Autonomous Research Assistants for Hybrid Intelligence: Landscape and Challenges

Giacomo Zamprogno[1], **Ilaria Tiddi**[1], **Bart Verheij**[2]
[1]Vrije Universiteit Amsterdam
[2]University of Groningen
g.zamprogno@vu.nl, i.tiddi@vu.nl, bart.verheij@rug.nl

Abstract

We present an overview of AI-based tools assisting the research process, analyzing them from the point of view of Hybrid (human-AI) Intelligence (HI). While Autonomous Research Assistants (RAs) are gaining new interest by the latest advancements in AI (cf. Large Language Models), limitations arise when deployed in real-world use-cases. Starting from the hypothesis that principles from the emerging field of HI could enhance the synergy between researchers and AI tools, we explore what requirements allow to create HI RAs, using a survey of existing systems. We performed a review of 47 relevant articles published in the last 10 years, and we analyzed them according to various capabilities and characteristics proposed in the Hybrid Intelligence literature. Finally, we identify which future research lines could be followed to develop assistive systems that better combine the capabilities of humans and artificial RAs in a synergistic way.

Introduction

The process of scientific discovery has seen an extremely fast growth in recent decades, aided by faster communication means and more potent analytical tools (National Science Board, National Science Foundation 2023). Such a growth created in turn unprecedented challenges: keeping up with the sheer number of yearly increasing new publications is practically unfeasible, which, in turn, increases the difficulty of creating novel hypotheses and validating them starting from existing literature (Sarewitz 2016).

Clearly, advances in Artificial Intelligence (AI) also have the potential to facilitate the creation of systems that support researchers sifting through literature and in general throughout the research process. The idea of scholarly research assistants is far from new, dating back to the early stage of AI research (Langley 2000; Walker 1987), but with the development of technologies like Knowledge Graphs (Hogan et al. 2021) and Large-Language Models (Meyer et al. 2023) for storing, annotating and writing documents, it is natural to expect to see their application on the scientific domain, too. Since collaboration among researchers is quite common in the creative process of scientific discovery, the increasing reliance on AI-based tools is likely to generate a large number of Human-AI co-creation scenarios (Wu et al. 2021).

Hybrid Intelligence (HI) (Dellermann et al. 2019; Akata et al. 2020) represents a promising field to guide and analyze the development of such co-creative processes: at its core, HI aims at creating systems where humans and autonomous agents co-exist and cooperate to enhance each other's performance. Research assistant tools represent therefore a typical HI use-case: the nature of assistive tools is collaborative by definition and usually requires adapting to the user or the task, while the scientific domain has requisites for explainability (due to the principles of reproducible science) and responsibility (science ethics) (Wilkinson et al. 2016).

To study how HI can support the development of RAs, in this work we gather recently developed systems that provide assistive capabilities in research and analyze them under the lenses of HI principles. We collect 47 publications proposing tools that tackle various parts of the scientific process, and we analyze them using a framework we adapt from qualities and features available in the HI literature.

The contributions of this work are the following: first, we propose and discuss an analytical framework for evaluating the hybrid dimensions of RAs based on available HI capabilities and characteristics. Second, we apply such framework on a snapshot of recently proposed RAs to get insights on the current HI landscape. Third, we identify challenges and research directions aimed at the development of HI RAs.

Research Methodology

This section presents the methodology we applied to perform the literature review.

Core Terminology

First, we define the main concepts of our analytical framework: the research process with its phases, research assistance tools, and the field of Hybrid Intelligence.

Research Process Philosophy of Science has described multiple types of research process, often further refined by different definitions. We adapt the widely accepted *hypothetico-deductive* method (Sekaran and Bougie 2016) to reflect tasks that have been automated, and generalize it to be applicable to a larger number of fields. We thus consider the following adapted phases:

- Review literature: missing in the referred source;

- Develop hypothesis: includes the *Identify the problem area*, *Define a problem statement* and *Develop hypothesis* phases;
- Experiment: joining the *Determine measures* and *Data collection* phases;
- Interpret results: joining the *Data analysis* and *Interpretation of data*;
- Report results: missing in the referred source.

Research Assistants We refer to '(artificial) *research assistant*' as any type of automated agent that provides support to researchers during the research process. Considering the nuanced nature of the process, we include any RA that takes part in at least one of the aforementioned phases.

Hybrid Intelligence Hybrid Intelligence (Dellermann et al. 2019; Akata et al. 2020), also referred to as Hybrid Human-AI Intelligence, is an emerging research domain that investigates scenarios where human and artificial agents are both necessary and complementary. In such systems, collaboration between humans and AI is essential to achieve performance levels that surpass the simple sum of their individual capabilities. This is grounded on the assumption that combining the different forms of *intelligence* can synergistically mitigate biases and weaknesses inherent to each type of agent. By definition, Hybrid Intelligence systems include at least one human and one artificial agent, along with mechanisms to facilitate their interaction and collaboration.

Notice that the term *hybrid* (AI) is often used with different meanings in the field of Artificial Intelligence research, mainly when referring to systems that combine symbolic and statistical methods. In this work, we only use *hybrid* to refer to human-AI systems, preferring the term *neuro-symbolic* for the latter definition.

Search and Selection Criteria

Search Criteria We performed a keyword search in the Scopus online library[1]. The formulated query (cf. Appendix) consists of two conjuncts: we intersect keywords related to AI or autonomous systems with results filtering either specific phases of the research process (i.e. *academic writing*, *hypothesis generation*, *hypothesis evaluation*, ...) or keyword focusing on automated assistants (*research assistants*, *self-driving labs*, *computer-aided discovery*, ...).

Notice that, for this work, the *interpret results* phase is omitted. This is because the phase is heavily related to data processing and analysis activities, which comprise an extremely large proportion of AI-related literature. In order to keep the query results tractable, and considering the often 'passive' nature of such methods, we decided to leave the analysis of this phase for future work.

Selection Criteria Results from the keyword search are further limited according to the following criteria:

- Timeframe: we include papers dating from 2010 to 2024. By applying a date cutoff, we are aware that we are excluding classical research support tools (examples of which are the foundational Dendral (Lindsay et al. 1993) and Bacon (Langley 1978)). However, we aim to focus on tools that are most likely to represent the current landscape, and assume that more recent systems expand on earlier work on the topic.
- We only select peer-reviewed articles which propose actual models, or extend existing ones: we thus exclude reviews, blue-sky, and arxiv-only papers. This filter, and the previous one, is directly included in the Scopus query.
- Clear scientific tasks: the selected tools must actively perform tasks that are intended or optimized to be part of the research process. As an example, any given tool that recommends relevant papers for a literature review will be included, but application of existing summarization methods in the scientific domain will not.

We initially refined the results by filtering the title and abstract and performed a further selection from full-text reading. The final selection includes 47 papers, cf. Table 6, Appendix .

Analytical Framework

In this, we present the analytical framework we developed by adapting the following metrics from HI literature:

CARE (Akata et al. 2020) presents four desired capabilities (and related sub-capabilities) for Hybrid Intelligence systems under the acronym CARE: Collaborative, Adaptive, Responsible, Explainable. These include:

- **Collaborativeness**
  - Initiating relationships
  - Establishing shared situational awareness
  - Personalized multi-modal user interaction
  - Collaborative group support
- **Adaptivity**
  - Learning through interaction
  - Learning how to interact
  - Incremental adaptivity
  - Integrate symbolic constraints
- **Responsibility**
  - Critically examining decisions of big-data applications
  - Validating whether legally or morally acceptable behavior is learned
  - Reasoning about the legal or ethical acceptability of behavior

  In our analysis, we slightly expand *responsibility* to include the possibility of defining scientific constraints beyond the legal and ethical ones. Although the scientific process is clearly bound by legal and ethical considerations, an analysis of performances according to them would be trivial or extremely complex. Given that *responsibility* in research is also represented by the rigor and feasibility of hypotheses and methods, we therefore extended the definition of *responsibility* to include methodological and domain constraints.

[1] https://www.scopus.com

- **Explainability**
  - Transferability of shared representations
  - Quality of the explanations
  - Interactive explanations

In principle, these features are further refined in discrete levels (see Appendix), but as one of the aim of our analysis is to understand which capabilities are under-explored, we keep an inclusive approach: a system with features fitting any capability level will be marked positively. As a consequence, positive marking should be read as the presented system having *at least some features for the capability X*.

HI Teams (Dell'Anna et al. 2024) proposes a qualitative analysis of HI teams from multiple perspectives. While the team is an essential component of an HI system, the analyzed RAs were not necessarily developed following the HI framework, and might not fully fit the team dimension. As a consequence, we chose a less granular approach and include in our analysis three of the proposed team qualities:

- Initiative: we consider the initiative characteristics of the artificial actors within the team interaction. Specifically, the initiative can be classified as *passive*, i.e. the agent(s) has no initiative nor interaction with the user(s) besides being launched or stopped, *reactive*, i.e. there is interaction started from the user that can change the behavior of the agent during activity, and *mixed* (Allen, Guinn, and Horvtz 1999) where the agent can actively interact and change behavior according to the user's, without the need of being prompted. *Passive* actors can be, by our definition, also fully autonomous actors, as the discriminating characteristic is the type of interaction they have with the human agent, if an human agent is even needed.
- Interdependence: (Johnson et al. 2014) distinguishes interdependence as *hard* and *soft*, where the former describes complementary relationships that are required to manage dependencies in a joint task, and the latter describes relationships where dependencies come from the possibility of improving results. By this definition, the majority of autonomous research assistants should fall into the second category, given that most scientific tasks are, in principle, solvable by the sole human effort. This was also reported in (Dell'Anna et al. 2024): while it is considered a well-understood and important feature, experts also specified that the dependency in these systems is usually directed from the agent to the human. Inspired from this, we thus focus on the mutual dependency between the human and the agent, considering whether the parties can communicate (low interdependence), coordinate (medium), and overcome limitations (high), where the latter is intended from the agent's perspective.
- Effectiveness - Member satisfaction: given the variety of approaches and tasks, generic performance and satisfaction could not easily be compared. However, given the importance of the human factor within the HI team, we annotate the number of RAs systems in which user satisfaction or evaluation is directly considered in the analysis, hoping this could work as a proxy for the intention of fostering collaboration between the human and the agent.

General Characteristics We finally extend the HI analytical framework by annotating various additional features of the tools, with the intention of getting a general idea of the tasks and solutions within the analyzed landscape.

- Field of research: we categorize the analyzed tools into broad scientific fields, following the high-level clustering in the OECD Frascati Manual[2]:
  - Natural Sciences
  - Engineering and Technology
  - Medical and Health Sciences
  - Agricultural and Veterinary Sciences
  - Social Sciences
  - Humanities and the Arts
- Phase of the scientific process: for each system, we annotate the related phases in which it is employed, as described in the Methodology Section, and excluding the *interpret results* phase;
- Type of AI solution: HI literature (Tiddi et al. 2023) considers the neuro-symbolic approach as a promising way to overcome the classical downsides of data-driven and symbolic systems, which would limit certain HI capabilities. We therefore annotate the type of model implemented by the analyzed tools as *symbolic*, *data-driven* or *neuro-symbolic*. Referring to (van Bekkum et al. 2021), we define them, respectively, as only including a *semantic* model; only including a *statistical* model; including at least a combination of the two.

Research Synthesis

We proceed to present the results of the analysis of the selected work, according to the analytical framework presented in Section .

CARE Capabilities The following figures illustrate the distribution of CARE capabilities within the analyzed systems (Figure 1) and the research phases (Figure 2). *Adaptivity* is the most present, followed by *collaborativeness* and *responsibility*, while *explainability* is present only in nine papers. About one third (16) of the analyzed works do not match any capability, and there is no occurrence of all CARE capabilities being present at the same time.

Figure 2 shows the distribution of the capabilities between the different research phases. Here, it is possible to notice how only *adaptivity* is present in all phases. In contrast, *responsibility* features are mainly observed in the hypothesis development and experimentation phases. The absence of *responsibility* features in the report results and review literature phases is likely due to phase-specific requirements and the chosen annotation procedure (see Section 2). However, other missing capability-phase combinations pose challenges for future HI systems to address.

[2](Organisation for Economic Co-operation and Development 2015)

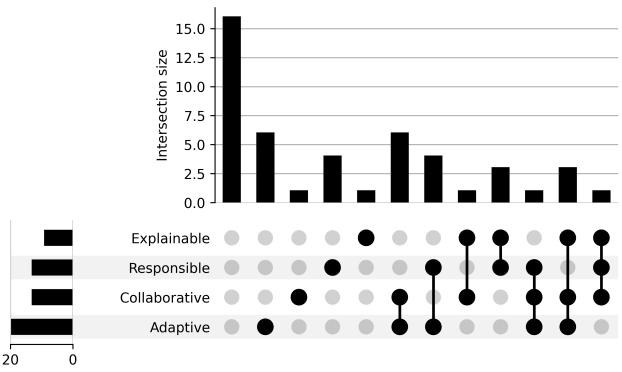

Figure 1: Number of papers per capability combination. Combinations that were not found are not shown.

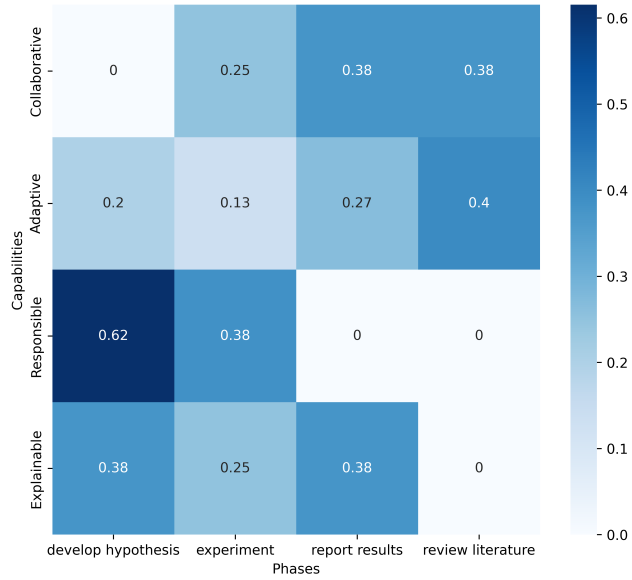

Figure 2: Capability distributions across research phases, normalized by capability.

Initiative Figure 3 and Table 1 shows how the types of initiative are distributed between the papers and phases. *Passive* tools are the most common and are mostly represented in the *develop hypothesis* phase. *Reactive* tools are evenly distributed, while only three mixed initiative systems were found, mainly in the *report results* and *review literature* phases. Notably, systems that tackle multiple parts of the scientific process seem to have a larger representation of reactive systems, opposite to the trend in phase-targeted tools.

Interdependence In Table 2, the number of interdependence types is shown. Figure 4 reports the correlation of interdependence types with the type of Initiative. Unsurprisingly, usually *reactive* initiative is needed to allow communication or coordination, and as a consequence higher levels of interdependence can mainly be found in such systems.

Initiative type	Nr. of papers
passive	31
reactive	13
mixed	3

Table 1: Number of papers per Initiative type.

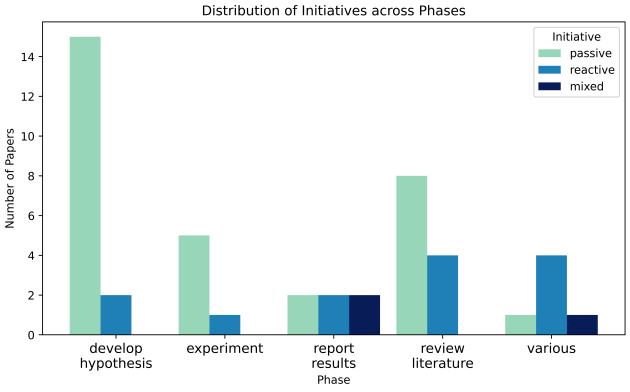

Figure 3: Initiative type distribution among research phases.

Only two RAs (20,31[3]) are able to integrate direct user feedback in order to improve performance.

Effectiveness - Member Satisfaction Among the 47 analyzed articles, 15 include some form of human evaluation of the results. As the majority of the systems fall under the data-driven category, it is not surprising that the main approach for evaluation is skewed toward benchmarking. However, considering the assistive nature of these tools, it is noteworthy that only about one third refer to user experiences to compare performance.

General Characteristics

Tables 4 and 5 show the number of systems related to research phase and research areas.

The *develop hypothesis* and *review literature* include most of the works, but all considered research phases are represented, and 6 works propose a more general approach, with their system covering at least two phases.

When considering fields of research, it is clear that the majority of systems are within STEM fields; this is likely due to the specific definition of the scientific process used for the search query. Nevertheless, a large number of field-independent systems is noticeable; while it is probably due to recommender tools for reviewing literature, it also highlights the need to have more comprehensive systems.

Table 3 shows the distribution of the AI solution types between articles. Purely data-driven models are in the majority, which can explain some of the aforementioned results (e.g. higher adaptivity and lower explainability). Despite the recent advances in neuro-symbolic research, these models account for a minority of the systems. Considering that there

[3]This numbering refers to Appendix , where complete references can be found.

Degree of interd.	Nr. of papers
None	32
Communicate	9
Coordinate	4
Overcome	2

Table 2: Number of papers per Interdependence type.

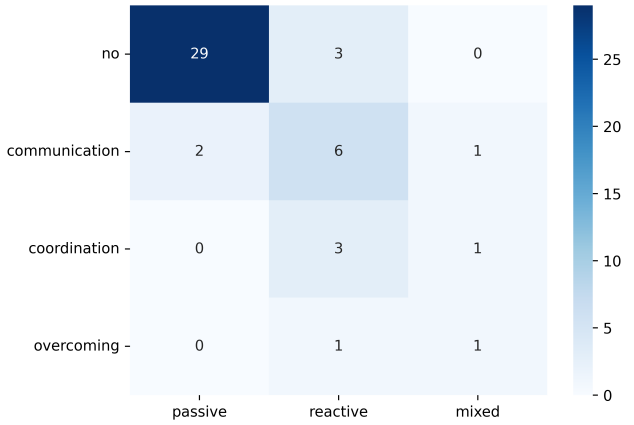

Figure 4: Distribution of interdependence characteristics across initiative types.

Model type	Nr. of papers
Symbolic	7
Data-driven	34
Neuro-symbolic	6

Table 3: Number of model types.

Research process phase	Nr. of papers
Review Literature	12
Develop Hypothesis	17
Experiment	6
Report Results	6
Various	6

Table 4: Number of papers per phase of the research process.

are many attempts to encode scientific knowledge symbolically, we suggest that this could be a noteworthy use-case for the application of neuro-symbolic systems.

Discussion

We further discuss the results of the analysis and draw some conclusions on their implications for the development of HI RAs. Furthermore, we reflect on how the framework can be extended to be better applied to the studied use-cases.

HI Features of the Systems In our analysis, all CARE capabilities are somewhat represented, but we found no tools including all of them. While this might be partly due to the specific qualitative annotation based on the HI tables (Appendix), it still poses the question of how it would be possible to combine them.

Regarding the specific qualities, *adaptivity* is the most commonly present, often appearing together with *collaborativeness* (see Figure 1): we suggest that this reflects the assistive nature of the analyzed tools. *Explainability* is on the other hand less present: while we expected that this capability would reflect the rigor of the scientific process, this result can also be understood considering the frequency of purely statistical models. We suggest that a fundamental cornerstone for the future development of the HI field is to pose attention to the actual applicability for the user and the requirements of the scientific domain.

Based on the HI capabilities found in the analyzed systems, we can divide them in systems matching 0, 1-2 or 3+ capabilities. This distinction, which we name "HI-ness", can be seen as adopting a relatively inclusive approach. We do not intend to claim that *hybrid* systems are close to the functionalities intended by the HI vision: as previously mentioned, each CARE capability also comes with defined levels, which were not taken into account in the analysis; nonetheless, with HI-ness, we are aiming to identify the minimal level of alignment of the tools to the design considerations suggested by HI. By the definition of HI-ness, most (26) of the selected articles are in the *low-hybrid* category, while only 5 systems (1, 18, 20, 35, 44) are *hybrid*: in this regard, we remark that all these share the *collaborativeness* capability. It is worth noting that, except for (1), all the *hybrid* systems include a LLM in their pipeline, which might hint at a promising role of the technology also in the RA domain, provided that relevant downsides in the *responsible* and *explainable* capabilities, such as hallucinations (Ji et al. 2023) and robustness of reasoning (Lappin 2024) are properly addressed. We also highlight how a large number of systems (16) fall under the *non-hybrid* category. This can find justification in the fact that some systems are either developed with the aim of being used in substitution of humans (16), or they target a specific task from a benchmark-based perspective (4, 12, 19). While these tasks primarily emphasize performance and advancing the state-of-the-art, we propose that their assistive potential may be consequently diminished. Incorporating CARE principles into these systems could open up promising new research directions for their usability by (or with) human researchers.

Our analysis of the type of initiative in the agents (Figure 3) showed an abundance of passive agents, but the majority is concentrated in tools tackling the *develop hypothesis* and *review literature* phase. We suggest that lower interaction in these phases can find an explanation in the nature of the tasks and their evaluation metrics, as the usual benchmark consists in predicting already known hypotheses or citations in test sets (e.g. 2, 4, 12, 13). This is clearly different from the 'real-life' methods of hypothesis generation and refinement (both for research hypothesis and relevant literature), and we suggest that HI systems should shift to more 'interactive' models, where user interaction and feedback could guide the discovery process.

In this regard, comparing the HI-ness of the systems with respect to the types of initiative of the agent (Figure 5),

Research field	Nr. of papers
Natural Sciences	15
Engineering and Technology	8
Medical and Health Sciences	3
Agricultural and Veterinary Sciences	0
Social Sciences	3
Humanities and the Arts	0
General	18

Table 5: Number of papers per research field.

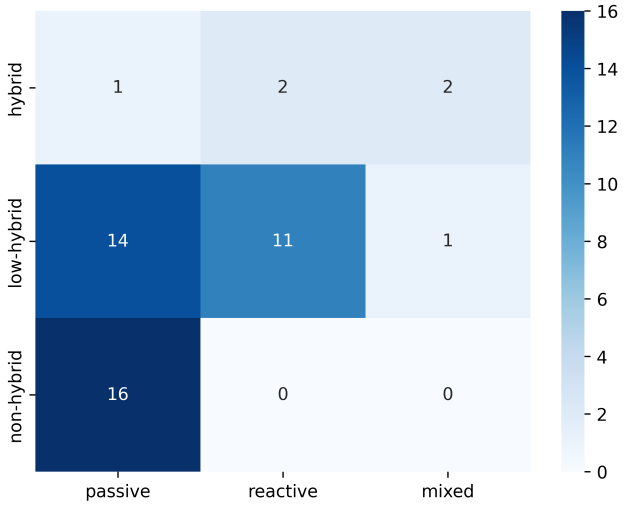

Figure 5: HI-ness of systems by initiative type.

shows how increasing the initiative in the interaction is usually linked to an increase in CARE capabilities. This is not surprising and probably influenced by how the *collaborativeness* and *adaptivity* capabilities are strongly intertwined with an agent's initiative, but we posit that intentionally considering reactive types of initiative in development of new tools would be a way to increase HI abilities 'by design'.

By the provided definition, interdependence is heavily linked and dependent on the initiative/interaction between the members of the team, so it is not surprising that passive agents do not present any interdependent features. Many passive models could even be defined as *independent*. We maintain that interdependent qualities can still provide different information with respect to to initiative, and provide further insights on the interaction between team members.

To complete the overview, we highlight the importance of user-evaluated effectiveness. We have highlighted how the agents we have tagged *hybrid* share the *collaborative* capability, making it imperative that users are involved, and considered in the development of HI systems.

Applicability of HI Characteristics We conclude with a few considerations on the HI characteristics of our framework. As current efforts to shape the field of HI are still in their infancy (with foundational papers like (Dellermann et al. 2019) and (Akata et al. 2020) being relatively recent), most of the available characteristics are intended to be generic. This in turn makes it harder to provide a rigorous analysis when considering use-cases. In this work, we adapted some of the capabilities and team features accordingly, but we had to accept the trade-off on annotation specificity. Thus, we suggest the need for guidelines for use-case-specific evaluations, especially when considering the *responsibility* and *explainability* capabilities. The former, currently focused on ethical and legal constraints, does not easily mirror the requirements of certain tasks in the scientific process: while scientific integrity is of paramount importance and ethical considerations can be made when generating hypotheses, defining ethical or legal constraints in phases like literature recommendation or automated experimentation is an unclear task (or trivialized by dataset or hardware requirements).

In the *explainability* case, the original definition (Appendix) already specifies that *quality of explanations* should be linked to the use-case. In this work, we focused on the intent of the authors in providing explanations, or in the nature of the system including shared representations by definition (i.e. symbolic systems), which in turn might have contributed to the lower number of systems annotated as *explainable*. Given the lack of requirements, and at the same time the large corpus of research in explainable AI, we further suggest the need for the involvement of users in understanding which types of explainable properties would be appropriate for the specific tasks and phases.

The HI team-based characteristics, as previously noted, are often intertwined with CARE capabilities. In the currently available definition, they seem to provide a lower level of detail compared to CARE, but we suggest that they can represent efficient guidelines for developing tools that already include at least minimal HI features. Despite this suggestion, the annotation of *interdependence* characteristics is often linked to certain *initiative*, *collaborativeness* and *adaptivity* features. We posit that this might be due to the (usual) lack of embodiment of the analyzed RAs, which simplifies *interdependence* at all the described levels, and, while still informative, might need further refinement to provide separate insights.

Conclusions and Future Work

In this work, we have reviewed a range of AI-based tools designed to support researchers in the scientific process, exploring how principles from Hybrid Intelligence can inform the development of future tools that interact synergistically with researchers. We proposed an analytical framework, derived from relevant HI literature, which includes applicable capabilities and characteristics adapted for the scientific discovery use case. We have reported a lack of HI-specific capabilities in many systems, and how interaction with the users is often very limited. Concurrently, we have suggested that designing for higher initiative and interaction is usually linked to increased alignment to HI principles. Finally, we have highlighted the need for more specific metrics or guidelines, further tailored to the use case, which should take into account the specific context and requirements of the tasks.

Despite our attempt to define a comprehensive query, we know that a number of possibly relevant articles were not

	Paper	Approach	Phase	Area	C	A	R	E	interd.	h. eval.	init.
1	(Doud and Yilmaz 2017)	ne-sy	experiment	natural sci	x		x	x	no	no	p
2	(Subramanian et al. 2020)	data-driven	develop hypotheses	med and health					no	no	p
3	(Zhang et al. 2018)	data-driven	report results	various		x			no	no	p
4	(Mandave and Pole 2017)	data-driven	literature review	various					no	no	p
5	(Chen and Ban 2019)	data-driven	literature review	various		x			no	no	p
6	(Ramirez et al. 2023)	data-driven	experimentation	natural sci		x	x		no	yes	p
7	(Safder, Hassan, and Aljohani 2018)	data-driven	literature review	eng and tech	x	x			cmm	no	r
8	(Behandish, Maxwell, and de Kleer 2022)	ne-sy	develop hypotheses	natural sci			x	x	no	no	p
9	(Sorkun et al. 2020)	data-driven	develop hypotheses	eng and tech			x		no	no	p
10	(Gao and Cheng 2015)	symbolic	develop hypotheses	various		x	x		cmm	no	r
11	(de Campos, Fernandez-Luna, and Huete 2024)	data-driven	report results	various	x			x	cmm	yes	r
12	(Choudhary and Connolly 2021)	data-driven	develop hypotheses	med and health					no	no	p
13	(Larson and Van Cleemput 2017)	symbolic	develop hypotheses	natural sci					no	no	p
14	(Hakuk and Reich 2020)	symbolic	develop hypotheses	natural sci				x	no	no	p
15	(Skirzynski, Jain, and Lieder 2024)	ne-sy	develop hypotheses	social sciences		x	x		cmm	yes	p
16	(Ament et al. 2021)	data-driven	experimentation	natural sci					no	no	p
17	(Reder et al. 2024)	symbolic	experimentation	natural sci			x		no	no	p
18	(Lim et al. 2024)	data-driven	report results	eng and tech	x	x		x	ovc	yes	m
19	(Sharma, Gopalani, and Meena 2017)	data-driven	literature review	various					no	no	p
20	(Shen et al. 2023)	data-driven	report results	eng and tech	x	x		x	crd	yes	m
21	(Majumder et al. 2024)	data-driven	varioys	various	x	x			cmm	no	r
22	(de Haan, Tiddi, and Beek 2021)	data-driven	develop hypotheses	social sci					no	yes	p
23	(Wang et al. 2022)	data-driven	various	varioys					no	no	p
24	(Rubio and Gulo 2016)	data-driven	literature review	eng and tech					no	no	p
25	(Abgaz et al. 2016)	ne-sy	develop hypotheses	eng and tech					no	yes	p
26	(Nguyen, Le, and Nguyen 2022)	data-driven	literature review	engi and tech	x				no	no	r
27	(Choi 2018)	data-driven	develop hypotheses	natural sci					no	no	p
28	(Kely De Melo et al. 2022)	data-driven	literature review	various					no	no	p
29	(Mucke et al. 2023)	data-driven	report results	various		x			no	no	r
30	(Venkatesan et al. 2023)	data-driven	literature review	various		x			no	no	p
31	(Goel and Joyner 2015)	data-driven	develop hypotheses	natural sci		x	x		crd	yes	r
32	(Karunananda et al. 2021)	data-driven	various	various	x	x			cmm	yes	r
33	(Anil et al. 2024)	data-driven	literature review	various		x			no	yes	r
34	(Li et al. 2024)	data-driven	various	various	x	x			cmm	yes	m
35	(Yoshikawa et al. 2023)	ne-sy	experimentation	natural sci	x	x	x		crd	no	r
36	(Gower et al. 2023)	symbolic	develop hypotheses	natural sci			x	x	no	no	p
37	(Segler and Waller 2017)	ne-sy	develop hypotheses	natural sci			x		no	no	p
38	(Alzoghbi et al. 2015)	data-driven	literature review	various		x			no	no	p
39	(Lemos et al. 2023)	data-driven	develop hypotheses	natural sci					no	no	p
40	(Ng 2020)	data-driven	develop hypotheses	natural sci					no	no	p
41	(Pollak et al. 2015)	data-driven	report results	various					cmm	yes	p
42	(Al-Natsheh et al. 2017)	data-driven	literature review	social sciences	x	x			ovc	yes	r
43	(Choe et al. 2024)	data-driven	various	various	x	x			crd	yes	r
44	(Xu, Ye, and Zhu 2023)	data-driven	various	natural sci	x	x		x	cmm	yes	r
45	(Garijo et al. 2019)	symbolic	experimentation	natural sci			x	x	no	no	p
46	(Liu, Goulding, and Brailsford 2015)	data-driven	develop hypotheses	various					no	no	p
47	(Wagner et al. 2024)	symbolic	develop hypotheses	med and health			x		no	yes	p

Table 6: Analyzed papers and annotated features. interd: *interdependence*; h. eval.: *human evaluation*; init.: *initiative*

captured: the *analyze results* phase is clearly one of the phases where automated tools have been used the most and while excluding it from the scope of the survey was required to maintain a tractable number of papers, we are aware its analysis could provide valuable insights.

In further extensions of this work, the annotation of HI systems with CARE capabilities could be done keeping into account the various levels of each capability for better granularity: specifically, this would allow further investigation of whether different types of model can consistently achieve better capabilities. Nevertheless, such an approach would further highlight the need to targeted adaptation of CARE capabilities levels to the use case considered.

Finally, an additional research direction would involve an extended exploration of the capabilities of LLM-based tools. In our analysis, we already encounter various such examples (18,20,21,28,29,35,43,44), 4 of which are included in the *hybrid* category. The capability for communication and user interaction represent an exciting feature for hybrid assistants, but warrants further exploration, particularly on the role of *responsible* and *explainable* aspects. Additionally, we are aware that a number of commercial LLM-based solutions have been developed, which usually require a certain level of trustworthiness to be deployed. Although they currently do not meet the inclusion criteria, they represent an interesting opportunity to extend and compare our results.

Scopus Query

KEY (
(“artificial intelligence” OR “AI” OR “automation” OR “automated systems” or “intelligent systems”)
AND (
((“idea generation” OR “research question generation” OR “literature recommendation” OR “literature discovery” OR “related works” OR “hypothesis generation” OR “hypothesis proposal” OR “theory generation” OR “theory proposal” OR “hypothesis testing” OR “theory testing” OR “experimentation” OR “hypothesis evaluation” OR “theory evaluation” or “report writing” or “paper writing”)
AND
(“scientific domain” OR “scholarly domain” OR “research domain” OR “academic domain” OR “scientific process” OR “scientific” or “academic” or “research method” or “scientific method” or “academic method”)
) OR
(“scientific writing” or “academic writing” or “scholarly writing” or “research writing” or “scientific text” or “academic text” or “scholarly text” or “research text” or “scientific publication” or “academic publication” or “scholarly publication” or “research publication” or “scientific paper” or “academic paper” or “scholarly paper” or “research paper” or “scientific discovery” or “academic discovery” or “scholarly discovery” or “research discovery” or “scientific contribution” or “academic contribution”)
OR
(“scientific assistant” OR “scholarly assistant” OR “academic assistant” or “research assistant” or “computer-aided discovery” OR “Self-driving labs” or “automated research”)
))
AND PUBYEAR $>$ 2014 AND PUBYEAR $<$ 2025 AND (LIMIT-TO (DOCTYPE , “ar”) OR LIMIT-TO (DOCTYPE , “cp”))

CARE Tables

(Hybrid Intelligence Centre Netherlands 2023) defines the four CARE capabilities, further refined in sub-capabilities and level. Due to formatting constraints, the full tables can also be found at https://github.com/Zamprognog/survey_autonomous_ra_hi.git.

Analyzed Papers

Table 6 presents a complete overview of the analyzed works, including the annotations from the described framework. Due to the number of references and formatting requirements, the full list of references is available at https://github.com/Zamprognog/survey_autonomous_ra_hi.git.

Acknowledgements

This work is supported by the Hybrid Intelligence programme (https://www.hybrid-intelligence-centre.nl/), funded by a 10-year Zwaartekracht grant from the Dutch Ministry of Education, Culture and Science (NWO).

References

Akata, Z.; Balliet, D.; De Rijke, M.; Dignum, F.; Dignum, V.; Eiben, G.; Fokkens, A.; Grossi, D.; Hindriks, K.; Hoos, H.; Hung, H.; Jonker, C.; Monz, C.; Neerincx, M.; Oliehoek, F.; Prakken, H.; Schlobach, S.; Van Der Gaag, L.; Van Harmelen, F.; Van Hoof, H.; Van Riemsdijk, B.; Van Wynsberghe, A.; Verbrugge, R.; Verheij, B.; Vossen, P.; and Welling, M. 2020. A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *Computer*, 53(8): 18–28.

Allen, J.; Guinn, C.; and Horvtz, E. 1999. Mixed-initiative interaction. *IEEE Intelligent Systems and their Applications*, 14(5): 14–23.

Dell’Anna, D.; Murukannaiah, P. K.; Dudzik, B.; Grossi, D.; Jonker, C. M.; Oertel, C.; and Yolum, P. 2024. Toward a Quality Model for Hybrid Intelligence Teams. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, AAMAS ’24, 434–443. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9798400704864.

Dellermann, D.; Ebel, P.; Söllner, M.; and Leimeister, J. M. 2019. Hybrid Intelligence. *Business & Information Systems Engineering*, 61(5): 637–643.

Hogan, A.; Blomqvist, E.; Cochez, M.; D’amato, C.; Melo, G. D.; Gutierrez, C.; Kirrane, S.; Gayo, J. E. L.; Navigli, R.; Neumaier, S.; Ngomo, A.-C. N.; Polleres, A.; Rashid, S. M.; Rula, A.; Schmelzeisen, L.; Sequeda, J.; Staab, S.; and Zimmermann, A. 2021. Knowledge Graphs. *ACM Comput. Surv.*, 54(4).

Hybrid Intelligence Centre Netherlands. 2023. Strategy Plan. Accessed: 2025-01-08.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, 55(12).

Johnson, M.; Bradshaw, J. M.; Feltovich, P. J.; Jonker, C. M.; van Riemsdijk, M. B.; and Sierhuis, M. 2014. Coactive design: designing support for interdependence in joint activity. *J. Hum.-Robot Interact.*, 3(1): 43–69.

Langley, P. 1978. Bacon. 1: A general discovery system. In *Proc. 2nd Biennial Conf. of the Canadian Society for Computational Studies of Intelligence*, 173–180.

Langley, P. 2000. The computational support of scientific discovery. *International Journal of Human-Computer Studies*, 53(3): 393–410.

Lappin, S. 2024. Assessing the strengths and weaknesses of Large Language Models. *Journal of Logic, Language and Information*, 33(1): 9–20.

Lindsay, R. K.; Buchanan, B. G.; Feigenbaum, E. A.; and Lederberg, J. 1993. DENDRAL: A case study of the first expert system for scientific hypothesis formation. *Artificial Intelligence*, 61(2): 209–261.

Meyer, J. G.; Urbanowicz, R. J.; Martin, P. C. N.; O'Connor, K.; Li, R.; Peng, P.-C.; Bright, T. J.; Tatonetti, N.; Won, K. J.; Gonzalez-Hernandez, G.; and Moore, J. H. 2023. ChatGPT and large language models in academia: opportunities and challenges. *BioData Mining*, 16(1): 20.

National Science Board, National Science Foundation. 2023. Publications Output: U.S. Trends and International Comparisons. Technical Report NSB-2023-33, National Science Foundation, Alexandria, VA.

Organisation for Economic Co-operation and Development. 2015. *Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development*. Paris: OECD Publishing. ISBN 978-92-64-23921-0.

Sarewitz, D. 2016. The pressure to publish pushes down quality. *Nature*, 533(7602): 147–147.

Sekaran, U.; and Bougie, R. 2016. *Research methods for business: A skill building approach*. john wiley & sons.

Tiddi, I.; De Boer, V.; Schlobach, S.; and Meyer-Vitali, A. 2023. Knowledge Engineering for Hybrid Intelligence. In *Proceedings of the 12th Knowledge Capture Conference 2023*, K-CAP '23, 75–82. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701412.

van Bekkum, M.; de Boer, M.; van Harmelen, F.; Meyer-Vitali, A.; and Teije, A. t. 2021. Modular design patterns for hybrid learning and reasoning systems. *Applied Intelligence*, 51(9): 6528–6546.

Walker, M. G. 1987. How Feasible Is Automated Discovery? *IEEE Intelligent Systems*, 2(01): 69–82.

Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; 't Hoen, P. A.; Hooft, R.; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. J.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Rocca-Serra, P.; Roos, M.; van Schaik, R.; Sansone, S.-A.; Schultes, E.; Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; van der Lei, J.; van Mulligen, E.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J.; and Mons, B. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1): 160018.

Wu, Z.; Ji, D.; Yu, K.; Zeng, X.; Wu, D.; and Shidujaman, M. 2021. AI Creativity and the Human-AI Co-creation Model. In Kurosu, M., ed., *Human-Computer Interaction. Theory, Methods and Tools*, 171–190. Cham: Springer International Publishing. ISBN 978-3-030-78462-1.