, Davide Grossi, Ehud Shapiro, and Nimrod Talmon. 2024. United for change: deliberative coalition formation to change the status quo. *Soc. Choice Welf.* 63, 3 (2024), 717–746. doi:10.1007/S00355-024-01561-Y

[41] Douglas C Engelbart. 2023. Augmenting human intellect: A conceptual framework. In *Augmented education in the global age*. Routledge, 13–29.

[42] Emre Erdogan, Hüseyin Aydin, Frank Dignum, Rineke Verbrugge, and Pinar Yolum. 2025. Mitigating Value Conflicts with Computational Theory of Mind. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2025, Detroit, MI, USA, May 19-23, 2025* (2025-01-01), Sanmay Das, Ann Nowé, and Yevgeniy Vorobeychik (Eds.). International Foundation for Autonomous Agents and Multiagent Systems / ACM, 695–703. doi:10.5555/3709347.3743586

[43] Emre Erdogan, Frank Dignum, and Rineke Verbrugge. 2024. Effective Maintenance of Computational Theory of Mind for Human-AI Collaboration. In *HHAI 2024: Hybrid Human AI Systems for the Social Good - Proceedings of the Third International Conference on Hybrid Human-Artificial Intelligence, Malmö, Sweden, 10-14 June 2024* (2024-01-01) *(Frontiers in Artificial Intelligence and Applications, Vol. 386)*, Fabian Lorig, Jason Tucker, Adam Dahlgren Lindström, Frank Dignum, Pradeep K. Murukannaiah, Andreas Theodorou, and Pinar Yolum (Eds.). IOS Press, 114–123. doi:10.3233/FAIA240188

[44] Emre Erdogan, Frank Dignum, Rineke Verbrugge, and Pinar Yolum. 2022. Computational Theory of Mind for Human-Agent Coordination. In *Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XV* (2022-01-01), Nirav Ajmeri, Andreasa Morris Martin, and Bastin Tony Roy Savarimuthu (Eds.). Springer International Publishing, 92–108. http://dx.doi.org/10.1007/978-3-031-20845-4_6

[45] Emre Erdogan, Rineke Verbrugge, and Pinar Yolum. 2024. Computational Theory of Mind with Abstractions for Effective Human-Agent Collaboration. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2024, Auckland, New Zealand, May 6-10, 2024* (2024-01-01), Mehdi Dastani, Jaime Simão Sichman, Natasha Alechina, and Virginia Dignum (Eds.). International Foundation for Autonomous Agents and Multiagent Systems / ACM, 2249–2251. doi:10.5555/3635637.3663123

[46] Leon Eshuijs, Shihan Wang, and Antske Fokkens. 2025. Short-circuiting Shortcuts: Mechanistic Investigation of Shortcuts in Text Classification. In *Proceedings of the 29th Conference on Computational Natural Language Learning* (2025-07-01), Gemma Boleda and Michael Roth (Eds.). Association for Computational Linguistics, Vienna, Austria, 105–125. doi:10.18653/v1/2025.conll-1.8

[47] Andrea Fanti, Francesco Frattolillo, Rosapia Laudati, Fabio Patrizi, and Luca Iocchi. 2025. Human-AI Collaboration Via Trust Factors: A Collaborative Game Use Case. In *HHAI 2025: Proceedings of the 4th International Conference on Hybrid Human-Artificial Intelligence*. SAGE Publications 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP, 60–73.

[48] Alejandro García-Castellanos, David Wessels, Nicky J van den Berg, Remco Duits, Daniël M Pelt, and Erik Bekkers. 2026. Equivariant Eikonal Neural Networks: Grid-Free, Scalable Travel-Time Prediction on Homogeneous Spaces. *Advances in Neural Information Processing Systems* 38 (2026), 151777–151817.

[49] Kevin Godin-Dubois, Olivier Weissl, Karine Miras, and Anna V. Kononova. 2024. Interactive Embodied Evolution for Socially Adept Artificial General Creatures. In *Evolution of Things Workshop at the ALife 2024 Conference* (2024-07-01). arXiv. Type: Workshop eprint: 2407.21357. doi:10.48550/arXiv.2407.21357

[50] Davide Grossi, Wiebe van der Hoek, and Louwe B. Kuijer. 2022. Reasoning about General Preference Relations. *Artif. Intell.* 313, C (2022). Place: GBR Publisher: Elsevier Science Publishers Ltd.. doi:10.1016/j.artint.2022.103793

[51] Shuai Han, Mehdi Dastani, and Shihan Wang. 2023. Model-based Sparse Communication in Multi-agent Reinforcement Learning. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems* (2023-01-01). 439–447. https://www.southampton.ac.uk/~eg/AAMAS2023/pdfs/p439.pdf

[52] Shuai Han, Mehdi Dastani, and Shihan Wang. 2024. Learning Reward Structure with Subtasks in Reinforcement Learning. In *ECAI 2024 - 27th European Conference on Artificial Intelligence, 19-24 October 2024, Santiago de Compostela, Spain - Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024)* (2024-01-01) *(Frontiers in Artificial Intelligence and Applications, Vol. 392)*. IOS Press, 2282–2289. doi:10.3233/FAIA240751

[53] Floris Hengst, Ralf Wolter, Patrick Altmeyer, and Arda Kaygan. 2024. Conformal Intent Classification and Clarification for Fast and Accurate Intent Recognition. In *Findings of the Association for Computational Linguistics: NAACL 2024* (2024-01-01). 2412–2432. doi:10.18653/v1/2024.findings-naacl.156

[54] Floris den Hengst, Martijn Otten, Paul Elbers, Frank van Harmelen, Vincent François-Lavet, and Mark Hoogendoorn. 2024. Guideline-informed reinforcement learning for mechanical ventilation in critical care. *Artificial Intelligence in Medicine* 147 (2024), 102742. doi:10.1016/j.artmed.2023.102742

[55] Maria Heuss, Maarten de Rijke, and Avishek Anand. 2025. RankingSHAP - Faithful Listwise Feature Attribution Explanations for Ranking Models. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025* (2025-01-01), Nicola Ferro, Maria Maistro, Gabriella Pasi, Omar Alonso, Andrew Trotman, and Suzan Verberne (Eds.). ACM, 381–391. doi:10.1145/3726302.3729971

[56] Maria Heuss, Fatemeh Sarvi, and Maarten de Rijke. 2022. Fairness of Exposure in Light of Incomplete Exposure Estimation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2022-01-01). Association for Computing Machinery, New York, NY, USA, 759–769. event-place: Madrid, Spain. https://irlab.science.uva.nl/wp-content/papercite-data/pdf/heuss-2022-fairness.pdf

[57] Pascal Hitzler and Md Kamruzzaman Sarker. 2022. Neuro-symbolic artificial intelligence: The state of the art. (2022).

[58] Loan Ho, Somjit Arch-int, Erman Acar, Stefan Schlobach, and Ngamnij Arch-int. 2022. An argumentative approach for handling inconsistency in prioritized Datalog± ontologies. *AI Commun.* 35, 3 (2022), 243–267. Place: NLD Publisher: IOS Press. doi:10.3233/AIC-220087

[59] Niklas Hoepner, David Kuric, and Herke van Hoof. 2025. Making Universal Policies Universal. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems* (2025-01-01). International Foundation for Autonomous Agents and Multiagent Systems, 2553–2555.

[60] Niklas Hoepner, Ilaria Tiddi, and Herke van Hoof. 2025. Data Augmentation for Instruction Following Policies via Trajectory Segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2025-01-01), Vol. 39. 17214–17222. Issue: 16. https://ojs.aaai.org/index.php/AAAI/article/view/33892

[61] Muhan Hou, Koen Hindriks, AE Eiben, and Kim Baraka. 2023. A Process-Oriented Framework for Robot Imitation Learning in Human-Centered Interactive Tasks. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (2023-01-01). IEEE, 1745–1752. https://ieeexplore.ieee.org/document/10309326/

[62] Muhan Hou, Koen V. Hindriks, Guszti Eiben, and Kim Baraka. 2024. "Give Me an Example Like This": Episodic Active Reinforcement Learning from Demonstrations. In *Proceedings of the 12th International Conference on Human-Agent Interaction, HAI 2024, Swansea, United Kingdom, November 24-27, 2024* (2024-01-01), Muneeb Imtiaz Ahmad, Katrin S. Lohan, Mary Ellen Foster, Patrick Holthaus, and Yukie Nagai (Eds.). ACM, 287–295. doi:10.1145/3687272.3688298

[63] Tiffany Matej Hrkalovic, Aria Li, Magnus Bopp, Yingling Li, and Daniel Balliet. 2025. Task affordances affect partner preferences. *Journal of Experimental Social Psychology* 119 (2025), 104751. https://www.sciencedirect.com/science/article/pii/S0022103125000320

[64] Ing Axel Hunger. 2009. Fundamentals of Software Engineering I. (2009).

[65] Filip Ilievski, Barbara Hammer, Frank van Harmelen, Benjamin Paassen, Sascha Saralajew, Ute Schmid, Michael Biehl, Marianna Bolognesi, Xin Luna Dong, Kiril Gashteovski, Pascal Hitzler, Giuseppe Marra, Pasquale Minervini, Martin Mundt, Axel-Cyrille Ngonga Ngomo, Alessandro Oltramari, Gabriella Pasi, Zeynep Gozen Saribatur, Luciano Serafini, John Shawe-Taylor, Vered Shwartz, Gabriella Skitalinskaya, Clemens Stachl, Gido M. van de Ven, and Thomas Villmann. 2025. Aligning generalization between humans and machines. *Nature Machine Intelligence* 7 (2025), 1378 – 1389. https://arxiv.org/abs/2411.15626

[66] Delaram Javdani Rikhtehgar, Ilaria Tiddi, Shenghui Wang, Stefan Schlobach, and Dirk Heylen. 2024. Assessing the HI-ness of Virtual Heritage Applications with Knowledge Engineering. In *HHAI 2024: Hybrid Human AI Systems for the Social Good*. IOS Press, 173–187. https://ebooks.iospress.nl/volumearticle/68010

[67] Delaram Javdani Rikhtehgar, Shenghui Wang, Hester Huitema, Julia Alvares, Stefan Schlobach, Carolien Rieffe, and Dirk Heylen. 2023. Personalizing Cultural Heritage Access in a Virtual Reality Exhibition: A User Study on Viewing Behavior and Content Preferences. In *Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization* (2023-01-01). 379–387. https://dl.acm.org/doi/pdf/10.1145/3563359.3596666

[68] Delaram Javdani Rikhtehgar, Shenghui Wang, Stefan Schlobach, and Dirk Heylen. 2025. Real-time gaze awareness in conversational agents: Enhancing collaboration and personalization in vr art experiences. In *HHAI 2025*. IOS Press, 153–168.

[69] Delaram Javdani Rikhtehgar, Shenghui Wang, Stefan Schlobach, and Dirk Heylen. 2025. Seeing and Speaking with Culture: How Visitor Profiles Shape Multimodal Interaction in a VR Art Exhibition. In *Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents (IVA '25)*. Association for Computing Machinery, New York, NY, USA, Article 26, 9 pages. doi:10.1145/3717511.3747084

[70] C Centeio Jorge, Siddharth Mehrotra, ML Tielman, and CM Jonker. 2021. Trust should correspond to trustworthiness: a formalization of appropriate, mutual trust in human-agent teams. In *Proceedings of the 22nd International Workshop on Trust in Agent Societies* (2021-01-01). London, UK. https://ceur-ws.org/Vol-3022/paper4.pdf

[71] Carolina Centeio Jorge, Emma M van Zoelen, Ruben Verhagen, Siddharth Mehrotra, Catholijn M Jonker, and Myrthe L Tielman. 2024. Appropriate context-dependent artificial trust in human-machine teamwork. In *Putting AI in the Critical Loop*. Elsevier, 41–60. https://doi.org/10.1016/B978-0-443-15988-6.00007-8

[72] A. E. Eiben Karine Miras, Decebal Mocanu. 2023. Hu-bot: promoting the cooperation between humans and mobile robots. *Neural Computing and Applications* (2023). https://link.springer.com/article/10.1007/s00521-022-08061-z

[73] Vijay Keswani, Elisa Celis, Krishnaram Kenthapadi, and Matthew Lease. 2023. Designing closed-loop models for task allocation. In *HHAI 2023: Augmenting Human Intellect: Proceedings of the Second International Conference on Hybrid Human-Artificial Intelligence*. SAGE Publications, 17–32.

[74] Urja Khurana, Eric Nalisnick, and Antske Fokkens. 2025. DefVerify: Do Hate Speech Models Reflect Their Dataset's Definition?. In *Proceedings of the 31st International Conference on Computational Linguistics* (2025-01-01), Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE, 4341–4358. https://aclanthology.org/2025.coling-main.293/

[75] Urja Khurana, Eric Nalisnick, Antske Fokkens, and Swabha Swayamdipta. 2024. Crowd-Calibrator: Can Annotator Disagreement Inform Calibration in Subjective Tasks?. In *First Conference on Language Modeling (COLM)*. Philadelphia, USA.

[76] Taewoon Kim, Michael Cochez, Vincent François-Lavet, Mark Neerincx, and Piek Vossen. 2023. A Machine with Short-Term, Episodic, and Semantic Memory Systems. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2023-02-01). doi:10.48550/ARXIV.2212.02098

[77] Julia Kiseleva, Ziming Li, Mohammad Aliannejadi, Shrestha Mohanty, Maartje ter Hoeve, Mikhail Burtsev, Alexey Skrynnik, Artem Zholus, Aleksandr Panov, Kavya Srinet, et al. 2022. Interactive grounded language understanding in a collaborative environment: Iglu 2021. In *NeurIPS 2021 Competitions and Demonstrations Track*. PMLR, 146–161.

[78] Annika Kniele, Lucia Donatelli, Catharine Oertel, and Piek Vossen. 2025. Human-Agent Co-Construction of Episodic Memories. In *4th International Conference on Hybrid Human-Artificial Intelligence, HHAI 2025* (2025-01-01) *(Frontiers in Artificial Intelligence and Applications, Vol. 408)*. 228–237. https://ebooks.iospress.nl/doi/10.3233/FAIA250640

[79] Matilda Knierim, Sahil Jain, Murat Han Aydogan, Kenneth Mitra, Kush Desai, Akanksha Saran, and Kim Baraka. 2024. Leveraging Prosody as an Informative Teaching Signal for Agent Learning: Exploratory Studies and Algorithmic Implications. In *Proceedings of the 26th International Conference on Multimodal Interaction, ICMI 2024, San Jose, Costa Rica, November 4-8, 2024* (2024-01-01), Hayley Hung, Catharine Oertel, Mohammad

Soleymani, Theodora Chaspari, Hamdi Dibeklioglu, Jainendra Shukla, and Khiet P. Truong (Eds.). ACM, 95–123. doi:10.1145/3678957.3685735

[80] Matilda Knierim, Sahil Jain, Murat Han Aydoğan, Kenneth D Mitra, Kush Desai, Akanksha Saran, and Kim Baraka. 2024. Leveraging prosody as an informative teaching signal for agent learning: Exploratory studies and algorithmic implications. In *Proceedings of the 26th International Conference on Multimodal Interaction*. 95–123.

[81] Ilir Kola, Ralvi Isufaj, and Catholijn M Jonker. 2022. Does personalization help? Predicting how social situations affect personal values. In *HHAI2022: Augmenting Human Intellect: Proceedings of the First International Conference on Hybrid Human-Artificial Intelligence*. 157–170.

[82] Abhishek Kolari, Mohammadhossein Khojasteh, Yifan Jiang, Floris den Hengst, and Filip Ilievski. 2025. ORBIT: An Object Property Reasoning Benchmark for Visual Inference Tasks. *arXiv preprint arXiv:2508.10956* (2025).

[83] Nikolaos Kondylidis, Andrea Rafanelli, Ilaria Tiddi, Annette ten Teije, and Frank van Harmelen. 2025. "Stop replacing salt with sugar!": Towards Intuitive Human-Agent Teaching. *arXiv preprint arXiv:2509.24651* (2025).

[84] Nikolaos Kondylidis, Anil Yaman, Frank van Harmelen, Erman Acar, and Annette ten Teije. 2025. Successful Misunderstandings: Learning to Coordinate Without Being Understood. doi:10.48550/ARXIV.2509.24660

[85] Tom Kouwenhoven, Max Peeperkorn, Roy De Kleijn, and Tessa Verhoef. 2025. Shaping Shared Languages: Human and Large Language Models' Inductive Biases in Emergent Communication. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2025, Montreal, Canada, August 16-22, 2025* (2025-01-01). ijcai.org, 10298–10306. doi:10.24963/IJCAI.2025/1144

[86] Lea Krause, Pia Sommerauer, and P. Vossen. 2022. Towards More Informative List Verbalisations. In *International Workshop on Knowledge Graph Summarization* (2022-01-01). https://ceur-ws.org/Vol-3257/paper14.pdf

[87] Lea Krause and Piek Vossen. 2020. When to explain: Identifying explanation triggers in human-agent interaction. In *2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence* (2020-11-01). Association for Computational Linguistics, Dublin, Ireland, 55–60. https://www.aclweb.org/anthology/2020.nl4xai-1.12

[88] Valerie Krug, Raihan Kabir Ratul, Christopher Olson, and Sebastian Stober. 2023. Visualizing deep neural networks with topographic activation maps. In *HHAI 2023: Augmenting Human Intellect: Proceedings of the Second International Conference on Hybrid Human-Artificial Intelligence*. SAGE Publications 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP, 138–152.

[89] Florian Kunneman and Koen V Hindriks. 2022. A Sequence-Based Dialog Management Framework for Co-Regulated Dialog.. In *HHAI*. 143–156.

[90] Nadin Kökciyan and Pinar Yolum. 2022. Taking Situation-Based Privacy Decisions: Privacy Assistants Working with Humans. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI)* (2022-07-01). 703–709. https://www.ijcai.org/proceedings/2022/0099.pdf

[91] Heliodoro Tejeda Lemus, Aakriti Kumar, and Mark Steyvers. 2022. An Empirical Investigation of Reliance on AI-Assistance in a Noisy-Image Classification Task.. In *HHAI*. 225–237.

[92] Roger X. Lera-Leri, Enrico Liscio, Filippo Bistaffa, Catholijn M. Jonker, Maite Lopez-Sanchez, Pradeep K. Murukannaiah, Juan A. Rodriguez-Aguilar, and Francisco Salas-Molina. 2024. Aggregating value systems for decision support. *Knowledge-Based Systems* 287 (2024), 111453. https://www.sciencedirect.com/science/article/pii/S0950705124000881

[93] Sergey Levine, Aviral Kumar, G. Tucker, and Justin Fu. 2020. Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems. *ArXiv* abs/2005.01643 (2020). https://api.semanticscholar.org/CorpusID:218486979

[94] Joseph CR Licklider. 1960. Man-computer symbiosis. *IRE transactions on human factors in electronics* HFE-1 (1960), 4–11.

[95] Mike Ligthart, Mark Neerincx, and Koen Hindriks. 2022. Memory-Based Personalization for Fostering a Long-Term Child-Robot Relationship. In *ACM/IEEE International Conference on Human-Robot Interaction, HRI 2022, Sapporo, Hokkaido, Japan, March 7 - 10, 2022* (2022-01-01), Daisuke Sakamoto, Astrid Weiss, Laura M. Hiatt, and Masahiro Shiomi (Eds.). IEEE / ACM, 80–89. doi:10.1109/HRI53351.2022.9889446

[96] Mike E.U. Ligthart, Simone M. de Droog, Marianne Bossema, Lamia Elloumi, Kees Hoogland, Matthijs Smakman, Koen V. Hindriks, and Somaya Ben Allouch. 2023. Design Specifications for a Social Robot Math Tutor. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (2023-01-01). doi:10.1145/3568162.3576957

[97] Adam Dahlgren Lindström, Leila Methnani, Lea Krause, Petter Ericson, Íñigo Martinez de Rituerto de Troya, Dimitri Coelho Mollo, and Roel Dobbe. 2024. AI Alignment through Reinforcement Learning from Human Feedback? Contradictions and Limitations. *CoRR* abs/2406.18346 (2024). arXiv: 2406.18346. doi:10.48550/ARXIV.2406.18346

[98] Enrico Liscio, Alin E. Dondera, Andrei Geadau, Catholijn M. Jonker, and Pradeep K. Murukannaiah. 2022. Cross-Domain Classification of Moral Values. In *Findings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics* (2022-01-01) *(NAACL '22)*. ACL, Seattle, WA, USA, 2727–2745. https://aclanthology.org/2022.findings-naacl.209.pdf

[99] Enrico Liscio, Michiel van der Meer, Luciano C. Siebert, Catholijn M. Jonker, Niek Mouter, and Pradeep K. Murukannaiah. 2021. Axies: Identifying and Evaluating Context-Specific Values. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021)* (2021-01-01). IFAAMAS, Online, 799–808. https://ii.tudelft.nl/~pradeep/doc/Liscio-2021-AAMAS-Axies.pdf

[100] Dou Liu, Claudia Alessandra Libbi, and Delaram Javdani Rikhtehgar. 2022. What Would You Like to Visit Next?–Using a Knowledge-Graph Driven Museum Guide in a Virtual Exhibition. In *HHAI2022: Augmenting Human Intellect*. IOS Press, 275–277.

[101] Robert Loftin and Frans A Oliehoek. 2022. On the Impossibility of Learning to Cooperate with Adaptive Partner Strategies in Repeated Games. In *International Conference on Machine Learning* (2022-01-01). PMLR, 14197–14209. https://arxiv.org/abs/2206.10614

[102] Robert Loftin, Mustafa Mert Çelikok, and Frans A. Oliehoek. 2023. Towards a Unifying Model of Rationality in Multiagent Systems. In *AAMAS Workshop on Optimization and Learning in Multiagent Systems* (2023-01-01). https://arxiv.org/abs/2305.18071

[103] Robert Loftin, Mustafa Mert Çelikok, Herke van Hoof, Samuel Kaski, and Frans A Oliehoek. 2024. Uncoupled Learning of Differential Stackelberg Equilibria with Commitments. In *Proceedings of the Twenty-Third International Conference on Autonomous Agents and Multiagent Systems (AAMAS)* (2024-05-01). IFAAMAS, Auckland, New Zealand. https://dl.acm.org/doi/abs/10.5555/3635637.3662984

[104] BF Lomov and VF Venda. 1978. Methodological principles of synthesis of hybrid intelligence systems. In *Proceedings of the International Conference on Cybernetics and Society*. 1026–1030.

[105] Jonne Maas. 2022. A Neo-Republican Critique of AI ethics. *Journal of Responsible Technology* 9 (2022), 100022. doi:10.1016/j.jrt.2021.100022

[106] Henry Maathuis, Daan Kolkman, Stefan Lehnen, and Danielle Sent. 2025. Formalizing Explanation Design through Interaction Patterns in Human-AI Decision Support. In *HHAI 2025: Proceedings of the 4th International Conference on Hybrid Human-Artificial Intelligence*. SAGE Publications 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP, 262–276.

[107] Elena Malnatsky. 2025. Towards Dialogic Education through Child-Robot Interaction: A Hybrid Child-Centered Framework. In *Proceedings of the 24th Interaction Design and Children*. 1214–1216.

[108] Elena Malnatsky and Mike E.U. Ligthart. 2025. Fitting Humor: Age-Based Personalization for Shaping Relatable Child-Robot Interactions. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction* (2025-03-01) *(HRI '25)*. IEEE Press, Melbourne, Australia, 331–341.

[109] Elena Malnatsky, Shenghui Wang, Koen V. Hindriks, and Mike E. U. Ligthart. 2025. Dialogic Learning in Child–Robot Interaction: A Hybrid Approach to Personalized Educational Content Generation. In *Proceedings of the 2025 AAAI Spring Symposium Series (AAAI-SS 2025)* (2025-01-01). AAAI Press, 416–420. https://ojs.aaai.org/index.php/AAAI-SS/article/view/35622

[110] Thomas W Malone and Michael Bernstein. 2015. *Handbook of collective intelligence*. MIT press.

[111] Adwitiya Mandal and Kim Baraka. 2024. Using Proxemics as a Corrective Feedback Signal during Robot Navigation. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction, HRI 2024, Boulder, CO, USA, March 11-15, 2024* (2024-01-01), Dan Grollman, Elizabeth Broadbent, Wendy Ju, Harold Soh, and Tom Williams (Eds.). ACM, 732–736. doi:10.1145/3610978.3640746

[112] Yaoli Mao, Janet Rafner, Yi Wang, and Jacob Sherson. 2023. A hybrid intelligence approach to training generative design assistants: partnership between human experts and AI enhanced co-creative tools. In *HHAI 2023: augmenting human intellect*. IOS Press, 108–123.

[113] Antoine Marot, Alexandre Rozier, Matthieu Dussartre, Laure Crochepierre, and Benjamin Donnot. 2022. Towards an AI Assistant for Power Grid Operators. In *HHAI2022: Augmenting Human Intellect*. IOS Press, 79–95.

[114] Margherita Martorana, Francesca Urgese, Mark Adamik, and Ilaria Tiddi. 2025. Bridging Bots: from Perception to Action via Multimodal-LMs and Knowledge Graphs. In *19th International Conference on Neurosymbolic Learning and Reasoning*.

[115] Tiffany Matej Hrkalovic. 2022. Designing Hybrid Intelligence Techniques for Facilitating Collaboration Informed by Social Science. In *Proceedings of the 2022 International Conference on Multimodal Interaction* (2022-01-01). 679–684. doi:10.1145/3536221.3557032

[116] Tiffany Matej Hrkalovic, Bernd Dudzik, Hayley Hung, and Daniel Balliet. 2025. Partner perceptions during brief online interactions shape partner selection and cooperation. *PloS one* 20, 4 (2025), e0318137.

[117] Michiel van der Meer. 2024. Facilitating Opinion Diversity through Hybrid NLP Approaches. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop, NAACL 2024, Mexico City, Mexico, June 18, 2024* (2024-01-01), Yang (Trista) Cao, Isabel Papadimitriou, Anaelia Ovalle, Marcos Zampieri, Francis Ferraro, and Swabha Swayamdipta (Eds.). Association for Computational Linguistics, 272–284. doi:10.18653/V1/2024.NAACL-SRW.29

[118] Michiel van der Meer, Neele Falk, Pradeep K. Murukannaiah, and Enrico Liscio. 2024. Annotator-Centric Active Learning for Subjective NLP Tasks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024* (2024-01-01), Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, 18537–18555. https://aclanthology.org/2024.emnlp-main.1031

[119] Michiel van der Meer, Enrico Liscio, Catholijn M. Jonker, Aske Plaat, Piek Vossen, and Pradeep K. Murukannaiah. 2024. A Hybrid Intelligence Method for Argument Mining. *J. Artif. Intell. Res.* 80 (2024), 1187–1222. doi:10.1613/JAIR.1.15135

[120] Michiel van der Meer, Piek Vossen, Catholijn M. Jonker, and Pradeep K. Murukannaiah. 2024. Value-Sensitive Disagreement Analysis for Online Deliberation. In *HHAI 2024: Hybrid Human AI Systems for the Social Good - Proceedings of the Third International Conference on Hybrid Human-Artificial Intelligence, Malmö, Sweden, 10-14 June 2024* (2024-01-01) *(Frontiers in Artificial Intelligence and Applications, Vol. 386)*, Fabian Lorig, Jason Tucker, Adam Dahlgren Lindström, Frank Dignum, Pradeep K. Murukannaiah, Andreas Theodorou, and Pinar Yolum (Eds.). IOS Press, 481–484. doi:10.3233/FAIA240231

[121] Siddharth Mehrotra, Chadha Degachi, Oleksandra Vereschak, Catholijn M Jonker, and Myrthe L Tielman. 2024. A Systematic Review on Fostering Appropriate Trust in Human-AI Interaction: Trends, Opportunities and Challenges. *ACM Journal on Responsible Computing* (2024). Publisher: ACM New York, NY. doi:10.1145/3696449

[122] Siddharth Mehrotra, Ujwal Gadiraju, Eva Bittner, Folkert van Delden, Catholijn M. Jonker, and Myrthe L. Tielman. 2025. "Even explanations will not help in trusting [this] fundamentally biased system": A Predictive Policing Case-Study. In *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization* (2025-01-01) *(UMAP '25)*. Association for Computing Machinery, New York, NY, USA, 51–62. doi:10.1145/3699682.3728343

[123] Siddharth Mehrotra, Catholijn M. Jonker, and Myrthe L. Tielman. 2021. More Similar Values, More Trust? - The Effect of Value Similarity on Trust in Human-Agent Interaction. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (2021-01-01) *(AIES '21)*. Association for Computing Machinery, New York, NY, USA, 777–783. event-place: Virtual Event, USA. doi:10.1145/3461702.3462576

[124] Siddharth Mehrotra, Carolina Centeio Jorge, Catholijn M. Jonker, and Myrthe L. Tielman. 2023. Building Appropriate Trust in AI: The Significance of Integrity-Centered Explanations. In *HHAI 2023: Augmenting Human Intellect* (2023-06-01) *(Frontiers in Artificial Intelligence and Applications, Vol. 368)*. IOS Press Ebooks, 436 – 439. doi:10.3233/FAIA230121

[125] Matteo Michelini, Adrian Haret, and Davide Grossi. 2022. Group Wisdom at a Price: Jury Theorems with Costly Information. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22* (2022-07-01), Lud De Raedt (Ed.). International Joint Conferences on Artificial Intelligence Organization, 419–425. doi:10.24963/ijcai.2022/60

[126] Shubhalaxmi Mukherjee, Catholijn M Jonker, and Pradeep K Murukannaiah. 2025. Exploring Human-AI Synergy for Complex Claim Verification. In *HHAI 2025: Proceedings of the 4th International Conference on Hybrid Human-Artificial Intelligence*. SAGE Publications 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP, 2–15.

[127] Pradeep K. Murukannaiah, Nirav Ajmeri, Catholijn M. Jonker, and Munindar P. Singh. 2020. New Foundations of Ethical Multiagent Systems. In *Proceedings of the 19th Conference on Autonomous Agents and MultiAgent Systems* (2020-01-01) *(AAMAS '20)*. Auckland, 1706–1710. https://ii.tudelft.nl/~pradeep/doc/Murukannaiah-2020-AAMASBlueSky-EthicalMAS.pdf

[128] Chiara Natali, Mohammad Naiseh, Federico Cabitza, and Brett Frischmann. 2025. Better AI with Designed Friction: Theories, Applications and Research Agenda. In *HHAI 2025: Proceedings of the 4th International Conference on Hybrid Human-Artificial Intelligence*. SAGE Publications 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP, 518–520.

[129] Michael Neely, Stefan F Schouten, Maurits Bleeker, and Ana Lucic. 2022. A Song of (Dis) agreement: Evaluating the Evaluation of Explainable Artificial Intelligence in Natural Language Processing. In *HHAI2022: Augmenting Human Intellect*. IOS Press, 60–78.

[130] Mark Neerincx, Jasper van der Waa, Myrthe Tielman, Chenxu Hao, Liv Ziegfeld, Davide Dell'Anna, and Shihan Wang. 2025. Social AI for a Healthier Lifestyle: Four Competencies to Manage and Prevent Chronic Diseases. In *Proceedings of the 4th International Conference on Hybrid Human-Artificial Intelligence, HHAI 2025: Hybrid Human AI Systems for the Social Good* (2025-01-01), Vol. 408. 317–332. doi:10.3233/FAIA250649

[131] Thong Nguyen, Mariya Hendriksen, Andrew Yates, and Maarten de Rijke. 2024. Multimodal Learned Sparse Retrieval with Probabilistic Expansion Control. In *Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part II* (2024-01-01) *(Lecture Notes in Computer Science, Vol. 14609)*, Nazli Goharian, Nicola Tonellotto, Yulan He, Aldo Lipani, Graham McDonald, Craig Macdonald, and Iadh Ounis (Eds.). Springer, 448–464. doi:10.1007/978-3-031-56060-6_29

[132] Annet Onnes. 2022. Monitoring AI systems: A Problem Analysis, Framework and Outlook. In *Proceedings of the First International Conference on Hybrid-Artificial Intelligence* (2022-01-01) *(Frontiers in Artificial Intelligence and Applications, Vol. 354)*. 238–240. doi:10.3233/FAIA220202

[133] Annet Onnes, Mehdi Dastani, and Silja Renooij. 2023. Bayesian Network Conflict Detection for Normative Monitoring of Black-Box Systems. In *Proceedings of the Thirty-Sixth FLAIRS Conference* (2023-01-01), Vol. 36. Florida Online Journals. https://journals.flvc.org/FLAIRS/article/view/133240/137859

[134] Annet Onnes and Silja Renooij. 2025. Maximum Entropy-based Quantification for Probability Elicitation in Bayesian Networks. In *The 18th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty* (2025-01-01). Accepted, in press.

[135] Nicole Orzan, Erman Acar, Davide Grossi, Patrick Mannion, and Roxana Radulescu. 2024. Learning in Multi-Objective Public Goods Games with Non-Linear Utilities. In *ECAI 2024 - 27th European Conference on Artificial Intelligence, 19-24 October 2024, Santiago de Compostela, Spain - Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024)* (2024-01-01) *(Frontiers in Artificial Intelligence and Applications, Vol. 392)*, Ulle Endriss, Francisco S. Melo, Kerstin Bach, Alberto José Bugarín Diz, Jose Maria Alonso-Moral, Senén Barro, and Fredrik Heintz (Eds.). IOS Press, 2749–2756. doi:10.3233/FAIA240809

[136] Nicole Orzan, Erman Acar, Davide Grossi, and Roxana Radulescu. 2024. Emergent Cooperation under Uncertain Incentive Alignment. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2024, Auckland, New Zealand, May 6-10, 2024* (2024-01-01), Mehdi Dastani, Jaime Simão Sichman, Natasha Alechina, and Virginia Dignum (Eds.). International Foundation for Autonomous Agents and Multiagent Systems / ACM, 1521–1530. doi:10.5555/3635637.3663012

[137] Nicole Orzan, Erman Acar, Davide Grossi, and Roxana Radulescu. 2025. Learning in public goods games: the effects of uncertainty and communication on cooperation. *Neural Comput. Appl.* 37, 23 (2025), 18899–18932. doi:10.1007/S00521-024-10530-6

[138] Patrizia Paci, Ilaria Tiddi, Daniel Preciado, and Kim Baraka. 2023. "Who's a Good Robot?!" Designing Human-Robot Teaching Interactions Inspired by Dog Training. In *HHAI 2023: Augmenting Human Intellect*. IOS Press, 310–319. https://ebooks.iospress.nl/pdf/doi/10.3233/FAIA230094

[139] H. Prakken and R. Ratsma. 2022. A top-level model of case-based argumentation for explanation: Formalisation and experiments. *Argument and Computation* 13 (2022), 159–194. https://content.iospress.com/articles/argument-and-computation/aac210009

[140] Muhammad A Rahman, Niklas Hopner, Filippos Christianos, and Stefano V Albrecht. 2021. Towards Open Ad Hoc Teamwork Using Graph-based Policy Learning. In *Proceedings of the 38th International Conference on Machine Learning* (2021-07-01) *(Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8776–8786. https://proceedings.mlr.press/v139/rahman21a.html

[141] Bogdana Rakova and Roel Dobbe. 2023. Algorithms as Social-Ecological-Technological Systems: an Environmental Justice Lens on Algorithmic Audits. In *2023 ACM Conference on Fairness, Accountability, and Transparency* (2023-01-01) *(FAccT '23)*. Association for Computing Machinery, Chicago, IL, USA. doi:10.1145/3593013.3594014

[142] Chirag Raman, Jose Vargas Quiros, Stephanie Tan, Ashraful Islam, Ekin Gedik, and Hayley Hung. 2022. ConfLab: A Data Collection Concept, Dataset, and Benchmark for Machine Analysis of Free-Standing Social Interactions in the Wild. In *Thirty-sixth Conference on Neural Information (NeurIPS) Processing Systems Datasets and Benchmarks Track* (2022-01-01). https://openreview.net/forum?id=CNJQKM5cV2o

[143] Maike Ramrath, Alisa Scharmann, Annika Ridder, Tristan Kühn, Sophie Weller, and Nicole Krämer. 2024. Trust in AI Chatbots: The perceived expertise of CHATGPT in subjective and objective tasks. In *HHAI 2024: Hybrid Human AI systems for the social good*. IOS Press, 264–280.

[144] Merle Reimann, Jesper van de Graaf, Nina van Gulik, Stephanie van de Sanden, Tibert Verhagen, and Koen Hindriks. 2024. Social Robots in the Wild and the Novelty Effect. In *Social Robotics* (2024-01-01), Abdulaziz Al Ali, John-John Cabibihan, Nader Meskin, Silvia Rossi, Wanyue Jiang, Hongsheng He, and Shuzhi Sam Ge (Eds.). Springer Nature Singapore, Singapore, 38–48. https://rdcu.be/dtXE5

[145] Merle M. Reimann, Florian A. Kunneman, Catharine Oertel, and Koen V. Hindriks. 2024. A Survey on Dialogue Management in Human-Robot Interaction. *J. Hum.-Robot Interact.* (2024). Place: New York, NY, USA Publisher: Association for Computing Machinery. doi:10.1145/3648605

[146] Merle M. Reimann, Catharine Oertel, Florian A. Kunneman, and Koen V. Hindriks. 2023. Predicting Interaction Quality Aspects Using Level-Based Scores for Conversational Agents. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents* (2023-01-01) *(IVA '23)*. Association for Computing Machinery, New York, NY, USA. event-place: , Würzburg, Germany,. doi:10.1145/3570945.3607332

[147] Silja Renooij. 2022. Relevance for Robust Bayesian Network MAP-Explanations. In *Proceedings of The 11th International Conference on Probabilistic Graphical Models* (2022-10-01) *(Proceedings of Machine Learning Research, Vol. 186)*, Antonio Salmerán and Rafael Rumí (Eds.). PMLR, 13–24. https://proceedings.mlr.press/v186/renooij22a.html

[148] Jennifer Renoux, Jasmin Grosinge, Marta Romeo, Kiran M. Sabu, Kim Baraka, and Victor Kaptelinin. 2024. Communication in Human-AI Interaction (Preface). In *Proceedings of the Workshops at the Third International Conference on Hybrid Human-Artificial Intelligence co-located with (HHAI 2024), Malmö, Sweden, June 10-11, 2024* (2024-01-01) *(CEUR Workshop Proceedings, Vol. 3825)*, Petter Ericson, Nina Khairova, and Marina De Vos (Eds.). CEUR-WS.org, 78–83. https://ceur-ws.org/Vol-3825/prefaceW3.pdf

[149] Bram M. Renting, Holger H. Hoos, and Catholijn M. Jonker. 2020. Automated Configuration of Negotiation Strategies. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems* (2020-05-01) *(AAMAS '20)*. International Foundation for Autonomous Agents and Multiagent Systems, 1116–1124. https://ifaamas.org/Proceedings/aamas2020/pdfs/p1116.pdf

[150] Arturo Rosenblueth, Norbert Wiener, and Julian Bigelow. 1943. Behavior, purpose and teleology. *Philosophy of science* 10, 1 (1943), 18–24.

[151] Fatemeh Sarvi, Maria Heuss, Mohammad Aliannejadi, Sebastian Schelter, and Maarten de Rijke. 2022. Understanding and Mitigating the Effect of Outliers in Fair Ranking. In *WSDM 2022: The Fifteenth International Conference on Web Search and Data Mining* (2022-02-01). ACM. https://arxiv.org/abs/2112.11251

[152] Andreas W. M. Sauter, Erman Acar, and Vincent Francois-Lavet. 2023. A Meta-Reinforcement Learning Algorithm for Causal Discovery. In *2nd Conference on Causal Learning and Reasoning* (2023-01-01). https://openreview.net/forum?id=p6NnDqJM_jL

[153] Jakob Schoeffer, Johannes Jakubik, Michael Voessing, Niklas Kuehl, and Gerhard Satzger. 2023. On the interdependence of reliance behavior and accuracy in AI-assisted decision-making. In *HHAI 2023: Augmenting Human Intellect: Proceedings of the Second International Conference on Hybrid Human-Artificial Intelligence*. SAGE Publications 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP, 46–59.

[154] Nienke Schrage-Prent, Daniel F. Preciado Vanegas, and Kim Baraka. 2024. Interactive Robot Programming Inspired by Dog Training: An Exploratory Study. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction, HRI 2024, Boulder, CO, USA, March 11-15, 2024* (2024-01-01), Dan Grollman, Elizabeth Broadbent, Wendy Ju, Harold Soh, and Tom Williams (Eds.). ACM, 965–969. doi:10.1145/3610978.3640655

[155] Ben Shneiderman. 2020. Human-centered artificial intelligence: Three fresh ideas. *AIS Transactions on Human-Computer Interaction* 12, 3 (2020), 109–124.

[156] Ruth Shortall, Anatol Itten, Michiel van der Meer, Pradeep Murukannaiah, and Catholijn Jonker. 2022. Reason against the machine? Future directions for mass online deliberation. *Frontiers in Political Science* 4 (2022). doi:10.3389/fpos.2022.946589

[157] Luciano C. Siebert, Enrico Liscio, Pradeep K. Murukannaiah, Lionel Kaptein, Shannon L. Spruit, Jeroen van den Hoven, and Catholijn M. Jonker. 2022. Estimating Value Preferences in a Hybrid Participatory System. In *HHAI2022: Augmenting Human Intellect* (2022-01-01) *(HHAI '22)*. IOS Press, Amsterdam, the Netherlands, 114–127. https://ebooks.iospress.nl/volumearticle/60861

[158] Nadiya Slobodenyuk. 2025. The Future of Hybrid Moral Decision-Making: Exploring the Alignment of Default Moral Reasoning Architectures in Humans and LLMs. In *HHAI 2025: Proceedings of the 4th International Conference on Hybrid Human-Artificial Intelligence*. 138–151.

[159] Cor Steging, Silja Renooij, and Bart Verheij. 2021. Rationale Discovery and Explainable AI. In *Legal Knowledge and Information Systems - JURIX 2021: The Thirty-fourth Annual Conference, Vilnius, Lithuania, 8-10 December 2021* (2021-01-01) *(Frontiers in Artificial Intelligence and Applications, Vol. 346)*, Schweighofer Erich (Ed.). IOS Press, 225–234. doi:10.3233/FAIA210341

[160] C. Steging, S. Renooij, and B. Verheij. 2022. Discovering the Rationale of Decisions. In *Proceedings of the First International Conference on Hybrid Human-Artificial Intelligence* (2022-06-01), S. Schlobach, M. Pérez-Ortiz, and M. Tielman (Eds.), Vol. 354. IOS Press, Amsterdam, the Netherlands, 255–257. doi:10.3233/FAIA220208

[161] C. Steging, S. Renooij, and B. Verheij. 2023. Improving rationales with small, inconsistent and incomplete data. In *Legal Knowledge and Information Systems - JURIX 2023: The Thirty-Sixth Annual Conference* (2023-12-01), G. Sileno, J. Spanakis, and G. van Dijck (Eds.), Vol. 379. IOS Press, Maastricht, the Netherlands, 53–62. https://ebooks.iospress.nl/volumearticle/65567

[162] Cor Steging, Silja Renooij, Bart Verheij, and Trevor Bench-Capon. 2023. Arguments, rules and cases in law: Resources for aligning learning and reasoning in structured domains. *Argument & Computation* 14 (2023), 235–243. Publisher: IOS Press. doi:10.3233/AAC-220017

[163] C. Steging and L. van Leeuwen. 2024. A hybrid approach to legal textual entailment. In *Eighteenth International Workshop on Juris-Informatics (JURISIN 2024)* (2024-05-01). Hamamatsu, Japan. https://jurisinformaticscenter.github.io/jurisin2024/jurisin2024local-proceedings.pdf

[164] Cor Steging and Tadeusz Zbiegien. 2025. Label Indeterminacy in AI & Law. In *Legal Knowledge and Information Systems - JURIX 2025: The Thirty-Eigth Annual Conference on Legal Knowledge and Information Systems* (2025-01-01). Proceedings not yet published.. https://doi.org/10.

48550/arXiv.2510.17463

[165] Rebecca Stower, Mike E.U. Ligthart, Micol Spitale, Natalia Calvo-Barajas, and Simone M. de Droog. 2023. CRITTER: Child-Robot Interaction and Interdisciplinary Research. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (2023-01-01). doi:10.1145/3568294.3579955

[166] Xin Sun, Dimosthenis Casula, Arathy Navaratnam, Anna Popp, Franziska Knopp, Giovanni Busini, Jan Wesolowski, Marie Van Reeth, Elke Reich, Reinout W Wiers, et al. 2023. Virtual Support for Real-World Movement: Using Chatbots to Overcome Barriers to Physical Activity.. In *HHAI 2023: Augmenting Human Intellect: Proceedings of the Second International Conference on Hybrid Human-Artificial Intelligence*. SAGE Publications, 201–214.

[167] Agnes Tegen, Paul Davidsson, and Jan A Persson. 2023. Human Factors in Interactive Online Machine Learning. In *HHAI 2023: Augmenting Human Intellect: Proceedings of the Second International Conference on Hybrid Human-Artificial Intelligence*. SAGE Publications 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP, 33–45.

[168] Ilaria Tiddi, Victor de Boer, Stefan Schlobach, and André Meyer-Vitali. 2023. Knowledge Engineering for Hybrid Intelligence. In *Proceedings of the 12th Knowledge Capture Conference 2023* (2023-01-01) *(K-CAP '23)*. Association for Computing Machinery, New York, NY, USA, 75–82. event-place: Pensacola, FL, USA,. doi:10.1145/3587259.3627541

[169] Emily Timmerman and Mike E.U. Ligthart. 2023. Let's Roll Together: Children Helping a Robot Play a Dice Game. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (2023-01-01). doi:10.1145/3568294.3580130

[170] Jakob Dirk Top, Harmen de Weerd, Abisharan Raveenthiran, Catholijn M Jonker, and Rineke Verbrugge. 2025. Cooperation, Deception and Theory of Mind in a Cyclic Game with Inter-Player Signalling. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (2025-01-01), Vol. 47. https://escholarship.org/uc/item/1466g95n

[171] Vito Trianni, Andrea Giovanni Nuzzolese, Jaron Porciello, Ralf HJM Kurvers, Stefan M Herzog, Gioele Barabucci, Aleksandra Berditchevskaia, and Fai Fung. 2023. Hybrid collective intelligence for decision support in complex open-ended domains. In *HHAI 2023: Augmenting Human Intellect: Proceedings of the Second International Conference on Hybrid Human-Artificial Intelligence*. SAGE Publications 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP, 124–137.

[172] Eric L Trist et al. 1981. *The evolution of socio-technical systems*. Vol. 2. Ontario Quality of Working Life Centre Toronto.

[173] Sergey Troshin, Wafaa Mohammed, Yan Meng, Christof Monz, Antske Fokkens, and Vlad Niculae. 2025. Control the Temperature: Selective Sampling for Diverse and High-Quality LLM Outputs. *CoRR* abs/2510.01218 (2025). doi:10.48550/ARXIV.2510.01218

[174] Sergey Troshin, Irina Saparina, Antske Fokkens, and Vlad Niculae. 2025. Asking a Language Model for Diverse Responses. *CoRR* abs/2509.17570 (2025). doi:10.48550/ARXIV.2509.17570

[175] Anna Trunk, Hendrik Birkel, and Evi Hartmann. 2020. On the current state of combining human and artificial intelligence for strategic organizational decision making. *Business Research* 13, 3 (2020), 875–919.

[176] Maria Tsfasman, Kristian Fenech, Morita Tarvirdians, Andras Lorincz, Catholijn Jonker, and Catharine Oertel. 2022. Towards Creating a Conversational Memory for Long-Term Meeting Support: Predicting Memorable Moments in Multi-Party Conversations through Eye-Gaze. In *Proceedings of the 2022 International Conference on Multimodal Interaction* (2022-01-01) *(ICMI '22)*. Association for Computing Machinery, New York, NY, USA, 94–104. event-place: Bengaluru, India. doi:10.1145/3536221.3556613

[177] Abdullah Aman Tutul, Theodora Chaspari, Sarah Ita Levitan, and Julia Hirschberg. 2025. Detecting Trust Calibration Traits in AI with EEG Signals for Speech Deception. In *HHAI 2025: Proceedings of the 4th International Conference on Hybrid Human-Artificial Intelligence*. SAGE Publications 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP, 122–137.

[178] Onuralp Ulusoy and Pınar Yolum. 2020. Agents for Preserving Privacy: Learning and Decision Making Collaboratively. In *Multi-Agent Systems and Agreement Technologies* (2020-01-01), Nick Bassiliades, Georgios Chalkiadakis, and Dave de Jonge (Eds.). Springer International Publishing, 116–131. doi:10.1007/978-3-030-66412-1_8

[179] Onuralp Ulusoy and Pinar Yolum. 2022. PANOLA: A Personal Assistant for Supporting Users in Preserving Privacy. *ACM Transactions on Internet Technology* 22, 1 (2022). Place: New York, NY, USA Publisher: Association for Computing Machinery. https://webspace.science.uu.nl/~yolum001/papers/panola-2022.pdf

[180] Martin Van Den Berg, Ouren Kuiper, Yvette Van Der Haas, Julie Gerlings, Danielle Sent, and Stefan Leijnen. 2023. A conceptual model for implementing explainable ai by design: Results of an empirical study. In *HHAI 2023: Augmenting Human Intellect: Proceedings of the Second International Conference on Hybrid Human-Artificial Intelligence*. SAGE Publications 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP, 60–73.

[181] Michiel van der Meer, Enrico Liscio, Catholijn M. Jonker, Aske Plaat, Piek Vossen, and Pradeep K. Murukannaiah. 2022. HyEnA: A Hybrid Method for Extracting Arguments from Opinions. In *HHAI2022: Augmenting Human Intellect* (2022-01-01) *(HHAI '22)*. IOS Press, Amsterdam, the Netherlands, 17–31. https://ebooks.iospress.nl/volumearticle/60855

[182] Ramira Van Der Meulen, Rineke Verbrugge, and Max Van Duijn. 2024. Common ground provides a mental shortcut in agent-agent interaction. *3rd International Conference on Hybrid Human-Artificial Intelligence, HHAI 2024* (2024), 281–290.

[183] Ramira van der Meulen, Rineke Verbrugge, and Max van Duijn. 2025. Finding Common Ground: Comparing Two Computational Models of Social Intelligence. *ToM4AI 2025* (2025), 64.

[184] Emile van Krieken, Erman Acar, and Frank van Harmelen. 2022. Analyzing differentiable fuzzy logic operators. *Artificial Intelligence* 302 (2022), 103602. Publisher: Elsevier. https://research.vu.nl/ws/portalfiles/portal/146020254/2002.06100v2.pdf

[185] Jan-Willem van Rhenen, Carolina Centeio Jorge, Tiffany Matej Hrkalovic, and Bernd Dudzik. 2022. Effects of Social Behaviours in Online Video Games on Team Trust. In *Extended Abstracts of the 2022 Annual Symposium on Computer-Human Interaction in Play* (2022-01-01). 159–165. https://pure.tudelft.nl/admin/files/146989190/vanRhenen2021_author.pdf

[186] Wijnand van Woerkom, Davide Grossi, Henry Prakken, and Bart Verheij. 2024. A Case-Based-Reasoning Analysis of the COMPAS Dataset. In *Legal Knowledge and Information Systems. JURIX 2024: The Thirty-seventh Annual Conference*. IOS Press. https://ebooks.iospress.nl/volumearticle/71009

[187] Maaike Venema-Los, Zoé Christoff, and Davide Grossi. 2025. On the graph theory of majority illusions: theoretical results and computational experiments. *Auton. Agents Multi Agent Syst.* 39, 2 (2025), 39. doi:10.1007/S10458-025-09720-W

[188] Bart Verheij. 2020. Artificial intelligence as law. *Artif. Intell. Law* 28, 2 (2020), 181–206. doi:10.1007/s10506-020-09266-0

[189] Mehul Verma and Erman Acar. 2022. Learning to Cooperate with Human Evaluative Feedback and Demonstrations. In *HHAI2022: Augmenting Human Intellect*. IOS Press, 46–59. https://ebooks.iospress.nl/volumearticle/60857

[190] Rajeev Verma and Eric Nalisnick. 2022. Calibrated Learning to Defer with One-vs-All Classifiers. In *ICML 2022 Workshop on Human-Machine Collaboration and Teaming* (2022-01-01). https://icml.cc/Conferences/2022/ScheduleMultitrack?event=18123

[191] Ludwig von Bertalanffy. 1968. *General Systems Theory: Foundations, Development, Applications*. George Braziller, New York.

[192] Piek Vossen, Selene Báez Santamaría, Lenka Bajčetić, and Thomas Belluci. 2025. Extracting triples from dialogues for conversational social agents. 408 (2025), 179–195. https://ebooks.iospress.nl/doi/10.3233/FAIA250636

[193] Veronika Vzigraiova, Aparajita Chowdhury, Aino Ahtinen, and Kaisa Vaananen. 2025. Trust in Warehouse Automation: A Generational Study of Workers and Automated Guided Vehicles. In *HHAI 2025*. IOS Press, 108–121.

[194] Shenghui Wang, Daria Kulyk, Delaram Javdani Rikhtehgar, Dirk Heylen, and Carolien Rieffe. 2023. Correlating Eye Gaze with Object to Enrich Cultural Heritage Knowledge Graph. In *CEUR workshop proceedings* (2023-01-01), Vol. 3632. Rheinisch Westfälische Technische Hochschule. https://ceur-ws.org/Vol-3632/ISWC2023_paper_429.pdf

[195] Sachini Weerawardhana, Michael Akintunde, and Luc Moreau. 2024. Exploring the dynamic nature of trust using interventions in a human-AI collaborative task. In *The Third International Conference on Hybrid Human-Artificial Intelligence*.

[196] Wijnand van Woerkom, Davide Grossi, Henry Prakken, and Bart Verheij. 2024. A Fortiori Case-Based Reasoning: From Theory to Data. *J. Artif. Intell. Res.* 81 (2024), 401–441. doi:10.1613/JAIR.1.15178

[197] Johanna Wolff, Victor de Boer, Dirk Heylen, and Birna van Riemsdijk. 2023. Using Non-Monotonic Reasoning for Understandable and Flexible User-Models. In *Proceedings of the 21st International Workshop on Non-Monotonic Reasoning co-located with the 20th International Conference on Principles of Knowledge Representation and Reasoning (KR 2023) and co-located with the 36th International Workshop on Description Logics (DL 2023), Rhodes, Greece, September 2-4, 2023* (2023-01-01) *(CEUR Workshop Proceedings, Vol. 3464)*, Kai Sauerwald and Matthias Thimm (Eds.). CEUR-WS.org, 133–136. https://ceur-ws.org/Vol-3464/short5.pdf

[198] Johanna Wolff, Victor De Boer, Dirk Heylen, and M Birna Van Riemsdijk. 2024. Using Default Logic to Create Adaptable User Models for Behavior Support Agents. In *HHAI 2024: Hybrid Human AI Systems for the Social Good*. IOS Press, 350–359. doi:10.3233/FAIA240207

[199] Phillip Wozny, Bram Renting, Robert Loftin, Claudia Wieners, and Erman Acar. 2023. AI4GCC-Team–Below Sea Level: Score and Real World Relevance. *arXiv preprint arXiv:2307.13892* (2023).

[200] Atefeh Keshavarzi Zafarghandi, Ji Qi, Laura Hollink, Erik Tjong Kim Sang, and Davide Ceolin. 2024. Investigating the Usefulness of Product Reviews Through Bipolar Argumentation Frameworks. In *ICWE* (2024-01-01). Springer. https://link.springer.com/chapter/10.1007/978-3-031-62362-2_21

[201] Jiayi Eurus Zhang, Bernhard Hilpert, Joost Broekens, and Jussi PP Jokinen. 2024. Simulating Emotions With an Integrated Computational Model of Appraisal and Reinforcement Learning. 1–12. https://dl.acm.org/doi/10.1145/3613904.3641908

[202] Yuzhe Zhang and Davide Grossi. 2022. Tracking Truth by Weighting Proxies in Liquid Democracy. In *21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2022, Auckland, New Zealand, May 9-13, 2022* (2022-01-01), Piotr Faliszewski, Viviana Mascardi, Catherine Pelachaud, and Matthew E. Taylor (Eds.). International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 1482–1490. doi:10.5555/3535850.3536015

[203] Yangyang Zhao, Mehdi Dastani, Jinchuan Long, Zhenyu Wang, and Shihan Wang. 2024. Rescue Conversations from Dead-ends: Efficient Exploration for Task-oriented Dialogue Policy Optimization. *Transactions of the Association for Computational Linguistics* 12 (2024), 1578–1596. Place: Cambridge, MA Publisher: MIT Press. doi:10.1162/tacl_a_00717

[204] Yangyang Zhao, Ben Niu, Mehdi Dastani, and Shihan Wang. 2024. Bootstrapped Policy Learning for Task-oriented Dialogue through Goal Shaping. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024* (2024-01-01). Association for Computational Linguistics, 4566–4580. https://aclanthology.org/2024.emnlp-main.263

[205] Yangyang Zhao, Ben Niu, Libo Qin, and Shihan Wang. 2025. An Efficient Task-Oriented Dialogue Policy: Evolutionary Reinforcement Learning Injected by Elite Individuals. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2025-07-01), Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 3429–3442. doi:10.18653/v1/2025.acl-long.171

[206] Yangyang Zhao, Ben Niu, Yuxuan Tan, Shihan Wang, and Libo Qin. 2025. Semantic-Aware Action Space Compression via LLM-DRL Synergy for Efficient Task-oriented Dialogue Policy Exploration. In *Findings of the Association for Computational Linguistics: EMNLP 2025* (2025-11-01), Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 17808–17820. doi:10.18653/v1/2025.findings-emnlp.968

[207] Changxi Zhu, Mehdi Dastani, and Shihan Wang. 2024. A survey of multi-agent deep reinforcement learning with communication. *Autonomous Agents and Multi-Agent Systems* 38, 1 (2024), 4. Publisher: Springer. https://doi.org/10.1007/s10458-023-09633-6

[208] Emma M. van Zoelen, Tina Mioch, Mani Tajaddini, Christian Fleiner, Stefani Tsaneva, Pietro Camin, Thiago S. Gouvêa, Kim Baraka, Maaike H. T. de Boer, and Mark A. Neerincx. 2023. Developing Team Design Patterns for Hybrid Intelligence Systems. In *HHAI 2023: Augmenting Human Intellect - Proceedings of the Second International Conference on Hybrid Human-Artificial Intelligence, June 26-30, 2023, Munich, Germany* (2023-01-01) *(Frontiers in Artificial Intelligence and Applications, Vol. 368)*, Paul Lukowicz, Sven Mayer, Janin Koch, John Shawe-Taylor, and Ilaria Tiddi (Eds.). IOS Press, 3–16. doi:10.3233/FAIA230071

[209] Mustafa Mert Çelikok, Frans A. Oliehoek, and Samuel Kaski. 2022. Best-Response Bayesian Reinforcement Learning with Bayes-Adaptive POMDPs for Centaurs. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems* (2022-01-01) *(AAMAS '22)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 235–243. event-place: Virtual Event, New Zealand. https://ifaamas.org/Proceedings/aamas2022/pdfs/p235.pdf

Appendix: Survey Methodology

We answer our questions (RQ1/Evolution, RQ2/Anatomy) using a lightweight PRISMA procedure, where the publications of the HI Centre and the HHAI community are analysed and synthesised with the lenses of the CARE framework.

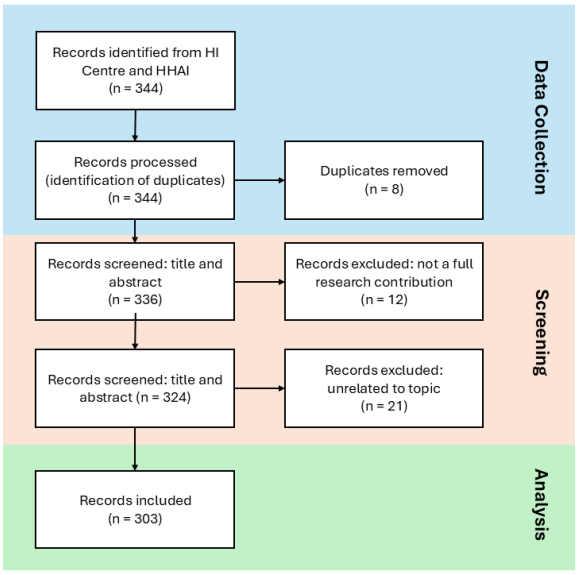

Fig. 1. PRISMA diagram as applied to our review.

Corpus Construction. We constructed a comprehensive corpus consisting of all publications produced within the HI consortium from the start of the project until the end of 2025. This corpus includes peer-reviewed material from journals, conferences, and workshops including: (i) formal theory papers; (ii) design/modelling papers presenting systems and (algorithmic) techniques; (iii) empirical work/hypothesis testing papers collecting, analysing and interpreting observations about a given hypothesis; (iv) surveys and (v) position papers. This selection is aimed to reflect the different stages of research maturity across the literature. In addition, we included full research papers from the Hybrid Human–Artificial Intelligence (HHAI) conference series in the same timespan. The HHAI conference series originated directly from the HI consortium and has functioned as its primary external and independent dissemination venue, attracting contributions that explicitly engage with the HI research agenda. As such, HHAI can be considered a focused extension of the consortium's research ecosystem rather than a generic human–AI interaction venue. Notice that the decision not to extend the corpus to a broader set of journals and conferences in AI or HCI was deliberate, to maintain conceptual coherence and analytical focus to the HI research field. Recall that our selection is meant to provide a reflexive analysis of the evolution of the HI field w.r.t. the HI research agenda and its research lines (cfr. RQ1), and not to capture the broader AI, HCI or Human-AI interaction landscape. In other words, this methodology is meant of provide a structured and transparent analysis of how the HI discipline and scientific community have evolved since the formulation of the original HI agenda, rather than an exhaustive literature survey.

This resulted in a total of 344 publications identified. Upon initial screening, we removed 8 duplicate entries. Two exclusion criteria were applied: records were removed when they were not a full research contribution (e.g. only an

abstract) and when the papers were topically not related to any of the HI concepts or research questions. This resulted in a total of 303 publications that were analysed (see Fig. 1).

Corpus Annotation. To enable a systematic and multi-dimensional analysis, we defined a structured annotation schema based on the CARE lines, capturing both conceptual and methodological aspects of HI research. Each publication was annotated along the following dimensions:

- Research questions addressed by the paper (explicitly stated or summarised by the annotator);
- Primary CARE line (Collaborativeness, Adaptiveness, Responsibility, Explainability);
- Secondary CARE research lines, where applicable;
- Key proposed methods and enabling technologies;
- Outcomes, described in terms of how the paper addresses or answers its research questions;
- Metrics and evaluation methods, when present;
- Case-study or application domain, if the paper is grounded in a concrete use case;
- Publication type, indicating research maturity;
- Relevant HI research questions, derived from the original research agenda.

Bibliographic metadata, including title, year, venue, authors, DOI, URL, and abstract, were automatically extracted from the associated BibTeX file using a scripted pipeline, which also handled deduplication based on unique identifiers and metadata similarity. This ensured consistency across records and minimised manual transcription errors.

Annotations were carried out by approx. 20 junior and senior researchers within the HI consortium, whose combined expertise spans across CARE research lines and the core HI challenges. This ensured that publications were interpreted by annotators with relevant theoretical, methodological, and domain-specific knowledge. The annotation was conducted through a combination of individual annotation and designated physical and online joint annotation sessions. The process started with an initial joint session aimed at establishing a shared understanding of the annotation schema, terminology, and categorisation criteria. Following this calibration phase, annotators independently annotated subsets of publications. Subsequent joint sessions were organised periodically and served primarily to clarify ambiguities, discuss borderline cases, and align interpretations across annotators. When differing views arose, annotators discussed their reasoning during these sessions and reached agreement through consensus. This collaborative calibration process functioned as an explicit quality control mechanism, helping to ensure consistency across annotations while accommodating the interpretive nature of the task. Given the conceptual breadth and interdisciplinary scope of HI research, we opted for consensus-based alignment during joint sessions rather than formal inter-annotator agreement metrics. This approach allowed for nuanced discussion of complex and multi-dimensional contributions, which are difficult to capture through purely quantitative agreement measures, while still ensuring transparency and consistency in the resulting annotations.

Received 10 July 2026