Complex Problem Solving through Human-AI Collaboration: Literature Review on Research Contexts

Lucas Memmert
University of Hamburg
lucas.memmert@uni-hamburg.de

Eva A. C. Bittner
University of Hamburg
eva.bittner@uni-hamburg.de

Abstract

Solving complex problems has been proclaimed as one major challenge for hybrid teams of humans and artificial intelligence (AI) systems. Human-AI collaboration brings immense opportunities in these complex tasks, in which humans struggle, but full automation is also impossible. Understanding and designing human-AI collaboration for complex problem solving is a wicked and multifaceted research problem itself. We contribute to this emergent field by reviewing to what extent existing research on instantiated human-AI collaboration already addresses this challenge. After clarifying the two key concepts (complex problem solving and human-AI collaboration), we perform a systematic literature review. We extract research contexts and assess them considering different complexity features. We thereby provide an overview of existing and guidance for designing new, suitable research contexts for studying complex problem solving through human-AI collaboration and present an outlook for further work on this research challenge.

1. Introduction

Solving complex problems has been proclaimed as one major challenge for hybrid teams of humans and artificial intelligent (AI) systems [1]. The nature of work changes towards higher knowledge intensity and complexity, often exceeding the abilities of individual humans [2]. In parallel, AI, in particular machine learning, has improved over the last years even outperforming humans in some areas. However, those AI systems are limited to specific, narrowly defined tasks, far from reaching a general level of intelligence [3].

[4] describe a combination of human and AI – named "hybrid intelligence" by different authors [4, 5] – as a likely paradigm for the next years. The underlying reasoning is that the strengths and weaknesses of humans and AI systems are complementary, resulting in a mutually beneficial relationship when combined, enabling the team to jointly achieve more than individually possible [4]. According to [6] for AI-enabled machines to be "effective teammates" they need to "engage in at least some of the steps in a complex problem solving process". However, how humans and AI systems collaborate to solve problems is still unclear [7].

Though the challenge of complex problem solving (CPS) through human-AI collaboration was posed comparatively recently [1], there is already an extensive body of knowledge available around the broader topic of human-AI collaboration. However, it is unclear, to what extent this research deals with CPS.

Addressing this open point, we review extant literature on instantiated human-AI collaboration and thereby answer the following research questions (RQ):
RQ 1: In which instantiated research contexts is research on human-AI collaboration conducted?
RQ 2: To what extent can these research contexts be considered to deal with complex problem solving?

[6] advocate for taking a socio-technical perspective on human-AI collaboration. Figure 1 illustrates a schematic view of socio-technical systems, consisting of *user(s)* interacting with *technology* to work on a *task* within a specific *context*. To what extent a study can be considered to deal with CPS depends on certain characteristics of the *task* that the human-AI team performs and the *context* this task is performed in. Therefore, the combination of *task* and *context* will be the focus of this study. It will be referred to as *research contexts* for the remainder of this paper (some authors use *scenario* as a term, e.g., *[8, 9]*). For the analysis, the research contexts will be extracted from the relevant publications, described briefly (RQ 1) and assessed against literature-based characteristics for CPS (RQ 2).

For researchers wanting to study CPS through human-AI collaboration it is crucial to select an appropriate research context, as the choice of the research context will define the class of the problem the derived insights can be generalized to. This paper provides researchers with an overview of potential research contexts and relevant underlying characteristics.

URI: https://hdl.handle.net/10125/79376
978-0-9981331-5-7

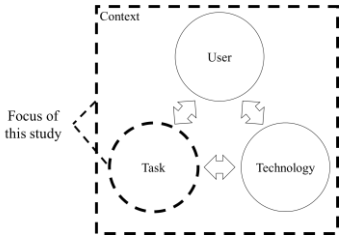

Figure 1. Schematic depiction of a socio-technical system (according to [10]); focus added

The analysis is limited to studies going beyond only providing theoretical descriptions of potential research contexts, focusing on studies with instantiated research contexts, i.e., actual implementations of research contexts within which user studies or experiments were conducted. This focus ensures that only research contexts are included which have already proven to be suitable for studying CPS through human-AI collaboration. They are therefore thought to be most helpful for scholars as a starting point, both for conducting studies within these already established research contexts and for inspiring the construction of new research contexts.

Additionally, this focus on instantiations was set in the light of the design science research tradition in IS research [11] and the call for design knowledge [1, 6], with instantiations playing a critical role in theory building (e.g., for developing cause-effect-theories). As [11] note, "[m]id-range theories or grand design theories" are usually not simply invited, but are built through abstraction from a number of specific instantiations.

The remainder of this paper is structured as follows: As part of the background section, "human-AI collaboration" and its adjacent concepts as well as "CPS" are explained. Within the methodology section the approach for reviewing the literature is described. The overview regarding research contexts and their complexities along with additional explanatory details is provided in the succeeding section. Subsequently, limitations are described and the paper closes with a conclusion and an outlook on relevant future research.

2. Background

Both concepts relevant in the context of this paper (human-AI collaboration, CPS) have a long research history outside Information System (IS) research and are described in detail in the following section.

2.1. (Human-AI) collaboration

Human-AI collaboration is the idea of humans and AI systems working together. The underlying reasoning is that humans and AI systems are complementary in their strengths and weaknesses [1, 5]. AI systems can perform clearly defined tasks, process large amounts of data, recognize patterns and make coherent predictions [1]. Already today AI systems can outperform humans who can only process a limited amount of information and are subject to cognitive bias [5]. Humans, on the other hand, are creative, empathic and have common sense [1]. They comparatively easily navigate dynamic environments, put unexpected events into context and react flexibly to changing circumstances [1, 5]. They can deal with fragmented information and solve abstract problems [10]. [1] provide an in-depth discussion and conceptualization of "hybrid intelligence".

While there are already many contributions around this topic, there is some uncertainty regarding the terminology. Several papers use the term collaboration without a shared definition [e.g., 4, 5]. Additionally, the idea of combining humans and AI systems is researched under a variety of labels, such as human-AI collaboration, human-AI teaming or hybrid intelligence. However, the concepts are not clearly distinguished, which might hinder the progress of research. This terminological challenge is not only faced by IS research but exists cross-disciplinarily [12]. [12] point out that "without shared understanding of the collaboration construct, it is difficult to theorize and empirically test […] outcomes associated with collaboration" and therefore state that "it is necessary to reduce construct confusion about collaboration to systematically and collectively advance knowledge".

[12] provide a comprehensive, multi-discipline literature review identifying the relevant aspects of collaboration, going beyond the idea of *working together* and defining collaboration as an "evolving process whereby two or more social entities actively and reciprocally engage in joint activities aimed at achieving at least one shared goal" [12]. This widely cited review is used as a foundation for the understanding of "collaboration" in this study. The relevant characteristics of the definition are briefly explained according to [12]:

Evolving process. Collaboration is neither structure nor outcome, but a *process* leading to a (collaborative) outcome. This process consists of emergent states (e.g., motivations or values of team members) and collaborative behaviors (e.g., adaptation or sense making) and is influenced by contextual factors (e.g., task or environment). These influences feed back into the collaborative process making it dynamic or *evolving* in nature.

Two or more social entities. Collaboration or "working together" requires at least two entities such as individuals, teams or organizations.

Reciprocal. Collaboration requires mutual engagement of the involved entities, contributing to the collaborative effort, i.e., collaboration is neither one-sided nor simply a delegation of work, but is a "back-and-forth" *reciprocal* process. The requirement for a "back-and-forth" reciprocal process is interpreted in this paper as

follows: a simple way of working together with an AI system, such as the AI making a recommendation and a human taking the decision (i.e., decision support system) does not qualify as *collaboration*. The reciprocity characteristic here is in alignment with the understanding of hybrid intelligence where "activities conducted by each agent are conditionally dependent" [1].

Participation in joint activities. Collaboration focusses on performing joint activities such as solving problems, performing actions, or executing tasks.

Achieving a shared goal. Entities participating in shared work might have a variety of goals, which might be conflicting. However, what sets collaboration apart, is, that all entities have at least one shared or mutually agreed upon goal.

According to the analysis of [12], teaming and collaboration differ to the extent, that collaboration might occur on different levels (e.g., individuals, teams or organizations), whereas teaming usually applies only on team-level (see Figure 2). Given this relationship, both concepts will be considered in this paper as part of the literature search. Coordination and cooperation lack some of the before mentioned features and are therefore not included (though considered during initial search).

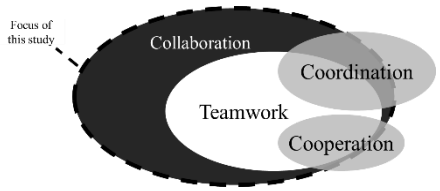

Figure 2. "Shared criterion space among collaboration and related constructs" according to [12]; focus added

A widely used means to analyze team effectiveness is the Input-Mediator-Output-Model (I-M-O-Model) [13]. According to [13], inputs can be considered "antecedent factors that enable and constrain members' interactions", mediators are processes and emergent states impacting outputs, and outputs are "results and by-products of team activity that are valued by one or more constituencies" [13]. While an in-depth analysis of team effectiveness is beyond the scope of this paper, we will use the model to point our relevant aspects of the studies with regards to CPS through human-AI collaboration.

2.2. Complex problem solving

Solving complex problems was proclaimed as the next challenge [1] for the collaboration of humans and AI systems. [6] ask for AI systems in order to be "effective teammates" to "*engage* in at least some of the steps in a complex problem solving *process* [emphasis added]" and [1] outline their understanding of complex problems as being "time variant, dynamic, requir[ing] much domain knowledge and hav[ing] no specific ground truth". They state that "these highly uncertain contexts require intuitive and analytic abilities, as well as human strengths such as creativity and empathy".

These descriptions provide a first impression of what could be studied under the label "CPS". In order to systematically review literature, however, a more detailed understanding of CPS rooted in theory is required.

CPS has a long research history in the psychology literature dating back to the mid-1970s [14]. It started with researching relatively simple problems in laboratory environments [9]. However, researchers realized that insights derived from those studies did not generalize well to more complex, real-life problems [9]. In reaction, European and North American scholars pursued different avenues, with European scholars focusing on increasingly complex, computerized laboratory tasks (so called "microworlds") whereas Northern American researchers focused on separate, natural knowledge domains (e.g., reading, writing, calculation) [9]. As a result, key terms such as "complexity" or "CPS" itself are not defined consistently in the field [8, 9].

Moreover, the transition from *simple* problem solving to *complex* problem solving is not clearly defined either [14]. [9] state, that these two concepts are qualitatively different. While *simple* problem solving requires to "overcome a single barrier; [*complex* problem solving], in contrast, deals with a large number of barriers that coexist simultaneously. Because there are multiple barriers, a single cognitive or behavioral activity may not be sufficient to reach the goal state". Problems solvable by *varying one thing at a time* (VOTAT) should not be considered complex problem solving [14].

Reflecting on this discourse, [8] provide a taxonomy of CPS tasks in order to enable an assessment of "the similarity between tasks and to generalize results from one type of task to another". The taxonomy contains both "formal" and "psychological" features. Using psychological features for the analysis would require "[postulating] processes and representations […] in participants facing the task" [8]. Given the novelty of the research challenge and the lack of understanding on both processes and representations of joint human-AI CPS in research, only "formal" features are used for the analysis.

The taxonomy consists of three feature-groups (time related, variable related, and system behavior related), each containing three features. Each feature is shortly explained below according to [8]:

Time related aspects are:

- *Time variance*. While static systems only change when participants intervene, dynamic or time variant systems can change independently of the participant, increasing complexity.

- *Time continuity*. Progression of time might be continuous, e.g., as in most real-life settings or might occur stepwise, e.g., as in games were participants take turns marking discrete "steps".
- *Degree of time pressure.* Time pressure refers to whether the research contexts require rapid decisions and actions, which increases complexity.

Variable related aspects are:

- *Number and type (discrete/continuous) of variables.* Provides a first indication for the size of the research context or problem, with more variables increasing complexity.
- *Number and pattern of relationships between variables.* Refers to the number of connections between variables. The more interconnected variables are, i.e., the more variables change as a consequence of a manipulation of a variable, the more complex. Certain patterns make the detection of causalities more difficult and therefore increase complexity.
- *Non-linearity*. Refers to the type of connection between variables. As humans do not perform well at estimating non-linear relations, those are considered to increase complexity.

System behavior related aspects are:

- *Opacity*. Refers to whether the complete state of the problem is visible or transparent to the participant. A system containing variables not directly observable (opaque) is considered more complex.
- *Stochasticity.* Refers to the changes within the system resulting from an action. If the same action is taken in the same state always results in the same succeeding state, the system is deterministic. If the resulting state might vary (stochastic), this increases complexity. Some opaque research contexts might appear to be stochastic to participants (because the participants do not see the "full picture"), however, during this literature review the research contexts are assessed with an outside view.
- *Delayed feedback*. Refers to the possibility of relating feedback to actions. If feedback (e.g., successful completion or increased score) cannot be related directly to an action, this increases complexity.

In CPS literature, typically a human is assessed regarding their problem solving competency and strategies. For this paper, we considered a team of at least one human and an AI system. Considering multiple (including non-human) stakeholders makes a clarification regarding *opacity* and *stochasticity* necessary.

Crucial for determining opacity for this study is the combined perception of the participant team, i.e., if the team members can jointly perceive the complete state of the research context, it is not considered to be opaque.

Additionally, stochasticity here is considered a problem characteristic. The fact that one of the team members (e.g., a human that is part of the human-AI team) might act stochastically, does not make the problem or research context stochastic. Simply speaking, a *problem* should not be considered *complex* because a human and AI work on it together but because the problem is inherently complex. Complexity introduced through working with AI systems is out-of-scope for this study as it was discussed already by [15].

This conceptualization of complexity is actor-independent, which is in alignment with CPS literature: "In the literature on CPS, it is mostly the structure of the external problem representation that is considered complex. So a problem usually is considered being of a certain complexity, even if it might seem less complex to problem solvers with more expertise" [16].

It should be noted, that none of the features are necessary or sufficient to qualify a research context as CPS [8]. However, they can be used to compare research contexts and help to understand, which findings might generalize across which research contexts [8].

3. Methodology

In order to ensure transparency and reproducibility, a systematic literature review following [17] is conducted. The search string used is *"hybrid intelligence" OR ("human" AND ("AI" OR "artificial intelligence") AND ("team*" OR "collaboration" OR "coordination" OR "cooperation"))*. The asterisk works as a wildcard for any number of characters providing flexibility. For databases without wildcard-support the terms *team*, *teams*, and *teaming* where used.

In alignment with [1] *hybrid intelligence* is used to find contributions describing combinations of humans and AI systems. However, as this is a relatively new term for describing human-AI collaboration, contributions containing *human* and *artificial intelligence* (or abbreviated) either in a *collaboration*- or *team*-setting are included additionally. These two concepts (*collaboration, team**) are chosen due to their similarity [12]. We presumed that some authors might use the concepts interchangeably and after initial screening we included the terms *coordination* and *cooperation* into our search.

The search string does not contain any reference to CPS as problems or tasks considered to qualify as CPS vary widely [8] and including specific aspects might lead to premature exclusion of relevant search results.

The search string is applied to title, abstracts and keywords. Where possible, the search is restricted to peer-reviewed contributions. Relevant publications for further analysis are identified in two review cycles.

1st review: By examining title, abstract and keywords, the following type of contributions are excluded: "hybrid intelligence" or "AI" have a different meaning than intended in the search string (e.g., "hybrid intelligence" refers to different types of AI algorithms being

used together), purely conceptual/ non-instantiated (e.g., theoretical frameworks, use case descriptions), focus on technical aspects of improving machine learning, non-collaborative (in alignment with the understanding outlined earlier), no usage of AI (e.g., suggestion of AI usage only as an outlook), and certain formats (e.g., workshops, talks).

2nd review: By examining the (user) study and research context descriptions within the contributions, only those providing sufficient information to assess the "CPS" characteristics outlined previously are included. This requires a specific description of the problem sought to be solved, not some (more generic) outlook as to which research contexts could potentially benefit from the research. It also requires an instantiation of the research context, e.g., accompanied by a user study in order to maintain the focus on suitable research contexts and provide a strong foundation for theorizing. Publications with duplicate research contexts (e.g., same context investigated in different publications) and primarily physical tasks (i.e., non-intellectual [1]) were excluded.

The searched domain-relevant databases, the number of results during the initial search, as well as after each of the review cycles are shown in Table 1. We performed a forward and backward search to identify additional relevant articles.

Table 1. Search results per review stage and database, and forward/backward (f/b) search

Database	Initial hits	After 1st review	After 2nd review + f/b
AISeL	19	1	-
ACM DL	201	53	16 + 3
ScienceDirect	265	23	2 + 1
Total	**485**	**77**	**18 + 4**

4. Results

A total of 22 relevant publications were identified. They were analyzed in order to extract the research contexts used to study human-AI collaboration (RQ 1). Similar research contexts were grouped into research context groups for further analysis. In order to understand to what extent these contexts can be classified as CPS, the CPS taxonomy of [12] was applied (RQ 2).

4.1. Research contexts of human-AI collaboration (RQ 1)

Each of the identified publications describes an instantiation of human-AI collaboration in a specific *research context*. Similar research contexts are grouped for reasons of clarity. Table 2 provides an overview regarding the research context groups, the identified examples and the references. We also report exemplary key inputs, mediators and outcomes (IMO) as well as the key findings for the studies, in order to provide readers with inspiration and to facilitate the analysis.

The following describes the research context groups along with information regarding the reasoning for using AI in this research context.

Creative content creation. Several studies deal with the creation of creative content (e.g., composing music or drawing a picture). Usually, a human performs an initial creative action (e.g., draws first lines of an idea onto a digital canvas) and can then evoke the AI system to add to the composition. Authors root their argument for choosing AI systems in their research context in literature citing several benefits from previous studies, e.g., [18] cite the effective provision of inspiration, stimulation of divergent thinking and support for innovation.

Turn-based cooperative games. There is a variety of cooperative games (e.g., card game or puzzles). Usually it is two players taking actions in alternating order. While the puzzles game [19] allows simultaneous action by both players, it does not necessitate simultaneous action-taking and is thus categorized as turn-based here. The reasoning for choosing AI systems is less related to their abilities within this context but more to the research interest for mediators like social perception [20], effective communication [21] or explanations [19].

Real-time cooperative games & simulations. There are several real-time (video) games requiring players to move around and perform actions swiftly. As will be described in the next section, these two types of games (i.e., turn-based, real-time) differ in their complexity and are thus sorted into different research context groups. Similarly to turn-based games, research in the real-time research contexts did primarily focus on collaboration aspects e.g., on how beliefs about teammates manifest in actions [22] or behavior [23]. Due to the similarity, simulations are included in this group.

Design. Five studies are related to design teams supported by AI systems, e.g., focusing on the creation of mood boards ("visual collages composed of images, text, and objects, that express concepts, ideas and emotions" [24]) or on engineers designing bridges [25] or drones [26]. While the humans are in the lead, the AI system provides suggestion for further ideas or improvements. AI systems were chosen to support in those research contexts due to their high task-specific performance expected to be effective when combined with human creativity and agility [25] and their ability to support the entire creative process including inspiration and reflection [24].

Negotiation. In one study, two sellers (two AI systems) competitively negotiate with a human buyer regarding a transaction. The human buyer and AI system seller might be seen as collaborators with the shared goal of reaching an agreement, making them better off.

With each of them aiming to maximize their own utility, they have conflicting goals. This, however, is common within collaborative settings [12]. The research context was chosen due to potential benefits for "retail, e-commerce, legal, business, and industrial sectors" [27].

Investigation. In one study, an AI system jointly with a team of conservation scientists, law enforcement and criminologists iteratively analyze data to identify instances of illegal trade of plants. AI is used to improve the effectiveness of search and analysis [28].

In summary, most of the human-AI collaboration research contexts are either real-life creative content creation or design tasks, or well-defined game environments (both turn-based and real-time). For the former, the AI systems generative ability is supposed to provide inspiration to humans supporting the creative process. In many cases, performance with and without AI is investigated (input – team composition). For the latter, the focus is usually not set on game-specific aspects but on researching aspects of collaboration such as effective communication, focusing more on mediating aspects.

Reflecting on the characteristics of collaboration described earlier, an *evolving process* and *reciprocity* are essential. All these research context require a certain degree of sense-making, i.e., trying to understand the situation as well as actions and intentions of the other(s) and adjusting accordingly (e.g., match the intentioned drawing style [29]). They are also inherently iterative and interdependent, requiring a back-and-forth between agents. Games and simulations allow researchers more fine-grained control over the environment enabling them to enforce those characteristics, e.g., by creating hard dependencies through providing agents with unique capabilities or restricting access to information, making communication necessary to achieve the joint goal [30]. This is more difficult for real-world tasks (e.g., writing, drawing, composing).

Measuring team effectiveness is a way to assess and compare designs options. While for all research contexts some performance outcome metric is used (see table 2), researchers should keep in mind, that for game environments measuring performance is usually comparatively easy, as a scoring system is (in most cases) inherently available. For real-world tasks an external (expert) judgment is required, increasing study effort. Self-reported metrics on (agent) performance (commonly used in the included studies) should be used with caution as these might not align with actual performance (e.g., [20]). For virtual research environments an in-depth analysis of the log data might be fruitful (e.g., [31]).

In the included studies outcomes were measured on team, but not on role or organizational level, which might be due to the novelty of the research field.

4.2. Complex problem solving (RQ 2)

In order to understand to what extent CPS is already addressed in the identified studies on human-AI collaboration, an assessment against the CPS taxonomy [8] introduced earlier is performed.

Each of the research contexts identified is evaluated according to the nine characteristics of CPS. Results are aggregated towards research context groups and feature groups (time related, variable related, system behavior related) for clarity. As a result, for each research context group the complexity along the three feature groups is displayed in Table 3. Complexity is assessed and marked as follows:

- *Complex* – marked with 'X': assigned to a feature group, if the majority of research contexts fulfilled all three features of complexity
- *Somewhat complex* – marked with '(X)': assigned, if the majority of research contexts fulfilled at least one feature of complexity

In the following, the reasoning for each assessment is explained briefly. The assessments only apply to the research contexts analyzed as part of this literature review without a claim to generalize to other research contexts that could be considered under these labels. Features not mentioned explicitly within the explanations do not fulfill the complexity features for the respective research context groups marked somewhat complex.

Creative content creation. Typically, humans work together with AI systems on the "drawing board" in these research contexts with no time related complexity. The exception among the assessed creative content creation research contexts is the live improvisation of music [32]. While drawing, many interrelated variables (e.g., position, shape, coloring, size) need to be considered, making it complex in this regard. With regards to the system behavior, the research contexts are somewhat complex due to the delayed feedback, however, without being opaque or stochastic.

Turn-based cooperative games. The turn-based games within the analysis follow discrete steps without changes to the environment (which are not player induced) and are thus not complex with regards to time. They require users to reason considering a variety of interrelated factors and can thus be considered somewhat complex. Given the adjusted understanding of opacity described earlier, the games analyzed here are not considered opaque as the team can perceive the entire game state. However, due to the delayed feedback they can be considered somewhat complex.

Real-time cooperative games & simulations. With time passing and the environment changing continuously, participants are constantly required to decide, which action to perform. Waiting is also considered an (implicit) action. Thus, real-time cooperative

Table 2. Research context groups with examples and exemplary study details

Research context groups/*Example(s)*	**Key Outcome(s)**	**Key Input(s)/ Mediator(s)**	**Key Finding(s)**
Creative content creation *Sketching [18, 29], composing [32, 33], writing [34, 35]*	*PO:* Novelty, integrity, interestingness & balance [18, 32], scariness [34] *PB:* Social & collaboration dynamics, flow [29, 32, 33] *AP:* Match writing style, novelty, creativity [35]	*I-TC:* AI vs. WoZ [29] *I-TC:* AI vs. no-AI [18, 33] *I-TC*: AI vs. human-AI vs. human [34] *M-TP:* Communication: inner state [32], suggestion style [35]	*Positive effect:* Visualization of machine confidence on flow & composition [32] *Positive effect:* AI on novelty [18] and social dynamics [33] *Positive effect:* Hybrid approach on scariness [34] Writer prefer more fine-grained control over outputs and editability [35] *Inconclusive:* Group comparison; focus on framework [29]
Turn-based cooperative games *Card game (Hanabi), puzzle, word guessing [19–21]*	*PO*: Score, win/loose, efficiency [20, 21] *AP*: Helpfulness, intelligence, sociability, humanness, likability, creativity, trustworthiness [19–21]	*I-TC*: AI vs. presumed-human [20] *M-TP*: Communication: implicature & explanations [19, 21]	*Positive effect*: Implicature on score and perceived humanness [21] *Positive effect*: Explanation style on helpfulness, trustworthiness and overall experience [19] *Negative attitude* towards AI despite equal performance [20]
Real-time cooperative games & simulations *Dearth, capture the gunner, defend the pass, don't starve together, reconnaissance missions overcooked [22, 23, 31, 36–38]*	*PO:* Score [36], efficiency [37], mission success [38] *PB*: Protecting or sacrificing teammate [22, 23] *AP:* Player identity (humanness) [31], perceived helpfulness, efficiency [37], trust, understandability [38]	*I-TC:* AI vs. human [31, 36] *I-TC*: AI vs. presumed-human [22, 23] I-E: Score visible [23] *M-TP*: Communication: different modalities [36]/ explanation types [37, 38]/ skill, reliability [38]	*Dependability:* Visibility of score on treatment of AI teammate [23] Humans act differently in presumed social context [22] Online players value co-presence [36] *Positive effect*: Staying close to other player on correct identification (AI or human) [31] *Positive effect*: explanation on non-experts' performance [37]/on performance if facilitating decision-making [38]
Design *Mood board-, bridge-, drone-, accessory-, game level- design [24–26, 39, 40]*	*PO*: Strength mass ratio [25], quality, novelty [40], range, velocity, payload, cost [26] *PB:* Design effort, search strategy, mental workload [26], solution exploration [24] *AP*: Usefulness, inspirational [24], fun, frustration, aid, creativity, adaptation [39], friendliness, engagement [40]	*I-TC:* AI vs. no-AI [25, 26] *I-E*: Task complexity [26] *M-TP*: task-related algorithms [39], embodiment [40], agency variants (different features) [24]	*Positive effect*: design on low performing teams (*negative* on high performing teams) [25] *Dependability* Agency preference of designer and design step; AI algorithm for outcome [24] *Positive effect:* embodiment on satisfaction, friendliness, engagement & helpfulness [40] *Dependability* Algorithm preference on designer [39] *Positive effect:* AI on efficiency of solution space exploration [26] *Impact:* Task complexity on AI systems effects [26]
Negotiation *Baking items [27]*	*PO*: Profit *PB*: Likeability/trust	*M-TP*: Agents variants – negotiation tactics	*Positive effect*: Specfic negotiation tactics on outcomes and likeability
Investigation *Illegal trade [28]*	*PO*: Efficiency	*I-TC:* AI vs. no-AI	*Presumed positive effect*: Hybrid approach on effiecency

PO = performance outcome, PB = performance behavior, AP = agent performance,
I-TC = input – team composition, I-E: input – environment, M-TP = mediator – team process; WoZ = Wizard of Oz

games are considered complex with regards to time. Similar to turn-based games, they also require participants to consider a variety of interconnected variables (somewhat complex). Lastly, besides delayed feedback, some research contexts are stochastic or opaque making them somewhat complex regarding system behavior.

Design. Similar to the creative content generation group, humans create their designs iteratively and the AI provides feedback or suggestions. The research contexts are set in a dedicated design program with no time related complexity. However, a plethora of different, non-linearly related variables need to be considered (e.g., determining if a constructed bridge can carry a certain weight) to successfully achieve the goal, making it complex in this regard. The system behavior can be described as somewhat complex, usually deterministic and transparent but with delayed feedback.

Negotiation. The negotiation is performed in discrete steps (offer, counteroffer, etc.) and is therefore not considered to be complex with relation to time. It requires participants to keep track of several interrelated variables, however, mostly linearly connected, thus considered somewhat complex in this regard. With another AI based system negotiating and the underlying policy of the agent being unknown to the human-AI team, the system behavior can be considered complex.

Investigation. Though time passes continuously and (bad) actors might continue to trade plants during participants deliberating, the system evolves slowly (i.e., low degree of time pressure) and is thus only somewhat time-complex. The team needs to keep track of a large number of interconnected variables and is therefore considered complex in this regard. Lastly, with other people acting within the system, the behavior is opaque, stochastic and provides only delayed feedback (e.g., after offline investigation), making it complex.

In summary, the research contexts are mostly simulations of problems, which are not complex with regards to time as they are usually static, meaning "changes occur only whether participant intervenes" [8]. The real-time cooperative games & simulations are an exception. In this research context group, while participants deliberate, the environment can change ("move on"). Many research contexts are not carefully constructed "microworlds" but model real world problems with many, non-trivially interrelated variables and are therefore at least somewhat complex with regards to variables. Similarly, the system behavior

is at least somewhat complex for all research contexts due to the delayed feedback, with some research contexts even being stochastic or opaque.

Table 3. Research context groups and complexity assessments

Research context group	T	V	SB
Creative content creation		X	(X)
Turn-based cooperative games		(X)	(X)
Real-time cooperative games & simulations	X	(X)	(X)
Design		X	(X)
Negotiation		(X)	X
Investigation	(X)	X	X

T = time related, V = variable related, SB = system behavior related, X = complex; (X) = somewhat complex

While in AI research reporting on the environment characteristics is common, only few authors explicitly report on the characteristics in the included studies. An exceptions is [26], reporting characteristics in detail and pointing out the before mentioned advantage of game and simulation environments: fine-grained control.

[26] study is particularly relevant in the context of this paper, as [26] investigate *complexity* as one of their input factors and find significant changes in the AI systems effect on performance: "AI assistance is most beneficial when addressing moderately complex objectives but exhibits a reduced advantage in addressing highly complex objectives". Only one other study investigates environment changes, similarly reporting changes in strategy [23]. This sensitivity of results to complexity is well-documented in CPS literature [9]. We therefore encourage researchers on CPS through human-AI collaboration to investigate the robustness of their findings with regards to complexity, e.g., by using systematic variation of system properties [9]. This will enable researchers to more precisely describe to which contexts their findings generalize to. While for real-world task practical value might be easier to demonstrate, for research results in games and simulations the argument must be made convincingly, as results on CPS in the laboratory might not necessarily transfer into practice [9].

While none of the research contexts are considered complex according to all features they might still be defined as complex as "[o]f all the features studied, none is necessary and sufficient to define a task as 'complex'" [8]. Thus, there is already human-AI collaboration research dealing with (aspects of) CPS that can be used as a foundation to address this recently posed challenge.

5. Limitations

Several limitations need to be considered. The search term was informed by the theoretical concepts surrounding the broad term "working together" and focused on "collaboration" and "teaming" due to their similarity and fit to what was described as a research challenge by [1]. However, joint work of humans and AI is studied under a variety of labels and even though we additionally considered "cooperation" and "coordination", potentially not all relevant works are covered. Additionally, this paper builds on a cross-disciplinary literature-based definition of collaboration from a widely cited article in reviewing and excluding search results. Using a different understanding of collaboration might lead to a different set of papers and potentially different results. In combination with the novelty of the research challenge only 22 articles were included. Therefore, we do not claim conclusiveness. On the contrary, our work should be complemented and extended, as the breadth of research contexts in the scientific literature evolves.

Additionally, the conceptualization of "CPS" is, even though having a long research history [14], still fuzzy [8]. There are many definitions and potential criteria inspired by different understandings of what should be studied under the label of "CPS" [9]. Here, the formal features of a literature-informed, widely cited taxonomy of CPS tasks was used. In favor of an integrated framework, frequently used criteria such as the necessity to use "creativity as opposed to routine behavior" [14] are neglected. Furthermore, some of those formal features of the taxonomy, while widely used for determining complexity (e.g., number of variables [8, 41]) lack clearly defined decision boundaries and are difficult to apply ex post, without insight into the research context instantiation process.

6. Conclusion and outlook

The nature of work is evolving and is increasingly characterized by complexity, requiring among other things solving complex problems [2]. Therefore, [1] define CPS through human-AI teams as a major challenge. The goal of this study was to understand to what extent this challenge is already addressed by extant research.

We offer a literature-informed clarification of the two key concepts: human-AI collaboration and CPS. Based on that, we perform a structured literature review to provide an overview regarding the instantiated research contexts used to perform research on CPS through human-AI collaboration. Most research contexts assessed are either creative content generation/ design settings or cooperative games. Applying the features of CPS from literature, most assessed research contexts are not complex with regards to time, but somewhat complex with regards to variables and system behavior. While none fulfill all complexity features, some might still be considered complex. Thus, extant human-AI collaboration literature can be leveraged as a foundation for the new challenge of complex problem solving.

Besides the identified research contexts, we provide insight into the suitability of research context for studying CPS through human-AI collaboration. Games and simulations offer the advantage of flexible adjustments, which is helpful for enforcing characteristics of *collaboration* (e.g., reciprocity) and exploring sensitivity to *complexity* variations (e.g., opacity). Real-world tasks (e.g., drawing) are more difficult to control and results more difficult to measure. However, justification of practical value might be easier for the latter as compared to lab environments (see CPS literature [9]).

We contribute to research in multiple ways. By describing the two key concepts and explicating their defining characteristics, we provide foundational terminology and frame the research challenge more clearly.

For researchers interested in human-AI collaboration through CPS (who are not committed to a specific use case) we provide inspiration by pointing out suitable research context examples. We provide guidance for research context selection or construction by explaining important aspects (e.g., flexibility, measurability) and study design by highlighting the sensitivity to complexity and recommending systematic variation of system properties to achieve more robust results [9].

For researchers who have conducted a study we provide a structure along which they can describe their contribution more precisely. Assuming findings generalize better to more similar situations, the taxonomy offers a way to compare research contexts and more precisely describe the relevant class of problems. Given the sensitivity of results to complexity (documented in CPS literature; also shown e.g., by [26]), this is important.

Conceptually, going forward we encourage the identification of additional suitable research contexts, e.g., through the evaluation of existing testbeds from reinforcement learning research and proven contexts from CPS literature regarding their suitability for CPS through human-AI-collaboration research. We also advocate to consider additional dimensions of complexity (e.g., technological or organizational) in future work. This should be done with the type of problems in mind that [1] mentioned when posing this research challenge: e.g., managerial, political or science problems – "tasks that so far seem to be at the core of human intellect" [1].

Acknowledgement. This research was funded by the German Federal Ministry of Education and Research (BMBF) in the context of the project *HyMeKI* (reference number: 01IS20057).

7. References

[1] Dellermann, D., P. Ebel, M. Söllner, and J.M. Leimeister, "Hybrid Intelligence", Business & Information Systems Engineering, 61(5), 2019, pp. 637–643.

[2] Jacobs, R.L., "Knowledge Work and Human Resource Development", Human Resource Development Review, 16(2), 2017, pp. 176–202.

[3] Grace, K., J. Salvatier, A. Dafoe, B. Zhang, and O. Evans, "Viewpoint: When Will AI Exceed Human Performance? Evidence from AI Experts", Journal of Artificial Intelligence Research, 62, 2018, pp. 729–754.

[4] Dellermann, D., A. Calma, N. Lipusch, T. Weber, S. Weigel, and P. Ebel, "The Future of Human-AI Collaboration: A Taxonomy of Design Knowledge for Hybrid Intelligence Systems", in Proceedings of the 52nd Hawaii International Conference on System Sciences, T. Bui, Editor. 2019.

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

[6] Seeber, I., E. Bittner, R.O. Briggs, T. de Vreede, G.-J. de Vreede, A. Elkins, R. Maier, A.B. Merz, S. Oeste-Reiß, N. Randrup, G. Schwabe, and M. Söllner, "Machines as teammates: A research agenda on AI in team collaboration", Information & Management, 57(2), 2020.

[7] Benbya, H., S. Pachidi, and S.L. Jarvenpaa, "Special Issue Editorial: Artificial Intelligence in Organizations: Implications for Information Systems Research", Journal of the Association for Information Systems, 22(2), 2021.

[8] Quesada, J., W. Kintsch, and E. Gomez, "Complex problem-solving: a field in search of a definition?", Theoretical Issues in Ergonomics Science, 6(1), 2005, pp. 5–33.

[9] Frensch, P.A. and J. Funke, Complex Problem Solving: The European Perspective, Taylor and Francis, Hoboken, 1995.

[10] Maedche, A., C. Legner, A. Benlian, B. Berger, H. Gimpel, T. Hess, O. Hinz, S. Morana, and M. Söllner, "AI-Based Digital Assistants", Business & Information Systems Engineering, 61(4), 2019, pp. 535–544.

[11] Gregor, S. and A.R. Hevner, "Positioning and Presenting Design Science Research for Maximum Impact", MIS Quarterly, 37(2), 2013, pp. 337–355.

[12] Bedwell, W.L., J.L. Wildman, D. DiazGranados, M. Salazar, W.S. Kramer, and E. Salas, "Collaboration at work: An integrative multilevel conceptualization", Human Resource Management Review, 22(2), 2012, pp. 128–145.

[13] Mathieu, J., M.T. Maynard, T. Rapp, and L. Gilson, "Team Effectiveness 1997-2007: A Review of Recent Advancements and a Glimpse Into the Future", Journal of Management, 34(3), 2008, pp. 410–476.

[14] Dörner, D. and J. Funke, "Complex Problem Solving: What It Is and What It Is Not", Frontiers in psychology, 8, 2017, p. 1153.

[15] Yang, Q., A. Steinfeld, C. Rosé, and J. Zimmerman, "Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design", in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020. ACM: NY, USA.

[16] Fischer, A., S. Greiff, and J. Funke, "The Process of Solving Complex Problems", The Journal of Problem Solving, 4(1), 2012.
[17] vom Brocke, J., A. Simons, B. Niehaves, K. Riemer, and A. Cleven, "Reconstructing the Giant: On the Importance of Rigour in Documenting the Literature Search Process", 17th European Conference on Information Systems (ECIS). 2009.
[18] Zhang, C., C. Yao, J. Liu, Z. Zhou, W. Zhang, L. Liu, F. Ying, Y. Zhao, and G. Wang, "StoryDrawer: A Co-Creative Agent Supporting Children's Storytelling through Collaborative Drawing", in Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. 2021. ACM: NY, USA.
[19] Tabrez, A., S. Agrawal, and B. Hayes, "Explanation-Based Reward Coaching to Improve Human Performance via Reinforcement Learning", in HRI '19: Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction. 2019. IEEE Press.
[20] Ashktorab, Z., Q.V. Liao, C. Dugan, J. Johnson, Q. Pan, W. Zhang, S. Kumaravel, and M. Campbell, "Human-AI Collaboration in a Cooperative Game Setting: Measuring Social Perception and Outcomes", Proc. ACM Hum.-Comput. Interact., 4(CSCW2), 2020.
[21] Liang, C., J. Proft, E. Andersen, and R.A. Knepper, "Implicit Communication of Actionable Information in Human-AI Teams", in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 2019. ACM: NY, USA.
[22] Merritt, T. and K. McGee, "Protecting Artificial Team-Mates: More Seems like Less", in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2012. ACM: NY, USA.
[23] Ong, C., K. McGee, and T.L. Chuah, "Closing the Human-AI Team-Mate Gap: How Changes to Displayed Information Impact Player Behavior towards Computer Teammates", in Proceedings of the 24th Australian Computer-Human Interaction Conference. 2012. ACM: NY, USA.
[24] Koch, J., N. Taffin, M. Beaudouin-Lafon, M. Laine, A. Lucero, and W.E. Mackay, "ImageSense: An Intelligent Collaborative Ideation Tool to Support Diverse Human-Computer Partnerships", Proc. ACM Hum.-Comput. Interact., 4(CSCW1), 2020.
[25] Zhang, G., A. Raina, J. Cagan, and C. McComb, "A cautionary tale about the impact of AI on human design teams", Design Studies, 72, 2021, p. 100990.
[26] Song, B., N.F. Soria Zurita, H. Nolte, H. Singh, J. Cagan, and C. McComb, "When faced with increasing complexity: The effectiveness of AI assistance for drone design", Journal of Mechanical Design, 2021, pp. 1–38.
[27] Geraghty, R., J. Hale, S. Sen, and T.S. Kroecker, "FUN-Agent: A 2020 HUMAINE Competition Entrant", in Proceedings of the 1st International Workshop on Multimodal Conversational AI. 2020. ACM: NY, USA.
[28] D. Whitehead, C.R. Cowell, A. Lavorgna, and S.E. Middleton, "Countering plant crime online: Cross-disciplinary collaboration in the FloraGuard study", Forensic Science International: Animals and Environments, 1, 2021, p. 100007.
[29] Davis, N., C. Hsiao, K.Y. Singh, B. Lin, and B. Magerko, "Quantifying Collaboration with a Co-Creative Drawing Agent", ACM Trans. Interact. Intell. Syst., 7(4), 2017.
[30] van den Bosch, K., T. Schoonderwoerd, R. Blankendaal, and M. Neerincx, "Six Challenges for Human-AI Co-learning", in Adaptive Instructional Systems, R.A. Sottilare and J. Schwarz, Editors. 2019. Springer International Publishing: Cham.
[31] Tulk, S., R. Cumings, T. Zafar, and E. Wiese, "Better Know Who You Are Starving with: Judging Humanness in a Multiplayer Videogame", in Proceedings of the Technology, Mind, and Society. 2018. ACM: NY, USA.
[32] McCormack, J., T. Gifford, P. Hutchings, M.T. Llano Rodriguez, M. Yee-King, and M. d'Inverno, "In a Silent Way: Communication Between AI and Improvising Musicians Beyond Sound", in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 2019. ACM: NY, USA.
[33] Suh, M., E. Youngblom, M. Terry, and C.J. Cai, "AI as Social Glue: Uncovering the Roles of Deep Generative AI during Social Music Composition", in Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 2021. ACM: NY, USA.
[34] Yanardag, P., M. Cebrian, and I. Rahwan, "Shelley: A Crowd-Sourced Collaborative Horror Writer", in Creativity and Cognition. 2021. ACM: NY, USA.
[35] Calderwood, A., V. Qiu, K.I. Gero, and L.B. Chilton, "How Novellists use Generative Language Models: An Exploratory User Study", IUI '20 Workshops, 2020.
[36] McGee, K., T. Merritt, and C. Ong, "What We Have Here is a Failure of Companionship: Communication in Goal-Oriented Team-Mate Games", in Proceedings of the 23rd Australian Computer-Human Interaction Conference. 2011. ACM: NY, USA.
[37] Gao, X., R. Gong, Y. Zhao, S. Wang, T. Shu, and S.-C. Zhu, "Joint Mind Modeling for Explanation Generation in Complex Human-Robot Collaborative Tasks", in 29th IEEE International Conference on Robot and Human Interactive Communication. 2020: NY, USA.
[38] Wang, N., D.V. Pynadath, and S.G. Hill, "The Impact of POMDP-Generated Explanations on Trust and Performance in Human-Robot Teams", in Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems. 2016. ACM: New York, USA.
[39] Guzdial, M., N. Liao, J. Chen, S.-Y. Chen, S. Shah, V. Shah, J. Reno, G. Smith, and M.O. Riedl, "Friend, Collaborator, Student, Manager: How Design of an AI-Driven Game Level Editor Affects Creators", in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 2019. ACM: NY, USA.
[40] Lin, Y., J. Guo, Y. Chen, C. Yao, and F. Ying, "It Is Your Turn: Collaborative Ideation With a Co-Creative Robot through Sketch", in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020. ACM: NY, USA.
[41] Funke, J., "Complex Problem Solving", in Encyclopedia of the Sciences of Learning, N.M. Seel, Editor. 2012. Springer US: Boston, MA.