Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

Measuring successful cooperation in human-AI teamwork: Development and validation of the perceived cooperativity and teaming perception scales

Christiane Attig, Christiane Wiebel, Patricia Wollstadt, Tim Schrills, Mourad Zoubir, Thomas Franke

2026

Preprint:

This is an accepted article published in International Journal of Human-Computer Interaction. The final authenticated version is available online at: https://doi.org/10.1080/10447318.2026.2739657

Measuring successful cooperation in human-AI teamwork: Development and validation of the perceived cooperativity and teaming perception scales

Christiane Attig[1,2], Christiane Wiebel-Herboth[1], Patricia Wollstadt[1], Tim Schrills[2], Mourad Zoubir[2], and Thomas Franke[2]

[1]Honda Research Institute Europe GmbH, Offenbach am Main, Germany

[2]Institute of Human-Centered Interactive Systems, University of Luebeck, Luebeck, Germany

Author Note

Christiane Attig https://orcid.org/0000-0002-6280-2530 Christiane Wiebel-Herboth https://orcid.org/0000-0002-6574-3820 Patricia Wollstadt https://orcid.org/0000-0002-7105-5207 Tim Schrills https://orcid.org/0000-0001-7685-1598 Mourad Zoubir https://orcid.org/0000-0003-4458-8443 Thomas Franke https://orcid.org/0000-0002-7211-3771

Correspondence concerning this article should be addressed to Christiane Attig, Honda Research Institute Europe GmbH, Carl-Legien-Str. 30, 63073 Offenbach am Main, Germany. E-mail: christiane.attig@honda-ri.de

The Version of Record of this manuscript has been published and is available in International Journal of Human–Computer Interaction (2026), https://doi.org/10.1080/10447318.2026.2739657.

Abstract

As human–AI cooperation becomes increasingly prevalent, reliable instruments for assessing the subjective quality of cooperative human–AI interaction are needed. We introduce two theoretically grounded scales: the Perceived Cooperativity Scale (PCS), grounded in joint activity theory, and the Teaming Perception Scale (TPS), grounded in evolutionary cooperation theory. The PCS captures an agent's perceived cooperative capability and practice within a single interaction sequence; the TPS captures the emergent sense of teaming arising from mutual contribution and support. Both scales were adapted for human–human cooperation to enable cross-agent comparisons. Across three studies (complete $N = 409$) encompassing a cooperative card game ($N_1 = 94$), LLM interaction ($N_2 = 182$), and a decision-support system ($N_3 = 133$), analyses of dimensionality, reliability, and validity indicated that both scales successfully differentiated between cooperation partners of varying cooperative quality and showed construct validity in line with expectations. The scales provide a basis for empirical investigation and system evaluation across a wide range of human–AI cooperation contexts.

Keywords: human-AI teaming, human-AI interaction, cooperation, user experience, scale development

Measuring successful cooperation in human-AI teamwork: Development and validation of the perceived cooperativity and teaming perception scales

1. Introduction

Over the past few years, developments in artificial intelligence (AI) have been profoundly transformative. What has once been the realm of science fiction—conversing in natural language with artificial agents that remember prior interactions, assist with a wide range of tasks, and can be perceived as team members, friends, or even relationship partners—has become part of daily life for millions of people (Wolf & Maier, 2024). This shift is largely driven by Large Language Models (LLMs) based on generative pre-trained transformers (GPTs; Akhtar, 2024). In parallel, other machine and deep learning systems have advanced, for instance, in visual pattern recognition (e.g., cancer detection in medical imaging; Javed et al., 2024, lane detection in autonomous vehicle control; Zakaria et al., 2023, threat detection in critical infrastructure; Budžys et al., 2024) and prediction (e.g., of pandemic diseases; Ajagbe and Adigun, 2024, market trends; Al-Khasawneh et al., 2025, or energy consumption; Zhang et al., 2024), broadening AI's reach across sectors such as private life, healthcare, transportation, and industry.

Given that AI systems are integrated in an increasing number of applications and become part of users' daily routines, understanding the psychological dynamics that arise during human-AI interaction is crucial to enable AI experiences that have a positive impact on individual user satisfaction as well as societal flourishing (Attig et al., 2024). The importance of this understanding is stressed by first findings indicating potentially adverse implications of AI use on individual, group, and societal levels. For instance, on the individual level, interacting with AI at work has the potential to heighten feelings of loneliness (Tang et al., 2023), boredom, and demotivation (Wu et al., 2025). On the group level, integrating an AI into a hybrid team might lead to deteriorating team cognition (McNeese et al., 2021; Sidji et al., 2024). On broader social structures, AI interaction might cause negative spillover effects, such as undermining social norms of reciprocity in traffic settings (Shirado et al., 2023) or of politeness in communicative situations (Tey et al., 2024). Thus, a core challenge for human factors and engineering psychology is to advance theoretical and methodological

foundations that can account for these multi-level effects and guide the development of human-AI interactions that are as effective as they are aligned with human psychology and social responsibility (Attig et al., 2024; Schmager et al., 2025; Sendhoff & Wersing, 2020).

Importantly, such individual and societal effects are likely shaped by the depth to which humans and AI systems are *integrated*. Regarding a single task—what we term *synchronic integration*—human-AI coupling can vary from minimal, task-specific assistance to highly interdependent integration of human and AI information processing and behavior (Johnson et al., 2018; O'Neill et al., 2022). Moreover, when looking beyond a single interaction—what we term *diachronic integration*—as in repeated collaborations and long-term cooperation, the degree of coupling can range from occasional, loosely connected joint actions to enduring partnerships in which humans and AI systems develop shared routines, mutual influence, and a joint history of interaction (Chang & Herath, 2025; Kirk et al., 2025; Qian & Wan, 2025). Particularly regarding cooperative tasks in performance-related settings (e.g., clinical decision-making, collaborative problem-solving, education), tighter synchronic and diachronic integration can blur boundaries between human and AI contributions, influencing roles, responsibilities, and the dynamics of teamwork (Basappa et al., 2025; Schelble et al., 2025). Thus, hybrid teams with high interdependence of tasks represent a context in which individual, team, and potentially societal dynamics are especially sensitive to disruption by AI, highlighting the urgent need to understand and guide these interactions through human-centered design.

A key prerequisite for understanding the factors that shape effective and socially responsible human-AI cooperation is the availability of context-specific instruments that can reliably capture humans' subjective perceptions of the cooperative dynamics. While a large variety of questionnaire scales have been developed over the last decade to assess users' general attitudes towards AI (Stein et al., 2024) and AI literacy (Lintner, 2024), there remains a need for specialized instruments capable of capturing users' subjective perceptions of specific cooperative human-AI interaction qualities and team dynamics. In the present work, we introduce two theoretically grounded scales designed to reliably assess users' perceptions of both the cooperative capabilities of artificial agents and the broader dynamics of hybrid

teamwork. Moreover, the scales were also adapted for human-human teamwork to enable comparisons of AI and human cooperation partners. Across three studies encompassing different contexts and levels of human-AI integration, we collected data to validate the questionnaires. The resulting scales not only enable rigorous empirical investigation of hybrid teamwork dynamics but also provide practical guidance for designing and evaluating AI systems that can integrate seamlessly into collaborative environments.

2. Background

2.1 Synchronic and diachronic dimensions of human-AI integration

To evaluate the success of human–AI cooperation on a psychological level, it is essential to first clarify the degree of integration between human and AI partners and the durability of their cooperation, since different psychological processes come into play at varying levels of integration and durability. For a systematic understanding of the cooperative structure, we propose a differentiation between a *synchronic* and a *diachronic* dimension of integration (see Figure 1). Borrowed from linguistics where they were introduced by Ferdinand de Saussure, these terms distinguish between phenomena examined at a single point in time (synchronic) and phenomena examined as they unfold and develop over time (diachronic; Buchanan, 2010).

With the term *synchronic integration* we refer to the degree of integration between human and AI information processing and behavior within a *single task sequence*. Single tasks with low interdependency between partners are typically characterized by loose synchronic integration, meaning that the task is mostly completed either by the human or by the AI system. For instance, low synchronic integration occurs when a user submits a simple prompt to an LLM-based chatbot (e.g., ChatGPT, Claude) to draft an email to their doctor and sends the generated text without modification. In contrast, high synchronic integration involves a more iterative and tightly coupled process, where the user provides a more elaborated initial prompt, evaluates the AI's response, refines the prompt, and repeats this cycle until the output meets their requirements. This illustrates cognitive interdependence (Alqasir, 2025), where human and AI information processing and action regulation are closely intertwined in shaping the final outcome (Schrills & Franke, 2026).

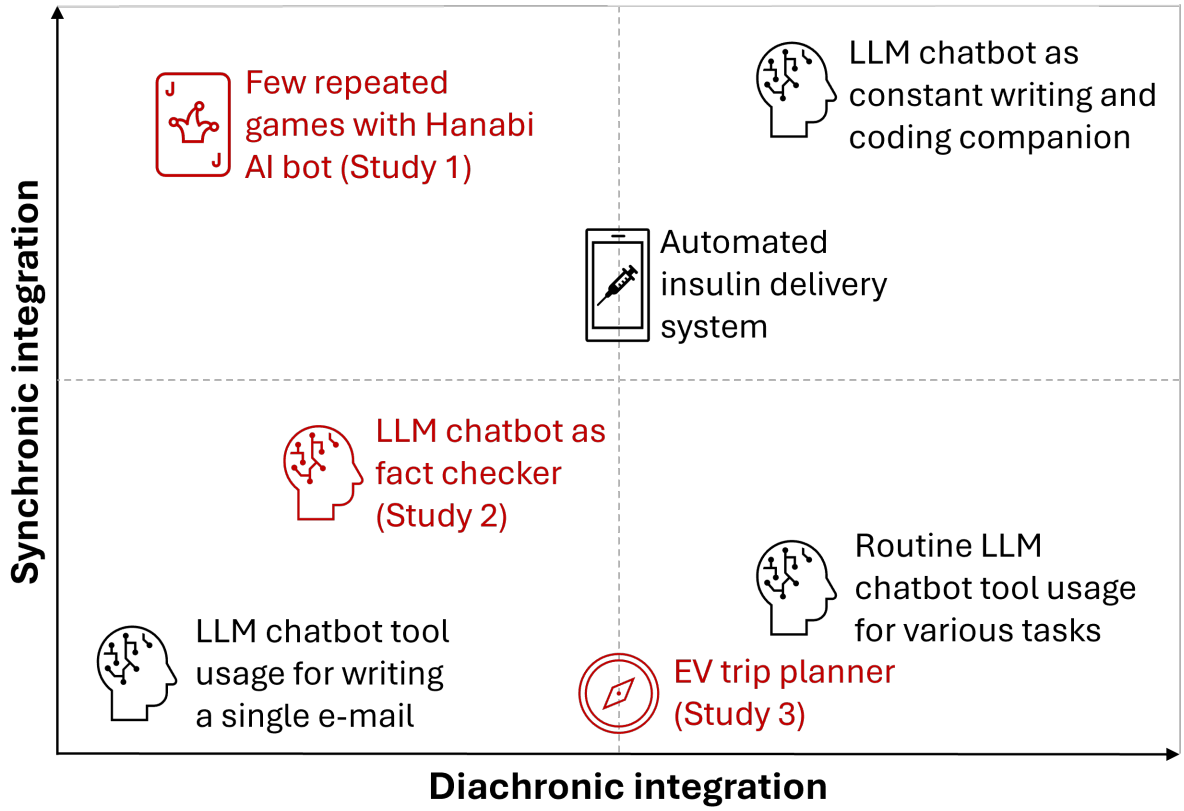

Figure 1

Illustration of two-axis model of synchronic and diachronic dimensions of human-AI integration. Synchronic integration refers to the degree of integration between human and AI information processing and behavior within a single task sequence. Diachronic integration refers to the degree of integration between human and AI information processing across different repeated task sequences.

In contrast, with the term *diachronic integration* we refer to the degree of integration between human and AI information processing across different *repeated task sequences*. Loose diachronic integration is characterized by occasional joint actions. For instance, imagine a user who consults an LLM-based chatbot every few months when drafting professional emails, relying on the AI only sporadically and performing most of the remaining work independently. Conversely, deep diachronic integration occurs when a user routinely consults the chatbot across a wide range of tasks over time, for instance, using it briefly for email drafting one day, brainstorming meeting agendas the next, and looking up information for a report the week after. Across these interactions, the user develops stable reliance patterns, learns the AI's capabilities and limitations, and gradually integrates the system as a

consistently present entity in their workflow—even though no single interaction may involve deep synchronic coupling. High diachronic integration, however, can also co-occur with high synchronic integration, as when a user collaborates with the chatbot continuously and iteratively across multiple demanding tasks such as manuscript improvement, coding, and data analysis, developing shared routines in which the AI consistently supports planning, drafting, and refining outputs. As another example, consider an automated insulin delivery (AID) system (Schrills et al., 2023). The system is continuously coupled with the human user at a physical level, while the integration of human and AI information processing is better characterized as episodic and interdependent (e.g., when the user checks measured glucose levels, when the system prompts the user for manual input regarding physical exercise). Even though the AID system is used over extended periods of time, the interaction remains confined to a single, physiologically defined task domain. Therefore, despite the continuity of use, diachronic integration remains at a moderate level.

In sum, while synchronic integration refers to the level of interdependence within a single task sequence, diachronic integration captures the continuity and stability of cooperation across multiple task sequences over time. Consequently, synchronic and diachronic integration are assumed to be *relatively* independent dimensions. This independence relates to the *level of integration* characterizing each dimension, rather than an absence of relationship between them. Diachronic integration is necessarily developed through a series of interaction episodes, yet the degree of synchronic integration within any given episode does not determine the degree of diachronic integration that emerges from the continuity and stability of the relationship across episodes, or vice versa. However, the assumed consequences of human-AI coupling on human information processing (e.g., long-term effects of cognitive offloading, Shukla et al., 2025) and hybrid team dynamics (e.g., team development, Schelble et al., 2025) are more likely to manifest when both synchronic and diachronic integration are high.

Our two-axis model has some parallels to classic models of human team development, such as Tuckman's framework of small-group development (Tuckman, 1965). Tuckman distinguishes two areas where group development takes place: a task realm (orientation,

emotional response to task demands, opinion exchange, emergence of solutions) and a social realm (testing-dependence, conflict, cohesion, functional roles), proposing that both progress through corresponding stages as a group persists over time. These two areas loosely map onto our distinction between synchronic (closer resemblance to the task realm) and diachronic integration (closer resemblance to the social realm). However, unlike Tuckman's model, our framework does not structure discrete developmental phases that human-AI teams are assumed to progress through; instead, it describes the degree of human-AI integration along two dimensions. The synchronic dimension is task-related in the sense that it characterizes a single task sequence, but it centers specifically on the interdependence (Klein et al., 2005) and the degree of integration between human and artificial information processing (Schrills & Franke, 2026) within that sequence. The diachronic dimension is likewise task-related, describing over what time period and for which purposes human-AI interaction emerges. Critically, however, we argue that diachronic integration functions as a *precursor* for social team dynamics to emerge: Interaction across episodes are assumed to make relational phenomena comparable to those described by Tuckman (e.g., development of mutual expectations, intragroup conflict, team cohesion; Tuckman, 1965) plausibly arise.

Knowledge about the depth of synchronic and diachronic integration in a given human–AI cooperation process is crucial for (a) identifying the potential psychological dynamics at play, (b) selecting appropriate theoretical frameworks to explain those dynamics, (c) and determining suitable metrics for assessing subjective cooperation success. When diachronic integration is absent or minimal—ranging from single interaction instances to only occasional joint actions—key questions concern the degree of synchronic integration, the extent to which cooperation partners are prepared for this level of interdependence, and the potential consequences that follow. For instance, strong synchronic integration with highly interdependent tasks typically heightens the need for coordination between partners (Klein et al., 2005). If either partner lacks the necessary capabilities for such coordination, the success of cooperation may be compromised (Attig et al., 2024). In contrast, when examining repeated cooperative interactions characterized by high diachronic integration, as for instance in hybrid work environments, consequences of integrating AI agents on team dynamics (e.g.,

team cohesion, Lakhmani et al., 2022; trust calibration over time, de Visser et al., 2020) become more relevant. Hence, different theoretical foundations are required to adequately capture the dynamics of human–AI cooperation, with synchronically integrated interactions benefiting from theories of interdependence, while diachronically integrated interactions call more strongly for longitudinal perspectives on cooperation.

2.2 From joint activity theory to perceived cooperativity

A fruitful foundation for systematically investigating synchronic human–AI integration is joint activity theory (Bradshaw et al., 2017; Klein et al., 2005), which conceptualizes joint activity as an interdependent process that requires the willingness of all parties to contribute to its successful completion. As it centers on the choreography of single joint activities, the theory is particularly well suited to examining the factors that determine the success of individual cooperative task sequences.

According to joint activity theory, successful teams are characterized by three factors: common ground, interpredictability, and directability (Klein et al., 2005). Common ground includes the mutual knowledge, beliefs, and assumptions that enable interdependent action, encompassing both what is shared beforehand and what develops throughout the activity. As such, common ground is the prerequisite for interpredictability, which refers to the accurate anticipation of others' actions. Skilled teams achieve this through shared knowledge and coordination practices developed over time. Finally, directability is the capacity to assess and adapt others' actions as conditions change. Successful coordination depends on participants' responsiveness and mutual commitment to addressing problems as they arise (Bradshaw et al., 2017; Klein et al., 2005).

These factors can be applied to individual teammates to delineate the attributes of an effective agent in joint activities. Bradshaw et al. (2017) identified eight characteristics that define such an agent: (1) observability—making one's own current state and intentions transparent, (2) monitorability—enabling progress appraisal and signaling potential issues, (3) communicativeness—being informative, polite, and context-aware, (4) self-assessment—knowing one's own limitations and balancing initiative with guidance, (5) reliability—being predictable and dependable, (6) directability—remaining responsive to

external guidance and authority, (7) selectivity—focusing attention on what is most important, and (8) coordination—managing dependencies and maintaining common ground.

Following joint activity theory, an agent—whether human or artificial—that embodies these capabilities has the potential to contribute to the success of a joint activity. During interaction, an agent's behavior serves as the key information source for perceiving it as a cooperative partner, since the relevant characteristics become directly observable in the activity and thus allow an assessment of an agent's cooperative potential. Yet potential alone is insufficient; it must be translated into actual cooperative behavior during the interaction itself. Both aspects are captured in the concept that we refer to as *perceived cooperativity*: the perceived capability and cooperative practice of a partner to contribute to a successful joint activity, which becomes particularly relevant in synchronic human-AI integration.

2.3 From evolutionary cooperation theory to teaming perception

While perceived cooperativity captures the subjective perception of a cooperation partner's qualities within a single joint activity (reflecting the synchronic dimension of human–AI integration), interactions with AI systems increasingly extend beyond single episodes into repeated collaborations and long-term partnerships, making the diachronic dimension of integration equally relevant. Given that even machines without human-like features can be perceived as social actors (Nass et al., 1994), and that this tendency is amplified in AI systems with natural language capabilities (DeVrio et al., 2025; Lee & Kim, 2025), the perception of AI as a teammate rather than a tool is an empirically grounded possibility rather than a mere assumption (Attig et al., 2024; Lindgren, 2025; Walliser et al., 2017). Hence, at the diachronic level, the question is no longer only whether a cooperation partner is perceived as cooperative in a given moment, but whether the accumulated experience of repeated cooperative episodes gives rise to a sense of genuinely being a team (Pacherie, 2014).

However, being part of a team entails more than the individual perception of another agent being a teammate. Teaming is a reciprocal relation that integrates a first-person experience of agency (i.e., being an active contributor whose actions matter for the outcome) with a global, collective framing of the activity as something *we* are doing together (Pacherie,

2014). Human–AI teaming thus relies on the simultaneous presence of first- and third-person perspectives, reflecting both (1) individual contributions of all actors and (2) shared, goal-directed actions. Together, the process of shared contribution of resources (e.g., time, cognitive effort, physical strength) towards mutually agreed goals is proposed to constitute the basis of the emergent state of teaming perception (Marks et al., 2001; Rieth et al., 2026).

The emphasis on (1) individual contributions can be theoretically grounded in evolutionary cooperation theory, which defines cooperation through the costs and benefits that actions impose on individuals: Cooperative acts incur a cost to the actor while providing a benefit to others. The emergence and stability of such behavior depend on interaction structures that allow these individually costly acts to be favored by selection, for example through repeated interaction, relatedness, or reputation (Nowak, 2006; West et al., 2007). In this view, to cooperate is to incur a cost that enables another individual to obtain a benefit, stressing that a prosocial investment is the central defining characteristic of cooperative behavior (Attig, Wollstadt, et al., 2026; Attig et al., 2025).

Shared, goal-directed action (2) is not merely the sum of individual efforts but a jointly constructed activity, making mutuality a structural feature of teaming. Genuinely shared action requires a collectively represented goal that both partners are jointly committed to pursuing (Bratman, 1992; Tomasello et al., 2012). This joint commitment requires each partner to actively support the other in performing their role and to deprioritize personal goals that might interfere with the shared objective (Bratman, 1992). From this perspective, the human partner's own cooperative contributions, support, and willingness to set aside individual goals are constitutive of the teaming experience itself: A sense of genuine teamwork cannot emerge from the perception of AI contributions alone but requires the human to also experience themselves as an active, investing member of the team.

Importantly, for this experience of joint commitment and mutual support to arise, the non-human partner must itself be perceived as a goal-directed agent to a minimum degree. Without attributing some form of agency to the AI partner (i.e., without perceiving it as an entity that pursues goals, makes decisions, and actively contributes to the shared objective rather than merely producing outputs), the psychological conditions necessary for genuine

teaming perception likely will not be met (Musick et al., 2021). Perceived agency of the AI partner thus constitutes a prerequisite for teaming perception to emerge, and distinguishes cooperative AI agents from purely instrumental tools in the context of human-AI interaction (Lyons et al., 2021).

We thus conceptualize teaming perception as a subjective state variable that emerges from the interplay of perceived AI prosocial contributions and mutual human–AI support. Based on this conceptualization, we organize the subjective measurement of *teaming perception* into three facets: one capturing the emergent sense of the team as a whole, one capturing the human partner's own cooperative investment and support, and one capturing the human's perception of the AI partner's supportiveness and contributions (Attig et al., 2024).

2.4 Integrating perceived cooperativity and teaming perception

Perceived cooperativity and teaming perception map onto complementary dimensions of the synchronic–diachronic integration framework (see Figure 1): Perceived cooperativity is particularly sensitive to the synchronic dimension, assessing the quality of a single cooperative task sequence regardless of interaction history, while teaming perception extends into the diachronic dimension, capturing the relational experience that accumulates as repeated cooperative episodes build a joint history. Both constructs refer to the individual level and are process-oriented. That is, they target the perceptual and experiential indicators of successful cooperation rather than its outcomes. In this sense, the two scales together offer a measurement approach that tracks the subjective experience of human–AI cooperation across both its synchronic and diachronic dimensions.

The interplay between synchronic and diachronic human–AI integration also implies a developmental relationship between the two constructs: the accumulation of positive cooperative episodes (i.e., episodes characterized by high perceived cooperativity of the AI partner) likely consolidates expectations about the partner's ability to cooperate while also generating positive cooperative experience (Attig, Wollstadt, et al., 2026), thereby giving rise to a genuine sense of teaming over time.

Notably, although teaming perception is theoretically most informative in the context of diachronic integration, it can also be meaningfully assessed after an initial encounter with

an AI system in a cooperative setting, provided that the cooperative episode reaches a sufficient degree of synchronic integration. Under such conditions (i.e., high interdependency between partners to complete the joint task sequence), even a single cooperation instance needs contributions from both partners and enables the potential experience of mutual coordination and joint goal pursuit with the AI partner.

3. Materials and methods

3.1 Scale construction

The Perceived Cooperativity Scale (PCS) and the Teaming Perception Scale (TPS) were developed to assess two complementary aspects of human–AI cooperation: the perceived cooperativity of an AI partner within a given interaction, and the broader sense of teaming that emerges as cooperative episodes accumulate over time. To allow for comparisons between human–AI and human–human cooperation, adapted versions of both scales were additionally developed in which references to the AI agent (PCS-A, TPS-A) were replaced with references to a human partner (PCS-H, TPS-H).

Initial item pools were generated based on the theoretical constructs and prior research on cooperation and teaming. For the PCS, items were developed to capture the perceived cooperativity of an AI agent within a single cooperative task sequence. Based on the eight characteristics of a good agent for joint activity (Bradshaw et al., 2017)—observability, monitorability, communicativeness, self-assessment, reliability, directability, selectivity, and coordination—initial items were formulated to operationalize each characteristic from the perspective of the human user, targeting their subjective perception of the AI partner's behavior during a cooperative task sequence. Item formulations followed closely the descriptions by Bradshaw et al. (2017), while being deliberately rephrased in accessible, everyday language to ensure that participants without a technical background could respond to them intuitively. Moreover, we added the item stem "I had the feeling that" to account for the fact that many characteristics must be inferred rather than directly observed, framing all items consistently as first-person perception judgments rather than objective assessments. For instance, the first characteristic reads as follows: "A good agent is observable. It makes its pertinent state and intentions obvious." (Bradshaw et al., 2017, p. 14). This characteristic was transformed into two scale items: "I had the feeling that the agent could communicate its status to me." and "I had the feeling that the agent made clear what it wanted to achieve.". This approach was repeated for all characteristics. Items were iteratively refined in discussions among the authors to ensure content validity and item clarity. The pre-validation PCS comprised 15 items. Item responses are provided on a 6-point Likert scale from 1 (*completely*

disagree) to 6 (*completely agree*). The full scales can be found in Table 1 (agent version) and Table 2 (human partner version).

Table 1

Perceived Cooperativity Scale (Agent Version, PCS-A)

Item #	English version	German version
	I had the feeling that ...	Ich hatte das Gefühl, dass ...
PCS01	... the agent could communicate its status to me.	... der Agent mir seinen Status mitteilen konnte.
PCS02	... the agent made clear what it wanted to achieve.	... der Agent klarmachte, was er erreichen wollte.
PCS03	... the agent kept me informed about its actions.	... der Agent mich über sein Vorgehen auf dem Laufenden gehalten hat.
PCS04	... the agent could detect difficulties on the way to achieving the goal.	... der Agent mögliche Schwierigkeiten bei der Zielerreichung erkennen konnte.
PCS05	... the agent knew enough of my goals to collaborate with me.	... der Agent genug über meine Ziele wusste, um mit mir zusammenzuarbeiten.
PCS06	... the agent could adapt its communication to a specific situation.	... der Agent seine Kommunikation an eine konkrete Situation anpassen konnte.
PCS07	... the agent was able to take or hand over the initiative by itself.	...der Agent von sich aus die Initiative ergreifen oder abgeben konnte.
PCS08	... the agent was able to react to external guidance.	... der Agent auf äußere Vorgaben reagieren konnte.
PCS09	... it was predictable how the agent would react in different situations.	... es vorhersehbar war, wie der Agent sich in unterschiedlichen Situationen verhalten würde.
PCS10	... the agent was dependable.	... der Agent verlässlich war.
PCS11	... the way in which the agent proceeded could be influenced by me.	... die Art und Weise wie der Agent vorging von mir beeinflusst werden konnte.
PCS12	... the agent was able to incorporate information I shared.	... der Agent von mir geteilte Informationen berücksichtigen konnte.
PCS13	... the agent was able to indicate the most important information.	... der Agent auf die wichtigsten Informationen hinweisen konnte.
PCS14[1]	... the agent was able to realize that we had different approaches to achieving the goal.	... der Agent erkennen konnte, dass wir unterschiedliche Ansätze zur Zielerreichung hatten.
PCS15	... the agent tried to solve problems that arose during our cooperation.	... der Agent versuchte, in unserer Zusammenarbeit auftretende Probleme zu lösen.

Note. Item responses are provided on a 6-point Likert scale from 1 (*completely disagree*) to 6 (*completely agree*). [1] Item excluded from final scale.

Table 2

Perceived Cooperativity Scale (Human Partner Version, PCS-H)

Item #	English version	German version
	I had the feeling that ...	Ich hatte das Gefühl, dass ...
PCS01	... the person could communicate their status to me.	... die Person mir ihren Status mitteilen konnte.
PCS02	... the person made clear what they wanted to achieve.	... die Person klarmachte, was sie erreichen wollte.
PCS03	... the person kept me informed about their actions.	... die Person mich über ihr Vorgehen auf dem Laufenden gehalten hat.
PCS04	... the person could detect difficulties on the way to achieving the goal.	... die Person mögliche Schwierigkeiten bei der Zielerreichung erkennen konnte.
PCS05	... the person knew enough of my goals to collaborate with me.	... die Person genug über meine Ziele wusste, um mit mir zusammenzuarbeiten.
PCS06	... the person could adapt their communication to a specific situation.	... die Person ihre Kommunikation an eine bestimmte Situation anpassen konnte.
PCS07	... the person was able to take or hand over the initiative on their own.	... die Person von sich aus die Initiative ergreifen oder abgeben konnte.
PCS08	... the person was able to react to external guidance.	... die Person auf äußere Vorgaben reagieren konnte.
PCS09	... it was predictable how the person would react in different situations.	... es vorhersehbar war, wie die Person sich in unterschiedlichen Situationen verhalten würde.
PCS10	... the person was dependable.	... die Person verlässlich war.
PCS11	... the way in which the person proceeded could be influenced by me.	... die Art und Weise wie die Person vorging von mir beeinflusst werden konnte.
PCS12	... the person was able to incorporate information I shared.	... die Person von mir geteilte Informationen berücksichtigen konnte.
PCS13	... the person was able to indicate the most important information.	... die Person auf die wichtigsten Informationen hinweisen konnte.
PCS14[1]	... the person was able to realize that we had different approaches to achieving the goal.	... die Person erkennen konnte, dass wir unterschiedliche Ansätze zur Zielerreichung hatten.
PCS15	... the person tried to solve problems that arose during our cooperation.	... die Person versuchte, in unserer Zusammenarbeit auftretende Probleme zu lösen.

Note. Item responses are provided on a 6-point Likert scale from 1 (*completely disagree*) to 6 (*completely agree*). [1] Item excluded from final scale.

For the TPS, item generation followed directly from the theoretical conceptualization of teaming perception as a subjective state variable emerging from the interplay of perceived AI prosocial contributions and mutual human–AI support (Bratman, 1992; Pacherie, 2014;

Tomasello et al., 2012). The three facets underlying this conceptualization (i.e., the emergent sense of the team as a whole, one's own contribution and support, and the human's perception of the system's contribution and support) were each transferred into a set of 3 items (i.e., 9 items in total). As with the PCS, items were phrased in accessible, everyday language to ensure comprehensibility. Unlike the PCS, however, no additional item stem was needed, as TPS items are inherently formulated as first-person experiential judgments. Rather than targeting specific, potentially non-observable AI characteristics, they capture the respondent's holistic subjective experience of the interaction as a cooperative episode. Items were iteratively refined in discussions among the authors to ensure content validity and item clarity. The pre-validation TPS comprised 9 items. Item responses are provided on a 6-point Likert scale from 1 (*completely disagree*) to 6 (*completely agree*). The full scales can be found in Table 3 (agent version) and Table 4 (human partner version).

Table 3

Teaming Perception Scale (Agent Version, TPS-A)

Item #	English version	German version
TPS01	The agent and I were a team.	Der Agent und ich waren ein Team.
TPS02	The agent and I worked together towards a goal.	Der Agent und ich haben gemeinsam auf ein Ziel hingearbeitet.
TPS03	The agent and I supported each other.	Der Agent und ich haben uns gegenseitig unterstützt.
TPS04	I contributed to achieving the common goal.	Ich habe zum Erreichen des gemeinsamen Ziels beigetragen.
TPS05	I put aside my own goals in favor of the agent.	Ich habe meine eigenen Ziele zugunsten des Agenten zurückgestellt.
TPS06	I supported the agent.	Ich habe den Agenten unterstützt.
TPS07	The agent contributed to achieving the common goal.	Der Agent hat zum Erreichen des gemeinsamen Ziels beigetragen.
TPS08	The agent put aside its goals in favor of me.	Der Agent hat seine eigenen Ziele zu meinen Gunsten zurückgestellt.
TPS09	The agent supported me.	Der Agent hat mich unterstützt.

Note. Item responses are provided on a 6-point Likert scale from 1 (*completely disagree*) to 6 (*completely agree*).

Table 4

Teaming Perception Scale (Human Partner Version, TPS-H)

Item #	English version	German version
TPS01	The person and I were a team.	Die Person und ich waren ein Team.
TPS02	The person and I worked together towards a goal.	Die Person und ich haben gemeinsam auf ein Ziel hingearbeitet.
TPS03	The person and I supported each other.	Die Person und ich haben uns gegenseitig unterstützt.
TPS04	I contributed to achieving the common goal.	Ich habe zum Erreichen des gemeinsamen Ziels beigetragen.
TPS05	I put aside my own goals in favor of the person.	Ich habe meine eigenen Ziele zugunsten der Person zurückgestellt.
TPS06	I supported the person.	Ich habe die Person unterstützt.
TPS07	The person contributed to achieving the common goal.	Die Person hat zum Erreichen des gemeinsamen Ziels beigetragen.
TPS08	The person put aside their goals in favor of me.	Die Person hat ihre eigenen Ziele zu meinen Gunsten zurückgestellt.
TPS09	The person supported me.	Die Person hat mich unterstützt.

Note. Item responses are provided on a 6-point Likert scale from 1 (*completely disagree*) to 6 (*completely agree*).

3.2 Translations

The translation of the German scales into English was conducted with the help of three native bilingual speakers. First, an initial translation was developed by the authors. Second, two bilingual speakers developed translations of the scales independently. Third, the first author discussed the translations with the other authors, resulting in one harmonized version of the scale translations with small adaptations. Fourth, the translated scales were given to the third bilingual speaker for blind translation back into German. Fifth, this back translation was checked against the original German versions by the first author. Seven instances were identified where the original scales and back translations across both scales were not in perfect accordance. These discrepancies were discussed by the first author and the third bilingual speaker. The discussion resulted in slight adaptations of the English translations, yielding a version in full accordance with the original German scales.

3.3 Scale evaluation procedure

The resulting scales were administered to diverse samples of users participating in three independent studies on human-AI and human-automation collaboration. The results of participants' self reports were then used to assess item properties (item difficulty, discrimination and Cronbach's α if removed), scale dimensionality (exploratory and confirmatory factor analysis), reliability (Cronbach's α and McDonald's ω), and construct validity (discriminant validity, ability to differentiate between systems). This approach is in line with the best practices of scale evaluation (Boateng et al., 2018).

3.3.1 Validity evaluation of the perceived cooperativity scale

For discriminant validity of the PCS, we selected measures that tap conceptually adjacent but distinct aspects of human–AI interaction. While we expect positive correlations with all variables, differences in effect sizes are expected to reflect the varying degrees of conceptual overlap with perceived cooperativity.

Particularly strong correlations are expected for subjective information processing awareness (SIPA; Schrills and Franke, 2023), which captures the degree to which a system is perceived as transparent, understandable, and predictable. Particularly items PCS01, PCS02, and PCS03 share conceptual overlap with the SIPA transparency subscale, while item PCS09 mirrors the SIPA subscale predictability. However, while SIPA focuses more narrowly on the cognitive, information-processing dimension of system perception, the PCS adopts a broader approach that additionally encompasses the agent's responsiveness to user input and its active management of the joint activity. Therefore, a perfect overlap between the two measures is not expected. Moreover, substantial correlations are expected for perceived competence (Leach et al., 2007) and system trustworthiness (Franke et al., 2015), as both reflect core aspects of an agent's capability and reliability that are central to cooperative potential.

By contrast, weaker positive correlations are expected for measures that are conceptually more distal from perceived cooperativity. System acceptance (Van Der Laan et al., 1997), although potentially influenced by cooperative quality, reflects a broader evaluative attitude toward the system as a whole rather than the episodic perception of cooperative capability and practice, and is therefore expected to show a lower correlation with

the PCS. Similarly, perceived usefulness (Attig, Winzer, et al., 2026) is included as a further indicator, capturing the broader instrumental dimension of the agent's actions. Finally, perceived warmth (Leach et al., 2007) captures the interpersonal and relational quality of the interaction—a dimension that is more closely tied to the experiential sense of teaming than to the assessment of cooperative capability that defines perceived cooperativity—and is therefore expected to show a weaker correlation with the PCS than the capability- and reliability-focused measures.

For evaluating the scale's ability to differentiate between cooperation partners with high and low perceived cooperativity, we compare scale mean scores for cooperation partners in a cooperative game setting that differ with respect to partner type (human vs. rule-based system vs. reinforcement learning based-system; see Study 1 and Attig et al., 2024). These three variations differ regarding human-likeness, transparency, and the degree to which they embody the characteristics of an effective cooperative agent as defined by joint activity theory as the basis for item formulation. Specifically, the human partner is expected to score highest on perceived cooperativity, as human partners naturally possess the full range of cooperative capabilities. The rule-based system is expected to score higher than the reinforcement learning-based system, because its behavioral rules were derived from human cooperative game strategies, making its actions more transparent, predictable, and comprehensible to the human partner. The reinforcement learning-based system, by contrast, developed cooperative game strategies autonomously through experience (Canaan et al., 2020), resulting in game styles that are not as easily interpretable by human partners, thereby limiting the human's ability to form accurate expectations and assess the agent's cooperative potential. If the PCS successfully differentiates between cooperation partners of varying cooperative quality, mean scores should reflect this ordering, with human partners rated highest, followed by the rule-based system, and the reinforcement learning-based system rated lowest.

3.3.2 Validity evaluation of the teaming perception scale

For discriminant validity of the TPS, we assessed perceived warmth and competence (Leach et al., 2007). Perceived warmth captures the interpersonal and relational quality of the interaction and is expected to align with the prosocial and mutual support facets of teaming

perception. Perceived competence is included because teaming perception requires that both partners are perceived as contributors whose efforts meaningfully advance the shared goal. An agent perceived as incapable is less likely to be experienced as a genuine teammate regardless of relational quality.

By contrast, weaker positive correlations are expected for measures that are conceptually more distal from teaming perception. SIPA's narrower focus on cognitive transparency and predictability (Schrills & Franke, 2023) makes it conceptually distant from the relational, experiential quality of teaming perception: The cooperative engagement that teaming requires extends well beyond what information processing awareness can capture. Moreover, system usefulness reflects the broader instrumental dimension of the agent's actions (Attig, Winzer, et al., 2026), capturing whether its outputs are practically valuable and facilitate task performance rather than the reciprocal and goal-directed experience of genuinely acting as a team, and is therefore also expected to show a weaker correlation with the TPS. Importantly, we expect stronger correlations with the partner-focused and team-focused facets of the TPS than with the self-focused facet, as the human partner's own cooperative investment and support reflect primarily their own motivation and goal prioritization rather than their perception of the AI's qualities.

For evaluating the scale's ability to differentiate between cooperation partners that vary in their potential for teaming perception, we compare scale mean scores across the same three partner types in a cooperative game setting (human vs. rule-based system vs. reinforcement learning-based system; see Study 1 and Attig et al., 2024). The expected ordering mirrors that of the PCS: The human partner is expected to score highest, as human partners are most naturally suited to engage in genuinely reciprocal cooperation and mutual support. The rule-based system is expected to score higher than the reinforcement learning-based system, as its human-derived behavioral strategies enable the human partner to develop more stable expectations about its game style and should therefore enable a smoother user experience and a sense of joint agency (Pacherie, 2014). In contrast, the reinforcement learning-based system, whose less interpretable game style limits the formation of such expectations, is expected to score lowest.

3.4 Samples and study designs

To enable analyses, we utilized data from three studies with various human-AI and human-automation applications, totaling $N = 409$ participants. Details on the studies—labeled by the application examined—are reported below. Table 5 shows key descriptive sample statistics. Beyond demonstrating the heterogeneity of use cases for the scales, the samples also differed greatly in age, gender, and affinity for technology interaction. Regarding the latter, we applied the Affinity for Technology Interaction (ATI) scale (Franke et al., 2019), which evaluates a person's self-reported tendency to engage actively in technology interaction and is rated on a response scale from 1 to 6. Samples are compared with a population distribution sample ($M = 3.61$, $SD = 1.08$, $N = 232$; Franke et al., 2019).

Table 5

Sample Demographics and Application Contexts

Study	*N*	Population	Application	Samples	Age *M (SD)*	ATI *M (SD)*	Gender (M:F:D)
1: Hanabi	94	Students	Cooperative card game	Han-HH Han-RB Han-RL	23.0 (4.8)	3.4 (1.0)	19:74:1
2: LLM	182	Prolific	LLM-based chatbot	Augmented Focused Conflicted Failed	38.5 (13.2)	4.0 (0.8)	91:89:1
3: LDC	133	EV Drivers	EV trip planner	LDC	46.2 (10.8)	4.9 (0.9)	127:6:0

Note. One participant in the LLM sample (Study 2) did not disclose their gender and is therefore not included in the gender ratio (M:F:D), though included in *N*.

3.4.1 Study 1: Playing the cooperative card game Hanabi

Study 1 used a within-subjects design to examine human-AI teaming using the cooperative card game *Hanabi*, played in dyads of human participants or dyads consisting of a human and one of two artificial agents. Hanabi is a cooperative game in which players jointly build sequences of correctly ordered cards without seeing their own cards, relying solely on

limited hints from their partners (Attig et al., 2024). Hanabi represents a particularly demanding challenge for AI because it inherently requires collaborative play and reasoning about a partner's mental state, managed through restricted communicative actions (Bard et al., 2020). This makes it not only a benchmark for AI development but also a prototypical setting for investigating human-AI teaming, as it fulfills the two requirements considered necessary to establish a joint activity (Bradshaw et al., 2017; Klein et al., 2005): When entering the game, both players implicitly agree to collaborate towards a shared goal (i.e., scoring a maximum of points) through interdependent actions (i.e., game moves that are interlinked between players). These two requirements are inherent and non-negotiable characteristics of the game; thus, Hanabi can only be played cooperatively (Bard et al., 2020). Nevertheless, it allows for meaningful variation in perceived cooperativity and teaming perception, as partners can adopt different playing styles (e.g., by playing more indirectly by giving hints that reduce uncertainty, or taking a more active, direct role in achieving success by playing cards) creating the conditions necessary to observe and measure differences in cooperative quality across partner types.

Playing Hanabi with artificial agents provides coverage of both dimensions of human-AI integration: the highly interdependent nature of in-game actions represents a high degree of synchronic integration, while the repeated interactions across two games per partner introduce at least a limited diachronic dimension (compared to naturalistic long-term partnerships with varying tasks over time). Nevertheless, it allows for the beginning of accumulated cooperative experience to emerge within each partner condition.

Study 1 was conducted using a web-based interface for the game application (Honda Research Institute EU and synyx, 2023). Communication was restricted to predefined in-game hints to maintain standardization. Participants (N = 94) played six rounds of Hanabi in total: two rounds each with another human player and with each of two artificial agents. The artificial agents were (1) the rule-based agent "Piers" (in the following referred to as RB, for rule-based), which was derived from human game play (Walton-Rivers et al., 2017), and (2) the reinforcement learning agent "Rainbow" (in the following referred to as RL, for reinforcement learning), which learned to play Hanabi by playing with the RB agent (Canaan

et al., 2020). RL agents that are trained with single game partners typically perform well when paired with their training partners, but this performance does not transfer to other game partners (i.e., fewer correct moves, more mistakes, higher rate of lost games; Canaan et al., 2020). Prior research has demonstrated that humans evaluate Hanabi agents that display dominated actions (i.e., moves that lead to loss of playable cards or life tokens) more negatively (Siu et al., 2025). Taken together, we expect the RL agent to display less transparent and less predictable behavior than the RB agent, and therefore to elicit lower ratings of perceived cooperativity and teaming perception.

Data collection was split into two sessions: one session in which participants played two games with a human partner, and one session in which they played two games each with both artificial agents. The order of sessions and the order of artificial agents within the agent session were partially counterbalanced across participants, with the constraint that both artificial agent interactions were completed within the same session. This resulted in four possible orderings: (1) human-human first followed by the RB then RL agent, (2) human-human first followed by the RL then RB agent, (3) RB then RL agent followed by human-human, and (4) RL then RB agent followed by human-human. For the human-human condition, the same pairing was maintained across both Han-HH rounds.

Of the 94 participants, 76 were paired with another study participant in the Han-HH session, resulting in 38 participant-participant pairs. The remaining 18 participants played with a member of the research team standing in as their human partner, due to participant no-shows.

After each interaction (i.e., two rounds of Hanabi with one partner) participants completed the PCS and TPS alongside additional scales assessing construct validity. As Study 1 was the first study in which the current scale versions were tested, we additionally asked participants to rate their confidence in responding to the single PCS and TPS items on a scale from 1 (*not confident at all*) to 10 (*extremely confident*) to assess the item fit to the type of interaction. Partner type (human vs. artificial agent) was disclosed to participants after completing the two games but before the assessment of subjective variables, ensuring that knowledge of partner type would not influence participants' game experience. Partner type

(human vs. artificial agent) was disclosed to participants after completing the two games but before the assessment of subjective variables, ensuring that knowledge of partner type would not influence participants' game experience. This disclosure was necessary to administer and psychometrically assess the agent- or human-referencing scale versions (i.e., PCS-A/TPS-A vs. PCS-H/TPS-H) and because participants' conscious awareness of partner type was itself of theoretical interest, allowing us to assess how a partner known to be human versus artificial is perceived in terms of cooperativity and teaming. Conditions are referred to throughout this paper by abbreviated labels (i.e., Han-HH for human partner, Han-RB for rule-based artificial agent, Han-RL for the reinforcement learning agent).

Participants consisted of undergraduate and graduate students who were offered course credit and that were recruited via email, student forums, and personal invitations in lectures and seminars. The sample's ATI differed significantly from the general population distribution ($t(93) = -2.55$, $p = .012$, $d = 0.26$, small effect size). The large majority of participants (90%) never played Hanabi before.

Ethical approval was obtained from the Ethics Committees of University of Luebeck (tracking number: 2023-570, July 17, 2023) and the Honda R&D Bioethics Committee (tracking number: 100HM-015H, July 21, 2023). The study was conducted in accordance with the Declaration of Helsinki; all participants provided written informed consent prior to taking part. Study 1 builds on a pilot study reported in Attig et al. (2024), extending it with a larger sample and the refined scale versions.

3.4.2 Study 2: Large-language model interaction

Study 2 used a within-subjects design to examine how AI autonomy and goal alignment influence user experience in the context of LLM use (Attig, Winzer, et al., 2026). Compared to Hanabi in Study 1, this setting represents a lower degree of synchronic human-AI integration, as the fact-checking task could in principle be completed through a single exchange rather than requiring tightly interdependent actions, while repeated trials per chatbot version introduce a limited diachronic integration.

Participants ($N = 182$) fact-checked encyclopedic articles on natural and sociocultural phenomena using four versions of an LLM-based chatbot, implemented via GPT-4o Mini and

iteratively refined to ensure alignment with experimental conditions. The chatbot versions were realized through instruction sets (i.e., system prompts; for details see Attig, Winzer, et al., 2026) and differed systematically in their degree of autonomy and alignment with the user's goal of correcting factual errors: *Focused* (supported the user goal without additional actions), *Augmented* (supported the user goal and autonomously provided sources for corrections), *Conflicted* (supported the user goal but also autonomously simplified the text, which conflicted with the user goal), and *Failed* (did not support the user goal but implied capability of correcting spelling errors). This systematic variation in cooperative quality—from a fully aligned and capable partner to one that fails to contribute meaningfully to the shared goal—provides the conditions necessary to examine whether the PCS and TPS are sensitive to differences in perceived cooperativity across interaction contexts.

Participants engaged with each chatbot across twelve fact-checking trials (three trials per chatbot version), and the PCS, TPS, and additional construct validity measures were assessed after each condition. Participants were recruited via Prolific (Prolific, 2024) and financially compensated according to Prolific's recommendations for ethical payment. The sample's ATI differed significantly from the general population ($t(181) = 6.92$, $p < .001$, $d = 0.51$, medium effect size). Ethical approval was obtained from the Ethics Committees of University of Luebeck (tracking number: 2024-452; August 28, 2024) and the Honda R&D Bioethics Committee (tracking number: 101HM-028H; August 23, 2024). The study was conducted in accordance with the Declaration of Helsinki; all participants provided written informed consent prior to taking part. Conditions are referred to throughout this paper by their respective labels (e.g., Focused indicates data from the Focused chatbot condition). As the present work focuses on scale development and validation, a detailed report of the experimental results (including ability to differentiate between cooperation partners with high and low perceived cooperativity and teaming perception) is provided elsewhere (Attig, Winzer, et al., 2026).

3.4.3 Study 3: Long-distance charging and trip planner interaction

Study 3 used a correlational design to examine perceived cooperativity in the context of real-world interaction with an algorithmic decision-support system, focusing on user

interaction with electric vehicle (EV) trip and charging stop planning systems (Stattkus-Fortange et al., 2026). Study 3 targeted users of the Tesla trip planner and Tesla Supercharger network. This setting differs fundamentally from the prior two studies in the nature of the human-system interaction it involves. Rather than a cooperative AI agent that actively contributes to a shared goal through interdependent actions, the Tesla trip planner is best characterized as a decision-support system: It processes inputs and returns optimized routing and charging recommendations based on real-time data, but exercises no salient agency (https://www.tesla.com/features/navigation). The TPS was not assessed in Study 3 because the system's lack of agency means that the conditions necessary for teaming perception to emerge (i.e., mutual contribution, reciprocal support, and the prioritization of a shared goal by both partners) are not sufficiently met in this interaction context (Lyons et al., 2021). This makes Study 3 a deliberate boundary case for the application of the PCS, allowing us to examine whether the scale remains meaningful in contexts where the interaction partner is an algorithmic tool rather than a cooperative agent.

Participants (N = 133) were recruited at Tesla Superchargers in northern Germany as well as through an online forum on electric mobility. Participants were compensated with EUR 10,00 for their participation. The sample's ATI was significantly higher than the general population distribution ($t(132) = 16.44$, $p < .001$, $d = 1.43$, large effect size). Ethical approval was obtained from the ethics committee of the University of Luebeck (tracking number 2024-462; August 26, 2024); the Honda R&D Bioethics Committee determined that ethical approval was not applicable for this study. The study was conducted in accordance with the Declaration of Helsinki; all participants provided written informed consent prior to taking part.

3.5 Construct validity measures

If not otherwise stated, participants responded to measures on a 6-point Likert scale from 1 (*completely disagree*) to 6 (*completely agree*). As reliability indicators, we report Cronbach's α as a range across the conditions for each sample, which is interpreted according to common practice (e.g., Cripps, 2017) as poor ($0.5 \leq \alpha < 0.6$), questionable ($0.6 \leq \alpha < 0.7$), acceptable ($0.7 \leq \alpha < 0.8$), good ($0.8 \leq \alpha < 0.9$), or excellent ($\alpha \geq 0.9$).

3.5.1 Subjective information processing awareness

To determine to what extent a system was perceived as transparent, understandable and predictable, the 6-item Subjective Information Processing Awareness Scale (SIPA, Schrills and Franke, 2023) was used. Internal consistency was good to excellent (α = .85 - .92). It was assessed in Hanabi RB/RL and all LLM samples.

3.5.2 Perceived system trustworthiness

We assessed the perceived trustworthiness of the chatbot with the 5-item Facets of System Trustworthiness scale (FOST, Franke et al., 2015). Internal consistency was acceptable to excellent (α = .79 - .90). It was assessed in the LLM and LDC samples.

3.5.3 User acceptance

The 9-item System Acceptance Scale Van Der Laan et al., 1997 assesses system acceptance along two dimensions: usefulness, which reflects the perceived effectiveness and practical benefits of the system, and affective satisfaction, which captures the user's emotional response to system interaction. Participants rated each item on a bipolar 5-point response scale, where each point represents opposing adjective pairs (e.g., "useless – useful", "annoying – pleasant"). Internal consistency was good (α = .80 - .81). It was assessed in the LDC sample.

3.5.4 System usefulness

For assessing usefulness (Attig, Winzer, et al., 2026), we used four self-constructed items ("The outputs of the chatbot are not meaningful.", "My decisions are facilitated when the system displays information.", "I can use the chatbot's output for my decisions.", "I cannot assess what role the chatbot's output plays in the task."). Internal consistency was questionable to acceptable (α = .67 - .72). It was assessed in the LLM samples.

3.5.5 Perceived warmth and competence

Leach et al. (2007) developed a scale to assess social perception in human-human interaction with 9 items on a 7-point response scale (from 1 – *strongly disagree* to 7 – *strongly agree*). We utilized the subscales perceived warmth and competence (consisting of three items each) for assessing social perception of both human and artificial cooperation partners.

Internal consistency was good to excellent (α = .80 - .97). Social perception was assessed in all Hanabi and all LLM samples.

4. Results

4.1 Perceived Cooperativity Scale

4.1.1 Dimensionality

Perceived cooperativity characterizes the perception of agents being equipped with requirements for cooperative teamwork according to joint activity theory (Bradshaw et al., 2017; Klein et al., 2005). The items were developed to reflect a single underlying construct of perceived cooperativity. To examine the dimensionality of the PCS empirically, EFAs were conducted on the Hanabi samples of Study 1. First, parallel analysis using principal axis factoring (PAF) with 100 Monte Carlo simulations was employed to determine the number of factors. Results indicated a one-factor solution, further supported by the Kaiser criterion, which confirmed that only the first factor had an eigenvalue exceeding 1. To inspect item-level factor structure, a subsequent PAF with Promax rotation was conducted. Factor loadings were generally strong, ranging from .44 to .87. An exception was PCS14, which showed weaker loadings in the Han-RB (.28) and Han-HH (.32) conditions while performing adequately in the Han-RL condition (.60), suggesting some inconsistency in how this item functions across contexts. The single factor explained between 47 % (Han-RL) and 55 % (Han-HH) of the variance.

Confirmatory factor analyses (CFA) assessed model fit across all eight samples of Studies 1, 2, and 3 (Table 6), comparing a 15-item one-factor model to a 14-item one-factor model (excluding PCS14, also based on item analysis results). Model fit was evaluated using Comparative Fit Index (CFI), Tucker–Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA), Standardized Root Mean Square Residual (SRMR), Akaike information criterion (AIC), and Bayesian information criterion (BIC). Following Browne and Cudeck (1992), RMSEA values above .10 were considered indicative of unacceptable fit. For CFI and TLI, values of .90 and above were treated as reflecting acceptable fit (Bentler & Bonett, 1980), and SRMR values below .08 were considered acceptable (Hu & Bentler, 1999). Consistent with Marsh et al. (2004), these thresholds were applied as approximate guidelines rather than fixed decision rules, with fit indices interpreted jointly and in light of the broader pattern of evidence. Moreover, recommended cut-offs are intended for the evaluation of a

single model in isolation rather than the comparative fit of nested models (Hu & Bentler, 1999).

Comparing the two models, the 14-item solution generally showed improved fit indices across most samples. Significant χ^2 differences in Han-RL, LDC, Focused, Augmented, and Conflicted indicated that PCS14 negatively impacted model fit, while no meaningful difference was found for the remaining samples. Across both models, CFI met the .90 threshold in most samples, with the notable exception of the Augmented condition, where fit was poor across all indices. TLI values indicated poor fit of the 15-item model across most samples, and a rather inconclusive pattern for the 14-item model. RMSEA values were almost consistently within the acceptable range across models, while SRMR indices all were at least acceptable. This pattern was further supported by AIC and BIC values, which are information-theoretic indices that allow for direct model comparison irrespective of sample size, with lower values indicating better model fit relative to model complexity. AIC and BIC consistently favored the 14-item solution. Taken together, and interpreted as approximate guidelines rather than fixed decision rules (Marsh et al., 2004), the CFA results are consistent with a unidimensional factor structure across most contexts and an improved fit of the 14-item model. However, differences across samples suggest that the empirical factor structure of the PCS might be to some degree context- and interaction-dependent.

4.1.2 Item analysis

Item analysis (i.e., item discrimination, item difficulty, α if item removed, and confidence rating z-scores) was conducted on the Hanabi samples to examine scale performance on the single item level (Table 7). Item discrimination coefficients (i.e., part-whole corrected item-total correlations) depict how well a single item differentiates between interaction ratings in the same direction as the overall scale. Item discrimination ranged on average across samples between $\bar{r}_{itc}$ = .39 (PCS14) to .81 (PCS15), indicating good item discrimination (Moosbrugger & Kelava, 2020), with the exception of item PCS14 in the Han-HH (r_{itc} = .32) and Han-RB samples (r_{itc} = .28). Item difficulty indices reflect the average degree of agreement across the response scale, ranging from 0 to 100 %, with higher values indicating that participants more readily agreed with the item on average. Item

Table 6

Comparison of 14-item and 15-item CFA models

	14-item model						15-item model						Comparison	
Sample	CFI	TLI	RMSEA	SRMR	AIC	BIC	CFI	TLI	RMSEA	SRMR	AIC	BIC	$\Delta\chi^2(13)$	p
Han-HH	.93	.92	0.09	0.05	3516.02	3587.23	.93	.92	0.09	0.05	3827.42	3903.72	15.52	.276
Han-RB	.91	.89[2]	0.10	0.06	3638.23	3709.44	.90	.88[2]	0.10	0.06	3915.54	3991.84	19.17	.118
Han-RL	.93	.92	0.08	0.07	3534.73	3605.94	.91	.89[2]	0.09	0.07	3779.70	3856.00	29.27	.006
LDC	.89[1]	.87[2]	0.09	0.07	3605.94	5525.34	.88[1]	.86[2]	0.09	0.07	5844.91	5931.62	22.91	.043
Focused	.92	.91	0.09	0.05	6753.30	6843.01	.91	.89[2]	0.10	0.05	7303.84	7399.96	32.44	.002
Augmented	.80[1]	.77[2]	0.12[3]	0.08	6551.10	6640.82	.77[1]	.74[2]	0.13[3]	0.08	7127.25	7223.37	46.72	<.001
Conflicted	.91	.90	0.10	0.05	6380.18	6469.90	.90	.88[2]	0.10	0.05	6922.49	7018.61	38.91	<.001
Failed	.95	.94	0.10	0.04	7170.03	7259.74	.95	.94	0.09	0.04	7625.39	7721.51	15.76	.262

Note. [1] CFI < .90, indicating poor fit. [2] TLI < .90, indicating poor fit. [3] RMSEA > .10, indicating unacceptable fit.

difficulty values ranged on average between 40.0 % (PCS14) to 56.1 % (PCS12), indicating adequate item difficulty (Moosbrugger & Kelava, 2020). Results further revealed systematic differences between samples: The average difficulty value in the Han-RL condition (28.1 %) was considerably lower than in the Han-HH (61.5 %) and Han-RB (54.9 %) conditions, suggesting that participants agreed the PCS items more readily when interacting with humans and the rule-based agent than with the reinforcement learning agent. This lower agreement in the Han-RL condition is also reflected in the item mean values (Table 7) and is further analyzed in Section 4.1.4. Cronbach's α if item removed showed that all items, except PSC14, contributed positively to scale reliability.

Mean confidence ratings ranged from $M = 6.1$ to $M = 7.8$ (overall $M = 7.0$), with standard deviations between $SD = 1.7$ and $SD = 2.7$ ($M_{SD} = 2.1$). To identify items with relatively low confidence, item-level means and standard deviations were averaged across Hanabi samples and standardized using z-scores. Items were flagged if their mean z-score was more than 2 standard deviations below the overall mean or if their standard deviation z-score exceeded +1 standard deviation. This approach accounts for sample-specific response tendencies. Two items were flagged: PCS03 (high variability) and PCS14 (high variability and below-average confidence).

Taking CFA and item analysis results into account, we decided to exclude item PCS14 from the scale. Thus, all further analyses were performed on the 14-item version of the PCS.

Table 7

PCS item analysis metrics (based on Hanabi samples of Study 1)

	Value *M*(*SD*)			Item discrimination (r_{itc})			Item difficulty			α if item removed			Confidence *z*	
Item	Han-HH	Han-RB	Han-RL	Han-HH	Han-RB	Han-RL	Han-HH	Han-RB	Han-RL	Han-HH	Han-RB	Han-RL	*M*	*SD*
PCS01	3.6 (1.3)	3.4 (1.2)	2.4 (1.0)	.73	.62	.65	52.7	48.9	27.5	.94	.93	.92	-1.0	0.4
PCS02	3.8 (1.2)	3.7 (1.2)	2.3 (1.1)	.80	.73	.61	56.6	53.6	25.5	.94	.93	.92	0.0	-0.3
PCS03	3.6 (1.2)	3.3 (1.2)	2.1 (1.1)	.80	.72	.68	52.8	46.4	22.6	.94	.93	.92	-0.8	1.1[5]
PCS04	4.2 (1.2)	3.4 (1.2)	2.2 (1.1)	.74	.70	.78	64.0	48.3	24.3	.94	.93	.92	-0.1	-0.5
PCS05	4.2 (1.3)	3.8 (1.3)	2.5 (1.3)	.78	.66	.65	64.5	55.5	30.6	.94	.93	.92	0.0	0.3
PCS06	3.9 (1.3)	3.7 (1.3)	2.3 (1.1)	.73	.73	.72	58.7	54.5	26.4	.94	.93	.92	-0.7	0.1
PCS07	4.4 (1.2)	3.9 (1.2)	2.9 (1.2)	.73	.52	.62	68.5	57.9	37.2	.94	.93	.92	-0.2	0.1
PCS08	4.4 (1.3)	4.0 (1.2)	2.7 (1.2)	.73	.72	.71	67.7	60.2	33.4	.94	.93	.92	0.2	0.2
PCS09	3.7 (1.3)	3.7 (1.3)	2.3 (1.3)	.64	.71	.45	53.6	53.6	26.6	.94	.93	.93	0.5	-1.1
PCS10	4.4 (1.2)	4.0 (1.4)	2.1 (1.1)	.75	.75	.71	67.5	59.8	21.7	.94	.93	.92	0.8	-0.1
PCS11	4.4 (1.2)	3.9 (1.2)	2.8 (1.3)	.57	.54	.53	67.0	58.9	36.8	.94	.93	.92	0.3	-0.6
PCS12	4.6 (1.3)	4.3 (1.2)	2.6 (1.3)	.82	.70	.61	71.7	65.5	31.1	.94	.93	.92	2.1	-2.0
PCS13	4.2 (1.3)	3.9 (1.3)	2.2 (1.2)	.71	.76	.70	63.2	58.9	24.0	.94	.93	.92	0.9	-0.3
PCS14	3.4 (1.3)	3.1 (1.1)	2.5 (1.1)	.32[1]	.28[1]	.56	48.1	41.9	30.0	.95[2]	.94[2]	.92	-2.4 [4]	2.6[5]
PCS15	4.3 (1.4)	3.9 (1.4)	2.2 (1.1)	.80	.83	.80	66.4	58.7	24.5	.94	.92	.92	0.4	0.0
Scale	4.1 (1.0)	3.7 (0.9)	2.4 (0.8)	.71	.66	.65	61.5	54.9	28.1	.94[3]	.93[3]	.92[3]		

Note. [1] Item discrimination $< .4$. [2] α if item removed $>$ overall α. [3] Scale Cronbach's α. [4] Below-average z-standardized mean confidence ratings ($z < -2$ *SD*). [5] High variability of z-standardized *SD* of confidence ratings ($z > 1$ *SD*).

4.1.3 Reliability

Reliability analysis of the 14-item PCS was conducted on all samples (see Table 8). The PCS yielded on average internal consistency values of Cronbach's α = .93 and McDonalds ω = .93. According to common guidelines for interpretation (e.g., Cripps, 2017), these values indicate excellent internal consistency, suggesting that the PCS is a highly reliable measure across diverse contexts of human-AI cooperation.

Table 8

Reliability for PCS and TPS (composite and subscales) across samples

	PCS		TPS							
			Composite scale		Team		Self		Partner	
Sample	α	ω	α	ω	α	ω	α	ω	α	ω
Han-HH	.95	.95	.86	.88	.89	.90	.40	.61	.69	.73
Han-RB	.94	.94	.84	.85	.88	.88	.50	.57	.78	.79
Han-RL	.92	.93	.81	.84	.90	.91	.37	.42	.80	.81
Focused	.94	.94	.92	.92	.92	.92	.67	.73	.79	.83
Augmented	.90	.90	.88	.90	.91	.91	.57	.67	.63	.74
Conflicted	.94	.94	.90	.91	.92	.92	.58	.73	.78	.81
Failed	.96	.96	.93	.94	.96	.96	.78	.79	.95	.95
LDC	.91	.91								

4.1.4 Validity

Table 9 shows the correlations of the 14-item PCS with construct validity indicators. All effect sizes were interpreted according to Cohen, 1992. We conducted multiple Shapiro-Wilk tests to assess normality and found that most variables were non-normally distributed. Therefore, we only report Spearman correlations for subsequent analyses.

All correlations between the PCS and construct validity indicators were positive, as expected. Effect sizes followed the anticipated pattern: The strongest correlations emerged for perceived competence ($\bar{r}_S$ = .74), SIPA ($\bar{r}_S$ = .68), and system trustworthiness ($\bar{r}_S$ = .66).

Indicators expected to correlate more weakly—system acceptance (r_S = .64), perceived warmth ($\bar{r}_S$ = .57), and perceived usefulness ($\bar{r}_S$ = .53)—nonetheless showed strong associations.

Table 9

Spearman correlations (and p-values) between PCS and validity indicators across samples

Variable	Han-HH	Han-RB	Han-RL	LDC	Augmented	Focused	Conflicted	Failed
SIPA		.76 (<.001)	.62 (<.001)		.65 (<.001)	.69 (<.001)	.62 (<.001)	.74 (<.001)
Warmth	.55 (<.001)	.59 (<.001)	.43 (<.001)		.50 (<.001)	.64 (<.001)	.62 (<.001)	.68 (<.001)
Competence	.84 (<.001)	.76 (<.001)	.67 (<.001)		.65 (<.001)	.70 (<.001)	.76 (<.001)	.78 (<.001)
Acceptance				.64 (<.001)				
FOST				.57 (<.001)	.58 (<.001)	.68 (<.001)	.69 (<.001)	.76 (<.001)
Usefulness					.48 (<.001)	.47 (<.001)	.51 (<.001)	.66 (<.001)

To evaluate the PCS's ability to differentiate between cooperation partners of varying cooperative capabilities, we repeated the statistical approach of Attig, Winzer, et al. (2026)—where differences in PCS mean scores between LLM-sampled Augmented, Focused, and Conflicted conditions had been tested—using the Hanabi samples from Study 1. Distributions of PCS scores across all samples are depicted in Figure 2, including histograms and boxplots. A repeated-measures ANOVA comparing PCS mean scores across three partner types was conducted: a human partner (Han-HH), a rule-based system (Han-RB), and a reinforcement learning-based system (Han-RL). Mauchly's test indicated that the assumption of sphericity was met (p = .194), and Shapiro-Wilk's test indicated that the residuals deviated from a normal distribution (p = .049). However, visual inspection of the QQ-plot supported an approximately normal distribution. No outliers were detected (Grubbs test, p = 1.00).

The ANOVA revealed a significant large main effect of partner type on PCS scores, $F(2,186) = 92.26$, $p < .001$, $\eta^2_G = .40$, indicating that perceived cooperativity differed substantially across conditions. A planned contrast analysis with weights $(-0.66, -0.33, 1)$ confirmed the hypothesized ordering Han-RL < Han-RB < Han-HH, $t(93) = 9.93$, $p < .001$, $d = 1.02$ (large effect). Pairwise comparisons with Bonferroni-Holm correction showed that PCS scores were significantly lower for Han-RL than for Han-RB, $t(93) = -11.39$, $p < .001$, $d = -1.22$, and lower for Han-RL than for Han-HH, $t(93) = -12.55$, $p < .001$, $d = -1.23$

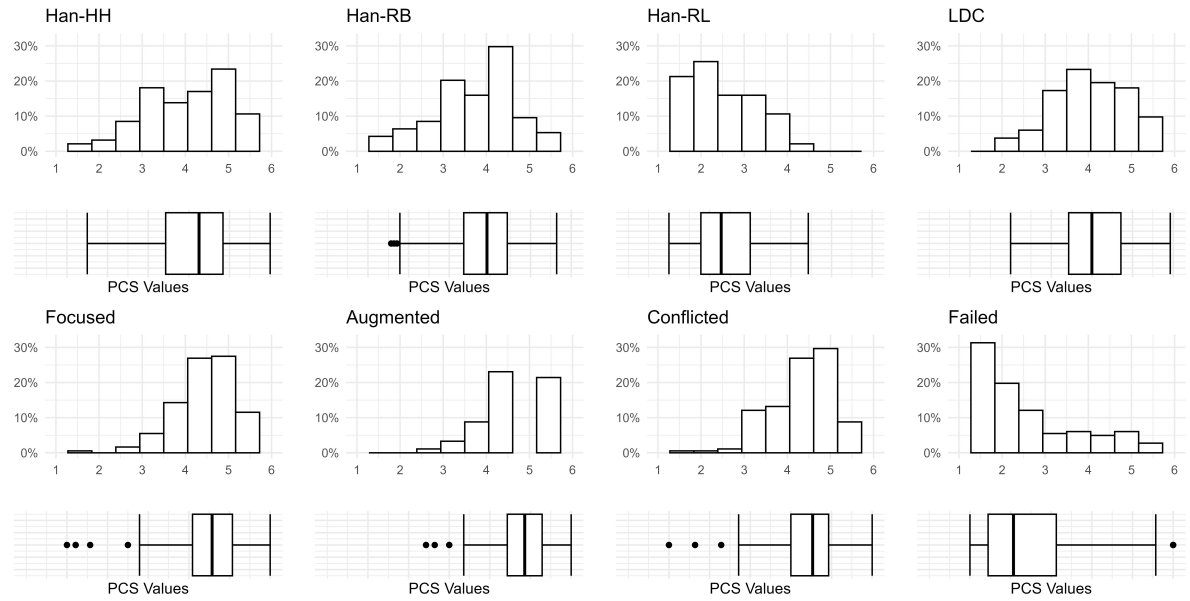

Figure 2

Histogram of PCS mean scores by sample. Boxplots below represent interquartile range (IQR), vertical lines are median, error bars are 1.5 times the IQR, and black dots are outliers beyond this threshold.

(both large effects). The difference between Han-RB and Han-HH was also significant, $t(93) = -2.34$, $p = .021$, $d = -0.23$ (small effect). Taken together, these results support the hypothesized ordering and indicate that the PCS successfully differentiates between cooperation partners of varying cooperative capability.

4.2 Teaming perception scale

4.2.1 Dimensionality

Teaming perception characterizes the emergent state of cooperation partners functioning as a team, arising from the interplay of perceived AI prosocial contributions and mutual human–AI support. To assess its factor structure, the TPS was developed with a theoretically grounded 3-factor model, which we first examined with EFA to assess scale dimensionality before confirming it with CFA.

As with the PCS, EFAs were conducted on the Hanabi samples of Study 1. First, parallel analysis using principal axis factoring (PAF) with 100 Monte Carlo simulations was employed to determine the number of factors. Results were inconclusive, supporting a two-factor solution for Han-HH sample (explaining 60 % of variance) and a three-factor

solution for both Han-RB and Han-RL samples (explaining 65 % and 60 % of variance). However, based on the Kaiser criterion, all three Hanabi samples indicated a one-factor solution with only the first factor having an eigenvalue exceeding 1 (explaining between 44 % and 49% of variance).

To inspect item-level factor structure, subsequent PAFs with Promax rotation were conducted. Factor loadings for the one-, two- and three-factor solutions showed that the two-factor model did not yield an easily interpretable factor structure (Factor 1: TPS01, TPS02, TPS03, TPS04, TPS07, TPS09; Factor 2: TPS05, TPS06, TPS08). The empirical three-model solution differed from the theoretical one, but was interpretable: On average across samples, Factor 1 comprised Items TPS01, TPS02, TPS03, TPS07, and TPS09 (items referring to general perception of teaming and partner contributions), Factor 2 comprised Items TPS04 and TPS06 (items referring to one's own support of the partner and team), and Factor 3 comprised items TPS05 and TPS08 (items referring to shared goal prioritization).

With CFAs, we assessed model fit of the one-factor solution, the three-factor solution based on EFA, and the three-factor solution based on theoretical considerations across the seven samples of Studies 1 and 2 (Table 10). Results indicate that the three-factor models showed consistently improved fit over a one-factor model across all samples, as indicated by significant χ^2-difference tests. The relative fit of the two three-factor solutions, however, varied across samples: The EFA-based model showed superior fit in the Hanabi samples, which is unsurprising given that the EFA was conducted on these data. More notably, the theory-based model performed comparably or better across the LLM-based samples (Focused, Augmented, Conflicted, Failed), supporting the generalizability of the theoretical factor structure across different human–AI cooperation contexts. This pattern was further supported by AIC and BIC values, which are information-theoretic indices that allow for direct model comparison irrespective of sample size, with lower values indicating better model fit relative to model complexity. AIC and BIC consistently favored the EFA-based model in the Hanabi samples and the theory-based model in the LLM-based samples. Crucially, however, the differences in fit between the two three-factor solutions were nuanced rather than substantial.[1]

[1] Please note that in this case, the model comparison is limited to comparing fit indices focusing on AIC and BIC information criteria, as traditional likelihood ratio tests are not suitable due to the non-nested model structure of

We therefore decided to retain the theory-based factor structure, as it is more directly grounded in our theoretical framework and yields the best interpretable subscales.

---

the two three-factor solutions (Merkle et al., 2016).

Table 10

Comparison of one-Factor and three-factor CFA models

	Three-Factor Model (Theory)						Three-Factor Model (EFA)						One-Factor Model				Comparison (Theory)		Comparison (EFA)	
Sample	CFI	TLI	RMSEA	SRMR	AIC	BIC	CFI	TLI	RMSEA	SRMR	AIC	BIC	CFI	TLI	RMSEA	SRMR	$\Delta\chi^2$ (3)	*p*	$\Delta\chi^2$ (3)	*p*
Han-HH	.91	.87[2]	0.14[3]	0.08	2188.63	2242.04	.97	.95	0.08	0.04	2163.39	2216.80	.86[1]	.82[2]	0.16[3]	0.10[4]	43.79	<.001	87.94	<.001
Han-RB	.94	.92	0.10	0.08	2163.44	2216.85	.99	.99	0.04	0.06	2143.20	2196.61	.90	.87[2]	0.13[3]	0.09[4]	15.08	.002	30.26	<.001
Han-RL	.95	.93	0.10	0.07	2264.31	2317.72	.94	.92	0.10	0.06	2267.25	2320.66	.91	.88[2]	0.12[3]	0.08	8.43	.038	10.44	.015
Focused	.97	.95	0.09	0.06	4135.26	4202.54	.92	.88[2]	0.15[3]	0.05	4189.36	4256.64	.90	.82[2]	0.16[3]	0.06	38.60	<.001	22.28	<.001
Augmented	.95	.92	0.11[3]	0.09[4]	4131.40	4198.69	.92	.88[2]	0.13[3]	0.06	4157.04	4224.33	.91	.88[2]	0.13[3]	0.07	17.83	<.001	5.27	.153
Conflicted	.93	.90	0.13[3]	0.07	4025.71	4093.00	.91	.86[2]	0.15[3]	0.05	4051.33	4118.62	.89[1]	.85[2]	0.16[3]	0.06	35.41	<.001	14.32	.003
Failed	.98	.97	0.10	0.05	4342.13	4409.42	.96	.94	0.13[3]	0.06	4373.53	4440.81	.92	.89[2]	0.18[3]	0.10[4]	41.63	<.001	48.08	<.001

Note. [1] CFI $<$ .90, indicating poor fit. [2] TLI $<$ .90, indicating poor fit. [3] RMSEA $>$ .10, indicating unacceptable fit. [4] SRMR $>$ .08, suggesting inadequate fit.

4.2.2 Item analysis

Item analysis followed the same procedure as for the PCS, with the exception that results are reported separately for the three TPS subscales (see Table 11).

For the Team subscale, item discrimination coefficients ranged on average across samples between $\bar{r}_{itc}$ = .75 (TPS03) and .79 (TPS01), indicating good item discrimination (Moosbrugger & Kelava, 2020). Item difficulty values ranged on average between 55.8 % (TPS01) and 61.6 % (TPS03), reflecting adequate item difficulty overall. As observed for the PCS, item difficulty values were considerably lower in the Han-RL condition (35.2 %) than in the Han-HH (73.3 %) and Han-RB (67.5 %) conditions, suggesting that participants agreed with Team subscale items more readily when interacting with the human partners and rule-based agent. Cronbach's α if item removed indicated that TPS02 and TPS03 would marginally increase subscale reliability in the Han-RL sample if removed. No items were flagged based on confidence rating z-scores.

For the Self subscale, item discrimination was substantially lower than for the Team and Partner subscales. Average item discrimination values across samples ranged between $\bar{r}_{itc}$ = .09 (TPS05) and .31 (TPS04), suggesting that the Self subscale measures a different construct than the overall scale. Item difficulty values ranged between 56.6 % (TPS05), 71.7 % (TPS04) and 80.6 % (TPS06), indicating that participants tend to agree readily with TPS04 and TPS06 across samples. Cronbach's α if item removed revealed that TPS05 would marginally increase subscale reliability across all three samples if removed. No items were flagged based on confidence rating z-scores.

For the Partner subscale, item discrimination coefficients ranged on average across samples between $\bar{r}_{itc}$ = .48 (TPS08) and .75 (TPS07), indicating good discrimination overall (Moosbrugger & Kelava, 2020). Item difficulty values ranged on average between 39.7 % (TPS08) and 59.5 % (TPS07), indicating adequate item difficulty. As with the Team subscale, item difficulty was considerably lower in the Han-RL condition (28.3 %) than in Han-HH (67.5 %) and Han-RB (57.3 %). Cronbach's α if item removed indicated that TPS08 would marginally increase subscale reliability across all three samples if removed. TPS08 was flagged on both confidence indicators: It showed below-average mean confidence ($z = -2.26$,

more than 2 *SD* below the overall mean) and high variability in confidence ratings ($z = 2.28$, more than 1 *SD* above the overall mean).

Table 11

TPS item analysis metrics (based on Hanabi samples of Study 1)

	Value $M(SD)$			Item discrimination (r_{itc})			Item difficulty			α if item removed			Confidence z	
Item	Han-HH	Han-RB	Han-RL	Han-HH	Han-RB	Han-RL	Han-HH	Han-RB	Han-RL	Han-HH	Han-RB	Han-RL	*M*	*SD*
TPS01	4.5 (1.5)	4.3 (1.3)	2.6 (1.3)	.79	.84	.81	69.2	66.0	32.1	.83	.79	.75	0.44	-0.43
TPS02	4.9 (1.1)	4.5 (1.3)	2.9 (1.3)	.76	.77	.71	77.7	70.0	37.2	.83	.80	.77[2]	0.27	0.07
TPS03	4.7 (1.2)	4.3 (1.2)	2.8 (1.2)	.79	.80	.74	73.2	66.6	36.2	.83	.80	.77[2]	0.13	-0.53
Team Subscale	4.7 (1.1)	4.4 (1.1)	2.8 (1.2)	.78	.80	.75	73.3	67.5	35.2					
TPS04	4.8 (1.0)	5.0 (0.8)	4.8 (0.9)	.53	.28[1]	.13[1]	76.6	79.8	76.2	.85	.85	.83	0.63	-1.01
TPS05	3.7 (1.3)	3.9 (1.1)	3.9 (1.3)	.15[1]	.04[1]	.07[1]	53.8	57.2	58.7	.89[2]	.87[2]	.85[2]	-0.92	0.64
TPS06	5.0 (0.8)	5.1 (0.7)	5.0 (0.9)	.58	.14[1]	.19[1]	79.6	82.1	80.2	.85	.86	.83	1.06	-0.91
Self Subscale	4.5 (0.7)	4.7 (0.6)	4.6 (0.7)	.42	.15	.13	70.0	73.1	71.7					
TPS07	4.9 (1.2)	4.4 (1.2)	2.6 (1.2)	.75	.79	.71	77.5	68.3	32.8	.83	.80	.77	0.40	-0.23
TPS08	3.7 (1.2)	3.1 (1.1)	2.2 (1.1)	.42	.45	.58	53.2	42.6	23.4	.86[2]	.83[2]	.79[2]	-2.26[4]	2.28[5]
TPS09	4.6 (1.1)	4.1 (1.3)	2.4 (1.2)	.69	.80	.70	71.9	61.1	28.7	.84	.79	.77	0.22	0.11
Partner Subscale	4.4 (0.9)	3.9 (1.0)	2.4 (1.0)	.62	.68	.66	67.5	57.3	28.3					
Composite Scale	4.5 (0.8)	4.3 (0.7)	3.3 (0.7)	.61	.55	.52	70.3	66.0	45.1	.86[3]	.84[3]	.81[3]		

Note. [1] Item discrimination $< .4$. [2] α if item removed $>$ subscale α. [3] Scale Cronbach's α. [4] Below-average z-standardized mean confidence ratings ($z < -2$ *SD*). [5] High variability of z-standardized *SD* of confidence ratings ($z > 1$ *SD*).

4.2.3 Reliability

Reliability analysis of the TPS was conducted for samples of Studies 1 and 2 (see Table 8 for all coefficients). All reliability values reported in the text below reflect averages across samples; interpretation followed common conventions (Cripps, 2017). The composite scale showed good internal consistency (α = .88, ω = .89). The Team subscale yielded excellent reliability (α = .91, ω = .91). The Partner subscale demonstrated acceptable to good reliability (α = .77, ω = .81). The Self subscale, in contrast, showed poor to questionable reliability (α = .55, ω = .65), with notably low values in the Hanabi samples. These results suggest that while the Team and Partner subscales are psychometrically robust across applications, the Self subscale shows higher internal consistency in non-game-based contexts (α = .65, ω = .73 across LLM samples) compared to the Hanabi samples (α = .42, ω = .53).

4.2.4 Validity

For construct validation of the TPS, we used only Spearman correlations for subsequent analyses due to non-normality (as for the PCS). Table 12 shows the Spearman correlations between the TPS subscales and construct validity indicators across samples. All effect sizes were interpreted according to Cohen (1992).

The Team and Partner subscales showed a consistent pattern of strong positive correlations across all samples and indicators, broadly in line with expectations. As anticipated, the strongest associations emerged for perceived competence (Team: $\bar{r}_S$ = .66; Partner: $\bar{r}_S$ = .68). Perceived warmth also correlated strongly with both subscales (Team: $\bar{r}_S$ = .56; Partner: $\bar{r}_S$ = .57), albeit with slightly lower correlation coefficients. The conceptually more distal indicators—SIPA and system usefulness—showed, however, similarly sized effects (SIPA – Team: $\bar{r}_S$ = .56; Partner: $\bar{r}_S$ = .59; system usefulness – Team: $\bar{r}_S$ = .51; Partner: $\bar{r}_S$ = .55).

The Self subscale showed a markedly different and more heterogeneous pattern, with substantially weaker correlations across samples, consistent with theoretical expectations. Excluding Han-RL, mean correlations were moderate for PCS ($\bar{r}_S$ = .44), SIPA ($\bar{r}_S$ = .37), warmth ($\bar{r}_S$ = .37), and competence ($\bar{r}_S$ = .37), with system usefulness showing the weakest association ($\bar{r}_S$ = .26). In Han-RB, several individual associations were non-significant. In the

Han-RL condition, negative correlations were found, specifically with perceived competence (r_S = -.36) and warmth (r_S = -.25).

To evaluate the PCS's ability to differentiate between cooperation partners of varying cooperative capabilities, we repeated the statistical approach of Attig, Winzer, et al. (2026)—where differences in TPS composite mean scores between LLM-sampled Augmented, Focused, and Conflicted conditions had been tested—using the Hanabi samples from Study 1. Distributions of TPS composite scores across samples (Study 1, Study 2) are depicted in Figure 3, including histograms and boxplots. A repeated-measures ANOVA comparing TPS composite mean scores across three partner types was conducted: a human partner (Han-HH), a rule-based system (Han-RB), and a reinforcement learning-based system (Han-RL). Mauchly's test indicated that the assumption of sphericity was met (p = .718). Shapiro-Wilk's test indicated a violation of the normality assumption (p = .018); however, visual inspection of the QQ-plot suggested approximate normality, and the ANOVA was conducted as planned given its robustness to mild normality violations. No outliers were detected (Grubbs test, p = 1.00).

The ANOVA revealed a significant main effect of partner type on TPS scores, $F(2, 186) = 79.16$, $p < .001$, $\eta^2_G = .36$, indicating that teaming perceptions differed substantially across conditions. A planned contrast analysis with weights $(-0.66, -0.33, 1)$ confirmed the hypothesized ordering Han-RL < Han-RB < Han-HH, $t(93) = 9.66$, $p < .001$, $d = 1.00$ (large effect). Pairwise comparisons with Bonferroni-Holm correction showed that TPS scores were significantly lower for Han-RL than for Han-RB, $t(93) = -9.90$, $p < .001$, $d = -1.05$, and lower for Han-RL than for Han-HH, $t(93) = -12.09$, $p < .001$, $d = -1.20$ (both large effects). The difference between Han-RB and Han-HH did not reach significance, $t(93) = -1.94$, $p = .056$, $d = -0.19$ (negligible effect). Taken together, these results largely support the hypothesized ordering and indicate that the TPS successfully differentiates between cooperation partners of varying cooperative capability. Distribution of TPS composite scores across samples (Study 1, Study 2) are depicted in Figure 3.

Table 12

Spearman correlations (and p-values) between TPS subscales and validity indicators across samples

Variable	Team subscale						
	Han-HH	Han-RB	Han-RL	Focused	Augmented	Conflicted	Failed
PCS	.82 (<.001)	.83 (<.001)	.76 (<.001)	.68 (<.001)	.63 (<.001)	.69 (<.001)	.73 (<.001)
SIPA		.66 (<.001)	.52 (<.001)	.63 (<.001)	.49 (<.001)	.45 (<.001)	.62 (<.001)
Warmth	.49 (<.001)	.59 (<.001)	.45 (<.001)	.57 (<.001)	.57 (<.001)	.57 (<.001)	.68 (<.001)
Competence	.76 (<.001)	.73 (<.001)	.53 (<.001)	.62 (<.001)	.55 (<.001)	.67 (<.001)	.77 (<.001)
System Usefulness				.48 (<.001)	.39 (<.001)	.50 (<.001)	.66 (<.001)
	Self subscale						
PCS	.47 (<.001)	.16 (.130)	-.20 (.054)	.56 (<.001)	.58 (<.001)	.60 (<.001)	.42 (<.001)
SIPA		.16 (.113)	.05 (.660)	.52 (<.001)	.42 (<.001)	.40 (<.001)	.35 (<.001)
Warmth	.25 (.017)	.07 (.482)	-.25 (.014)	.49 (<.001)	.53 (<.001)	.47 (<.001)	.39 (<.001)
Competence	.33 (.001)	.09 (.392)	-.36 (<.001)	.43 (<.001)	.46 (<.001)	.50 (<.001)	.38 (<.001)
System Usefulness				.26 (<.001)	.25 (.001)	.26 (<.001)	.27 (<.001)
	Partner subscale						
PCS	.76 (<.001)	.77 (<.001)	.79 (<.001)	.77 (<.001)	.67 (<.001)	.71 (<.001)	.77 (<.001)
SIPA		.63 (<.001)	.55 (<.001)	.64 (<.001)	.56 (<.001)	.55 (<.001)	.61 (<.001)
Warmth	.53 (<.001)	.64 (<.001)	.48 (<.001)	.58 (<.001)	.53 (<.001)	.53 (<.001)	.71 (<.001)
Competence	.70 (<.001)	.70 (<.001)	.59 (<.001)	.71 (<.001)	.61 (<.001)	.68 (<.001)	.80 (<.001)
System Usefulness				.50 (<.001)	.43 (.001)	.55 (<.001)	.71 (<.001)

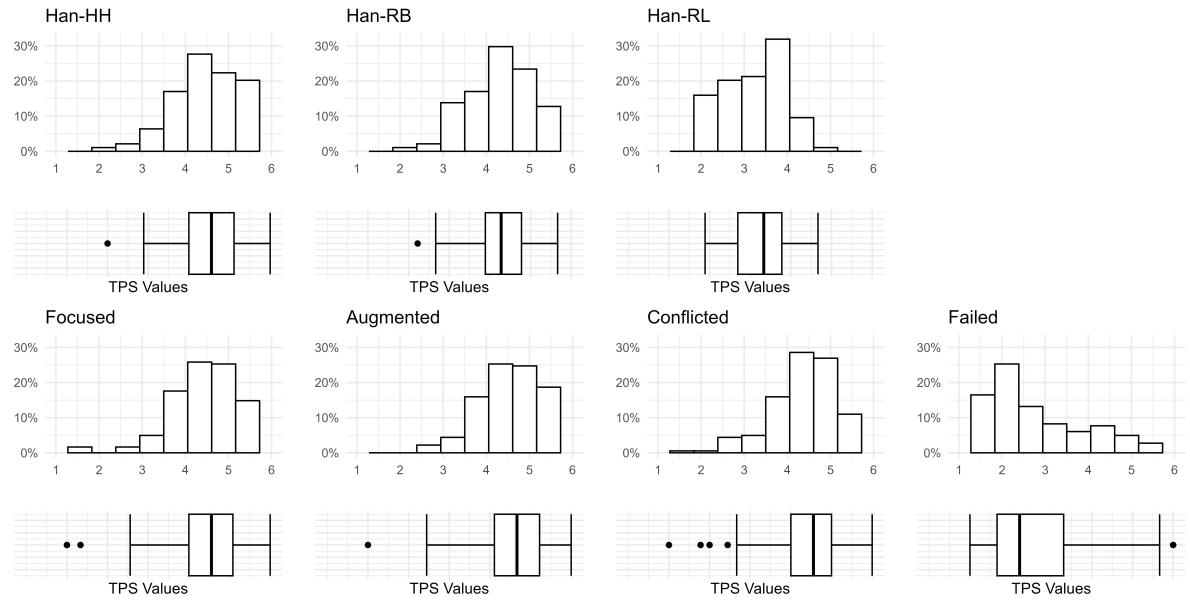

Figure 3

Histogram of TPS composite mean scores by sample. Boxplots below represent interquartile-range (IQR), vertical lines are median, error bars are 1.5 times the IQR, and black dots are outliers beyond this threshold.

5. Discussion

5.1 Summary of results

The objective of the present research was to develop and validate two new scales to assess perceived cooperativity and teaming perception in human–AI cooperation, grounded in joint activity theory and evolutionary cooperation theory respectively. Tests of the scales across three studies with diverse samples and interaction contexts ($N = 409$) showed satisfying results with regard to dimensionality, reliability, validity, and the ability to differentiate between cooperation partners. Specifically, the results can be summarized as follows:

- Factor analyses indicated unidimensionality for the PCS and a three-factor structure for the TPS, with subscales reflecting team perception, self-contribution, and partner contribution.
- Item analyses revealed good item discrimination and adequate item difficulty for the majority of items in both scales, with the Self subscale of the TPS showing notably weaker discrimination compared to the Team and Partner subscales.
- Reliability analyses showed excellent internal consistency for the PCS across all samples and contexts, and good to excellent internal consistency for the TPS Team and Partner subscales, while the Self subscale showed acceptable internal consistency primarily in non-game-based contexts.
- Construct validity analyses support expected relationships to subjective information processing awareness, perceived competence and warmth, system trustworthiness, system acceptance, and system usefulness, with effect sizes generally following the anticipated pattern of conceptual proximity to the respective constructs. Additionally, both scales successfully differentiated between cooperation partners of varying cooperative quality, with human partners rated highest, followed by the rule-based agent, and the reinforcement learning-based agent rated lowest, confirming the hypothesized ordering.
- Boundary case analyses using the Tesla trip planner in Study 3 suggest that the PCS

remains applicable in contexts of lower agent agency.

5.2 Critical discussion of scale evaluation results

While the overall psychometric properties of the PCS and TPS are promising, several limitations emerged from the analyses that warrant closer examination. For the PCS, the most striking psychometric issue was the inconsistent performance of item PCS14 across contexts. This item ("I had the feeling that the person/agent was able to realize that we had different approaches to achieving the goal.") showed weaker item discrimination and lower confidence ratings than the remaining items. CFA results further suggested the 14-item solution without PCS14 to be more appropriate than the 15-item solution. Consequently, we excluded the item from the scale, reducing the PCS to 14 items, and used the final 14-item version for reliability and validity analyses.

The correlational validity analyses of both the PCS and the TPS revealed that, while correlations with conceptually proximal constructs were strongest as hypothesized, some constructs that were expected to show weaker associations nonetheless yielded strong correlations. As Moradbakhti et al. (2024) argue for a similar pattern of findings, these unexpectedly high correlations do not necessarily reduce the validity of the scales as such: A degree of correlation between psychological constructs is to be expected in the social sciences, as virtually all theoretically meaningful variables share some common variance (Meehl, 1990). Moreover, given that all measures in the present studies were administered within the same session and in close succession, common method bias (Podsakoff et al., 2003) may have inflated some of the observed associations. Future studies should therefore further examine the discriminant validity of the PCS and TPS by using designs that reduce the risk of common method bias (e.g., by separating the administration of the focal scales and the validity indicators across time points). Additionally, criterion validity should be established by linking scale scores to objective indicators of cooperative success (e.g., task performance, error rates) as well as to implicit measures of cooperative experience (e.g., spontaneous use of first-person plural language when describing the interaction, the degree to which participants follow or override AI suggestions, indices of behavioral synchrony between human and AI partner actions over time).

Regarding the TPS, the most striking psychometric issues concern the Self subscale. Unlike the Team and Partner subscales, which ask participants to evaluate the interaction and the AI partner from a third-person perspective, the Self subscale turns the evaluative lens inward, asking participants to assess their own cooperative contribution and support from a first-person perspective. This fundamental difference in both content and perspective means that the Self subscale is, by design, measuring something distinct from the other two subscales. The substantially lower item discrimination coefficients observed across the three Hanabi conditions in Study 1 are consistent with this: Low discrimination values indicate that the Self subscale measures a different aspect of the construct than the other two subscales (Moosbrugger & Kelava, 2020), which is theoretically expected. However, when also taking the item difficulty values into account, a more revealing pattern emerges. Items TPS04 and TPS06 showed consistently high item difficulty values across conditions, suggesting that participants tended to rate themselves as highly cooperative, regardless of the interaction partner they were paired with. This gives rise to the hypothesis that the Self subscale—although designed as an interaction-specific self evaluation measure—may, under certain conditions, capture a more stable and positively biased evaluation of one's own cooperative tendency rather than a genuine reflection of the specific interaction (possibly prone to social desirability bias; Chung and Monroe, 2003). Particularly telling in this regard are the raw item score means (see Table 11): While the Team and Partner subscale scores decreased systematically across conditions in line with the decreasing cooperative quality of the interaction partner, the Self subscale mean scores remained stable across all three conditions (Han-HH: 4.5, Han-RB: 4.7, Han-RL: 4.6), providing further support in the raw data that participants' self-evaluations were largely unaffected by the nature of the actual interaction.

Supporting this interpretation, negative correlations between the Self subscale and agent evaluations such as perceived warmth and competence emerged specifically in the Han-RL condition: When the interaction partner had lower perceived cooperativity, participants still rated their own cooperative investment highly, resulting in an inverse relationship between self- and partner-focused evaluations in this condition. At the same time, the structural characteristics of the Hanabi task may have contributed to this pattern. Because

the game enforces cooperative behavior structurally (i.e., players cannot opt out of cooperating with their game partner), there is comparatively restricted variance in the extent to which participants can meaningfully vary their cooperative investment across conditions. The Self subscale items may therefore have captured a genuine psychological reality (i.e., participants did feel they were contributing and supporting their partner) that resulted in a ceiling effect across participants. Taken together, these observations are in line with the idea that when task structure restricts the degrees of freedom regarding one's own cooperative investment, what remains visible is a tendency to evaluate oneself positively regardless of the specific interaction. Consistent with this interpretation, internal consistency of the Self subscale improved in the LLM-based interaction context of Study 2. A possible explanation is that participants had considerably more freedom in deciding how actively to engage with the AI partner in this context, which may have allowed for larger variation in their own cooperative investment than the structurally enforced cooperation of the Hanabi task. However, this interpretation remains tentative in the absence of a direct behavioral measure of engagement.

Another psychometric issue of the TPS concerns a context-dependency in the performance of the two goal prioritization items (TPS05 and TPS08), which should be taken into account when applying the TPS in future research. Item analysis revealed near-zero discrimination values for TPS05 across all three Hanabi conditions, and notably low confidence ratings and high response variability for TPS08. Moreover, both items loaded onto a shared factor in the EFAs for the Han-RB and Han-RL samples, suggesting that, under certain conditions, they form a distinct dimension that is not well integrated with the remaining TPS items in these contexts. Specifically, genuinely prioritizing a shared goal over individual goals might not have been a meaningful experience in the Hanabi setting, which is a context where game partners share an identical goal by design, leaving little room for individual goals to compete with shared ones. As a consequence, participants' responses to TPS05 may have reflected a self-serving evaluation of their own cooperative tendency rather than a genuine interaction-specific experience, while TPS08 might have confronted participants with a judgment that was simply not answerable given the constraints of the task. Together, these psychometric issues suggest that the Self subscale, in its current form, likely

requires refinement to more reliably capture interaction-specific cooperative experience rather than stable self-evaluations.

Beyond its current psychometric limitations, the observed pattern in the Self subscale—namely, that participants tended to rate their own cooperative investment consistently and positively regardless of their interaction partner's cooperative quality—may itself carry theoretical value. It raises the question of how sensitive individuals' self-perceived cooperative investment is to their actual cooperation experience, and whether this sensitivity plays a role in shaping how people attribute responsibility for cooperation failures (Kares et al., 2026; Yu et al., 2026) or calibrate their trust in AI systems over time (Li & Lee, 2026; Schlicker et al., 2025). Future research explicitly designed to manipulate behavioral engagement could help clarify whether the Self subscale is able to capture this dynamic.

A further limitation concerns the composition of the human-human condition in Study 1: While the majority of participants (76 of 94) were paired with another study participant, 18 participants played with a member of the research team standing in as their human partner due to participant no-shows. Although the same procedural instructions and game format applied across both pairing types, research team members may have exhibited more consistent or practiced play patterns than participant-participant pairs, given their repeated exposure to the game across sessions. To examine whether this introduced systematic bias, we compared PCS and TPS scores between participants paired with a confederate and those paired with another participant, both within the Han-HH condition alone and across all three Hanabi conditions combined. No statistically significant differences emerged for either scale in either comparison (all $p > .290$). These results suggest that the inclusion of confederate pairings did not meaningfully bias the Han-HH condition.

Finally, it is worth noting that all three validation studies relied on relatively short interaction episodes, limiting the degree to which diachronic integration could be realized. While the Hanabi study introduced a limited diachronic element through repeated game rounds, and the LLM study included multiple trials per chatbot version, none of the studies examined genuine long-term partnerships characterized by deep diachronic integration over a variety of tasks. Future studies might therefore adopt longitudinal designs to examine how

perceived cooperativity and teaming perception develop, stabilize, and interact over time, whether early cooperative experiences predict later teaming perception, and whether the Self subscale proves more sensitive and reliable when participants have had sufficient opportunity to develop a genuine sense of their own role and investment in a long-term human-AI partnership.

5.3 Recommendations for scale use

The following recommendations are intended to guide researchers in applying the PCS and TPS in a way that is informed by both the theoretical framework and the psychometric findings of the present validation studies.

- **Assess the interaction context and choose the appropriate scale(s).** Before selecting a scale, assess the degree of synchronic and diachronic integration of the target interaction. The PCS is appropriate across a wide range of contexts, including first encounters, relatively low synchronic integration, and systems with limited agency. When the primary interest concerns the perceived cooperative quality of the AI agent and study resources are limited, the PCS alone is recommended: Given its psychometric robustness and unidimensional factor structure, it can be easily administered and scored, making it a economical option for applied evaluation purposes. The TPS is appropriate when a minimum degree of synchronic integration and agent agency are present and the shared task is genuinely cooperative (i.e., when the task requires meaningful contributions from both partners and a sense of joint goal pursuit can plausibly emerge). When the research question calls for a more complete picture of the cooperative experience (e.g., when examining how single cooperative sequences and long-term teaming interact), both scales should be administered.

- **Conduct a pilot study.** Where feasible, pre-test the scales with a small sample prior to main data collection, including item-level confidence ratings (e.g., by asking participants to rate their confidence in responding to single items on a scale from 1 [*not confident at all*] to 10 [*extremely confident*], as we did in Study 1). Although the scales were developed with the intention to be broadly applicable to cooperative settings of all

kinds, we cannot claim for certainty that all items generalize equally well across each setting. Hence, researchers should pay particular attention to items whose applicability may depend on context-specific features: For instance, TPS05 and TPS08 (goal prioritization) might not be meaningful in settings where individual and shared goals cannot genuinely diverge, PCS12 might be less relevant in settings coordinated primarily through implicit or non-verbal means, and PCS15 might not be applicable when no problems arose during the interaction. Items flagged by low confidence ratings or high response variability should be considered for omission before the main study.

- **Administer the scales at the right time.** In repeated-measures or longitudinal designs, administer the scales after each interaction episode rather than retrospectively—particularly when the cooperation partners differ between episodes (e.g., LLM-based chatbots with different system prompts; Attig, Winzer, et al., 2026). Cooperative experience is likely episode-specific, and retrospective ratings risk conflating distinct episodes into a single undifferentiated judgment.

- **Examine the factor structure and reliability.** Particularly when applying the scales in a new cooperative setting, dimensionality and internal consistency should be assessed before computing composite or subscale scores, as item performance may vary across contexts. Item-level diagnostics such as item-total correlations and Cronbach's α if item removed can help identify items that are not performing well in a given setting and may warrant exclusion. For the TPS, conduct exploratory and, where sample size permits, confirmatory factor analyses on the collected data. The theory-based three-factor structure (Team, Self, Partner) should be used as the default, particularly when the empirical factor structure is inconclusive or when the theoretically proposed subfacets show improved model fit. Retaining the theory-based structure ensures cross-study comparability. However, if the data clearly supports another factor structure, calculate subscale scores for the empirical factors.

- **Report and interpret subscales appropriately.** Report the three TPS subscales separately, as they capture meaningfully distinct aspects of the teaming experience. A

composite TPS score may be reported as a broad indicator of overall teaming perception, but we would recommend to only do so if the internal consistency of the composite reaches at least a good level (Cronbach's $\alpha \geq .80$). The Self subscale should be interpreted with particular caution given its susceptibility to positivity bias and context-dependent psychometric properties.

5.4 Implications and conclusion

The present research provides evidence to conclude that the PCS and TPS are promising tools for assessing the subjective experience of human–AI cooperation on both synchronic and diachronic dimensions. Together, the scales address a gap in the existing measurement landscape: While a large variety of scales has been developed to assess general attitudes towards AI (Schepman & Rodway, 2025) and AI literacy (Lintner, 2024), theory-based instruments capable of capturing the episodic, process-oriented experience of cooperative interaction with AI agents are more rare. By grounding the PCS in joint activity theory (Bradshaw et al., 2017; Klein et al., 2005) and the TPS in evolutionary cooperation theory (Nowak, 2006; Tomasello et al., 2012) as well as shared action frameworks (Bratman, 1992; Pacherie, 2014), the scales allow for connecting empirical findings to established literature on cooperation and joint activity. In the following, we will highlight specific implications from the present results.

Informing AI system design through diagnostic items. Rather than capturing users' overall disposition towards AI or their general satisfaction with a system, the proposed scales target the perceived quality of a specific cooperative episode and the emergent sense of teaming that arises from it. This specificity makes them sensitive to differences that broader instruments may miss, as evidenced by the scales' ability to discriminate between cooperation partners that differ in their cooperative quality. Viewed from an applied perspective, the PCS in particular has diagnostic potential beyond its use as an outcome measure. Because its items were derived directly from the eight characteristics of an effective cooperative agent as defined by joint activity theory (i.e., observability, monitorability, communicativeness, self-assessment, reliability, directability, selectivity, and coordination; Bradshaw et al., 2017), subscale or item-level scores can reveal which specific cooperative capabilities an AI agent is

perceived to lack, enabling actionable design feedback. For instance, if users consistently rate an agent low on items related to predictability and directability, but higher on items related to communicativeness, this pattern points to specific aspects of agent behavior that designers should prioritize for enhanced user experience and acceptance. The TPS complements this by capturing whether the overall interaction gives rise to a genuine sense of teaming, including the human partner's own sense of contribution and mutual support. Thus, including the TPS in system evaluations might be particularly beneficial in highly interdependent human-AI cooperation settings where both human and artificial contributions determine task success. For instance, in industrial human-robot collaboration, where the quality of the shared outcome depends on both partners continuously adapting to each other's actions, a positive overall task outcome may mask an unsatisfying cooperative experience (Siu et al., 2021), and the TPS provides a means to capture this experiential dimension independently of performance metrics.

Enabling cross-agent comparisons. A practical benefit of the scales is the availability of both AI and human partner versions, which use identical item formulations with only the referent changed. This allows to directly benchmark human-AI cooperation against human-human cooperation in a given context using the same instrument, enabling more nuanced conclusions about whether and to what degree AI agents can serve as genuine cooperation partners. Such comparisons are particularly relevant in hybrid team settings where humans collaborate with both human and artificial teammates.

Applicability to diverse study designs. The PCS and TPS are sufficiently economical to be embedded in larger study designs without imposing undue participant burden. They can serve as dependent variables for evaluating AI system quality, as manipulation checks in experimental designs that vary cooperative conditions, or as moderator and mediator variables in models of human-AI interaction outcomes. The availability of both AI and human partner versions further extends their utility to studies with mixed or hybrid team designs. Given the heterogeneity of the three samples in the present validation studies (i.e., ranging from students playing a cooperative card game to Prolific participants interacting with LLM-based chatbots to electric vehicle drivers using a real-world trip planning system), the scales appear suited for deployment across a wide range of populations and application

contexts. Future research should extend this work by testing the PCS and TPS in additional cooperative settings (e.g., real-time physical or virtual coordination tasks inspired by Overcooked; Salikutluk et al., 2024), to further establish the generalizability of the scales beyond the contexts examined here.

Disclosure of potential conflicts of interest

All authors declare that they have no conflicts of interest.

Author contributions: CRediT

Christiane Attig: Conceptualization; Funding acquisition; Investigation; Methodology; Project administration; Resources; Visualization; Writing – original draft; Writing – review & editing. Christiane Wiebel-Herboth: Conceptualization; Project administration; Resources; Supervision; Writing – review & editing. Patricia Wollstadt: Conceptualization; Resources; Writing – review & editing. Tim Schrills: Conceptualization; Methodology. Mourad Zoubir: Formal Analysis; Visualization; Writing – original draft. Thomas Franke: Conceptualization; Funding acquisition; Project administration; Resources; Supervision; Writing – review & editing.

Acknowledgments

We want to thank Laura Duwe, Sophia von Elbwart, Simon Flintrop, Bennet Hut, and Jacob Mellin for their support in collecting data for the Hanabi study as well as Beate Stattkus-Fortange for the LDC study. Further, we want to thank Tobias Harms, Jan Heidinger, Maged Mortaga, Lukas Rudat, and Luisa Winzer for their support in developing the study environments for the Hanabi, resp. LLM study, and Michelle Wrage for her support in data analysis. Finally, we want to thank Mark Dunn and Silvia Wersing for supporting the scale translations.

This work was supported by the Honda Research Institute Europe GmbH (project CoCharge) and by the first author's DenkRaum Fellowship from the Universities of Luebeck and Kiel (2023–2025).

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this work the authors used the large language model Claude (Sonnet 4.6) in order to improve language quality. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Data Availability Statement

The data that support the findings of this study are available at Dryad (https://doi.org/10.5061/dryad.5qfttdzpg).

Author Biographies

Christiane Attig is a senior scientist at Honda Research Institute Europe in Offenbach. Her research centers on understanding cooperation between humans and AI systems. She focuses on scale development and user diversity, examining how culture, social role perception, and design shape human–AI interaction that supports basic psychological needs.

Christiane Wiebel-Herboth is a senior scientist at Honda Research Institute Europe in Offenbach, where she works in the field of human–machine interaction. Her research focuses on understanding what constitutes successful human–AI cooperation and how to define concepts and measures for evaluating and designing such systems in a human-centered way.

Patricia Wollstadt is a principal scientist at Honda Research Institute Europe in Offenbach, where she studies human–machine interaction and cooperation. Her research combines statistical modeling with information-theoretic methods to quantitatively characterize successful cooperation between humans and AI.

Tim Schrills is a postdoctoral researcher at the Institute for Human-Centered Interactive Systems, University of Luebeck. His research focuses on human–AI interaction, particularly how people deal with AI-generated information. He develops theoretical models of information processing in human–AI collaboration and studies prerequisites of human oversight of AI.

Mourad Zoubir is a former postdoctoral researcher at the Chair of Engineering Psychology and Cognitive Ergonomics at University of Luebeck, where he completed his PhD

in 2025. His research addresses psychological basic needs, user diversity, automation psychology, and the development of questionnaire-based methods for studying human–automation interaction.

Thomas Franke is a professor of Engineering Psychology and Cognitive Ergonomics at University of Luebeck. His research centers on understanding and supporting resource regulation in human–technology environments. He focuses on theory and method development related to user diversity and the effects of system design on user experience and behavior.

References

Ajagbe, S. A., & Adigun, M. O. (2024). Deep learning techniques for detection and prediction of pandemic diseases: A systematic literature review. *Multimedia Tools and Applications*, *83*(2), 5893–5927. https://doi.org/10.1007/s11042-023-15805-z

Akhtar, Z. B. (2024). Unveiling the evolution of generative AI (GAI): A comprehensive and investigative analysis toward LLM models (2021–2024) and beyond. *Journal of Electrical Systems and Information Technology*, *11*(1), 22. https://doi.org/10.1186/s43067-024-00145-1

Al-Khasawneh, M. A., Raza, A., Khan, S. U. R., & Khan, Z. (2025). Stock market trend prediction using deep learning approach. *Computational Economics*, *66*(1), 453–484. https://doi.org/10.1007/s10614-024-10714-1

Alqasir, A. (2025). The relational shift: Why we need "AI psychology" now as a core field. *Journal of Psychology and AI*, *1*(1), 2573928. https://doi.org/10.1080/29974100.2025.2573928

Attig, C., Sarkisian, A., Chew, J. Y., & Wiebel-Herboth, C. (2025). Beyond reciprocity: Psychological needs as a foundation for human-AI cooperation. In *3th International Conference on Human-Agent Interaction (HAI '25)*. ACM. https://doi.org/10.1145/3765766.3765854

Attig, C., Winzer, L., Schrills, T., Zoubir, M., Mortaga, M., Wollstadt, P., Wiebel-Herboth, C., & Franke, T. (2026). Understanding successful human–AI teaming: The role of goal alignment and AI autonomy for social perception of LLM-based chatbots. *Computers in Human Behavior: Artificial Humans*, *7*, 100246. https://doi.org/10.1016/j.chbah.2025.100246

Attig, C., Wollstadt, P., Sarkisian, A., Chew, J. Y., & Wiebel-Herboth, C. (2026). Human–AI Cooperation Reconsidered: Integrating Reciprocity and Psychological Needs. In H. Degen & S. Ntoa (Eds.), *Artificial Intelligence in HCI* (pp. 235–261, Vol. 16743). Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-30846-7_15

Attig, C., Wollstadt, P., Schrills, T., Franke, T., & Wiebel-Herboth, C. B. (2024). More than task performance: Developing new criteria for successful human-AI teaming using the

cooperative card game Hanabi. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–11. https://doi.org/10.1145/3613905.3650853

Bard, N., Foerster, J. N., Chandar, S., Burch, N., Lanctot, M., Song, H. F., Parisotto, E., Dumoulin, V., Moitra, S., Hughes, E., Dunning, I., Mourad, S., Larochelle, H., Bellemare, M. G., & Bowling, M. (2020). The Hanabi challenge: A new frontier for AI research. *Artificial Intelligence*, *280*, 103216. https://doi.org/10.1016/j.artint.2019.103216

Basappa, R., Lancaster, C., Mallick, R., Flathmann, C., & McNeese, N. (2025). Mind the gaps: How AI shortcomings and human concerns may disrupt team cognition in human-AI teams (HATs). *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 10711813251361002. https://doi.org/10.1177/10711813251361002

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*(3), 588–606. https://doi.org/10.1037/0033-2909.88.3.588

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in public health*, *6*, 149. https://doi.org/10.3389/fpubh.2018.00149

Bradshaw, J. M., Feltovich, P. J., & Johnson, M. (2017). Human–agent interaction. In *The Handbook of Human-Machine Interaction* (pp. 283–300). CRC Press.

Bratman, M. E. (1992). Shared cooperative activity. *The Philosophical Review*, *101*(2), 327. https://doi.org/10.2307/2185537

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*(2), 230–258. https://doi.org/10.1177/0049124192021002005

Buchanan, I. (2010). *A dictionary of critical theory*. Oxford University Press. https://doi.org/10.1093/acref/9780199532919.001.0001

Budžys, A., Kurasova, O., & Medvedev, V. (2024). Deep learning-based authentication for insider threat detection in critical infrastructure. *Artificial Intelligence Review*, *57*(10), 272. https://doi.org/10.1007/s10462-024-10893-1

Canaan, R., Gao, X., Chung, Y., Togelius, J., Nealen, A., & Menzel, S. (2020). Behavioral evaluation of Hanabi Rainbow DQN agents and rule-based agents. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, *16*(1), 31–37. https://doi.org/10.1609/aiide.v16i1.7404

Chang, F., & Herath, D. (2025). From interaction to relationship: The role of memory, learning, and emotional intelligence in AI-embodied human engagement. *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 1269–1273. https://doi.org/10.1109/HRI61500.2025.10973813

Chung, J., & Monroe, G. S. (2003). Exploring social desirability bias. *Journal of Business Ethics*, *44*(4), 291–302. https://doi.org/10.1023/A:1023648703356

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Cripps, B. (2017). *Psychometric testing: Critical perspectives*. John Wiley & Sons.

de Visser, E. J., Peeters, M. M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a theory of longitudinal trust calibration in human–robot teams. *International Journal of Social Robotics*, *12*(2), 459–478. https://doi.org/10.1007/s12369-019-00596-x

DeVrio, A., Cheng, M., Egede, L., Olteanu, A., & Blodgett, S. L. (2025). A taxonomy of linguistic expressions that contribute to anthropomorphism of language technologies. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3706598.3714038

Franke, T., Attig, C., & Wessel, D. (2019). A personal resource for technology interaction: Development and validation of the affinity for technology interaction (ATI) scale. *International Journal of Human–Computer Interaction*, *35*(6), 456–467. https://doi.org/10.1080/10447318.2018.1456150

Franke, T., Trantow, M., Günther, M., Krems, J. F., Zott, V., & Keinath, A. (2015). Advancing electric vehicle range displays for enhanced user experience: The relevance of trust and adaptability. *Proceedings of the 7th International Conference on Automotive User*

Interfaces and Interactive Vehicular Applications, 249–256. https://doi.org/10.1145/2799250.2799283

Honda Research Institute EU and synyx. (2023, August 12). *Hanabi4hmi* (Version 1.1.0). https://github.com/HRI-EU/Hanabi4HMI

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Javed, R., Abbas, T., Khan, A. H., Daud, A., Bukhari, A., & Alharbey, R. (2024). Deep learning for lungs cancer detection: A review. *Artificial Intelligence Review*, *57*(8), 197. https://doi.org/10.1007/s10462-024-10807-1

Johnson, M., Bradshaw, J. M., & Feltovich, P. J. (2018). Tomorrow's human–machine design tools: From levels of automation to interdependencies. *Journal of Cognitive Engineering and Decision Making*, *12*(1), 77–82. https://doi.org/10.1177/1555343417736462

Kares, F., König, C. J., Polly, M., & Langer, M. (2026). Responsibility attributions in AI-assisted decisions: The interplay of the level of automation and system reliability. *Computers in Human Behavior: Artificial Humans*, *9*, 100336. https://doi.org/10.1016/j.chbah.2026.100336

Kirk, H. R., Gabriel, I., Summerfield, C., Vidgen, B., & Hale, S. A. (2025). Why human–AI relationships need socioaffective alignment. *Humanities and Social Sciences Communications*, *12*(1), 728. https://doi.org/10.1057/s41599-025-04532-5

Klein, G., Feltovich, P. J., Bradshaw, J. M., & Woods, D. D. (2005). Common ground and coordination in joint activity. *Organizational simulation*, *53*, 139–184. https://doi.org/10.1002/0471739448.ch6

Lakhmani, S. G., Neubauer, C., Krausman, A., Fitzhugh, S. M., Berg, S. K., Wright, J. L., Rovira, E., Blackman, J. J., & Schaefer, K. E. (2022). Cohesion in human–autonomy teams: An approach for future research. *Theoretical Issues in Ergonomics Science*, *23*(6), 687–724. https://doi.org/10.1080/1463922X.2022.2033876

Leach, C. W., Ellemers, N., & Barreto, M. (2007). Group virtue: The importance of morality (vs. competence and sociability) in the positive evaluation of in-groups. *Journal of personality and social psychology*, *93*(2), 234. https://doi.org/10.1037/0022-3514.93.2.234

Lee, Y., & Kim, S.-H. (2025). Exploring dimensions of human likeness in conversational AI: Implications for human identity threat and dehumanization. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 10711813251358801. https://doi.org/10.1177/10711813251358801

Li, M., & Lee, J. D. (2026). Purpose outweighs performance: Trust drops more when AI teammates fail to cooperate, but explanations can repair it. *International Journal of Human–Computer Interaction*, *42*(6), 4378–4399. https://doi.org/10.1080/10447318.2025.2539449

Lindgren, H. (2025). Emerging roles and relationships among humans and interactive AI systems. *International Journal of Human–Computer Interaction*, *41*(17), 10595–10617. https://doi.org/10.1080/10447318.2024.2435693

Lintner, T. (2024). A systematic review of AI literacy scales. *npj Science of Learning*, *9*(1), 50. https://doi.org/10.1038/s41539-024-00264-4

Lyons, J. B., Sycara, K., Lewis, M., & Capiola, A. (2021). Human–autonomy teaming: Definitions, debates, and directions. *Frontiers in Psychology*, *12*, 589585. https://doi.org/10.3389/fpsyg.2021.589585

Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. *The Academy of Management Review*, *26*(3), 356. https://doi.org/10.2307/259182

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(3), 320–341. https://doi.org/10.1207/s15328007sem1103_2

McNeese, N. J., Schelble, B. G., Canonico, L. B., & Demir, M. (2021). Who/what is my teammate? Team composition considerations in human–AI teaming. *IEEE Transactions on Human-Machine Systems*, *51*(4), 288–299. https://doi.org/10.1109/THMS.2021.3086018

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*(1), 195–244. https://doi.org/10.2466/pr0.1990.66.1.195

Merkle, E. C., You, D., & Preacher, K. J. (2016). Testing nonnested structural equation models. *Psychological methods*, *21*(2), 151–163. https://doi.org/10.1037/met0000038

Moosbrugger, H., & Kelava, A. (Eds.). (2020). *Testtheorie und Fragebogenkonstruktion*. Springer. https://doi.org/10.1007/978-3-662-61532-4

Moradbakhti, L., Leichtmann, B., & Mara, M. (2024). Development and validation of a basic psychological needs scale for technology use. *Psychological Test Adaptation and Development*, *5*(1), 26–45. https://doi.org/10.1027/2698-1866/a000062

Musick, G., O'Neill, T. A., Schelble, B. G., McNeese, N. J., & Henke, J. B. (2021). What happens when humans believe their teammate is an AI? An investigation into humans teaming with autonomy. *Computers in Human Behavior*, *122*, 106852. https://doi.org/10.1016/j.chb.2021.106852

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 72–78. https://doi.org/10.1145/191666.191703

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science*, *314*(5805), 1560–1563. https://doi.org/10.1126/science.1133755

O'Neill, T., McNeese, N., Barron, A., & Schelble, B. (2022). Human–autonomy teaming: A review and analysis of the empirical literature. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *64*(5), 904–938. https://doi.org/10.1177/0018720820960865

Pacherie, E. (2014). How does it feel to act together? *Phenomenology and the Cognitive Sciences*, *13*(1), 25–46. https://doi.org/10.1007/s11097-013-9329-8

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Prolific. (2024). Prolific – online participant recruitment [Accessed: 2024-10-08].

Qian, Y., & Wan, X. (2025). Co-experiencing with AI: Effects on social bonding and empathy in human-AI relationships. *Technology in Society*, *83*, 103009. https://doi.org/10.1016/j.techsoc.2025.103009

Rieth, M., Ontrup, G., Kluge, A., & Hagemann, V. (2026). Unveiling team emergent states in the age of human–AI teaming. *International Journal of Human–Computer Interaction*, 1–28. https://doi.org/10.1080/10447318.2026.2635683

Salikutluk, V., Schöpper, J., Herbert, F., Scheuermann, K., Frodl, E., Balfanz, D., Jäkel, F., & Koert, D. (2024). An evaluation of situational autonomy for human-AI collaboration in a shared workspace setting. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–17. https://doi.org/10.1145/3613904.3642564

Schelble, B. G., Mallick, R., Hauptman, A., & McNeese, N. (2025). Should AI teammates give all the answers? Examining the role of different AI information-sharing techniques on team cognition in human–AI teams. *International Journal of Human–Computer Interaction*, 1–26. https://doi.org/10.1080/10447318.2025.2528988

Schepman, A., & Rodway, P. (2025). The measurement of attitudes towards artificial intelligence: An overview and recommendations. In C. Montag & R. Ali (Eds.), *The Impact of Artificial Intelligence on Societies* (pp. 9–24). Springer. https://doi.org/10.1007/978-3-031-70355-3_2

Schlicker, N., Baum, K., Uhde, A., Sterz, S., Hirsch, M. C., & Langer, M. (2025). How do we assess the trustworthiness of AI? Introducing the trustworthiness assessment model (TrAM). *Computers in Human Behavior*, *170*, 108671. https://doi.org/10.1016/j.chb.2025.108671

Schmager, S., Pappas, I. O., & Vassilakopoulou, P. (2025). Understanding human-centred AI: A review of its defining elements and a research agenda. *Behaviour & Information Technology*, 1–40. https://doi.org/10.1080/0144929X.2024.2448719

Schrills, T., & Franke, T. (2023). How do users experience traceability of AI systems? Examining subjective information processing awareness in automated insulin delivery (AID) systems. *ACM Transactions on Interactive Intelligent Systems*, *13*(4), 1–34. https://doi.org/https://doi.org/10.1145/3588594

Schrills, T., & Franke, T. (2026). A model of integrated information processing in human-AI interaction. https://doi.org/10.48550/ARXIV.2606.07283

Schrills, T., Gruner, M., Peuscher, H., & Franke, T. (2023). Safe environments to understand medical AI – Designing a diabetes simulation interface for users of automated insulin delivery. In V. G. Duffy (Ed.), *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management* (pp. 306–328, Vol. 14029). Springer. https://doi.org/10.1007/978-3-031-35748-0_23

Sendhoff, B., & Wersing, H. (2020). Cooperative intelligence – A humane perspective. *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*, 1–6. https://doi.org/10.1109/ICHMS49158.2020.9209387

Shirado, H., Kasahara, S., & Christakis, N. A. (2023). Emergence and collapse of reciprocity in semiautomatic driving coordination experiments with humans. *Proceedings of the National Academy of Sciences*, *120*(51), e2307804120. https://doi.org/10.1073/pnas.2307804120

Shukla, P., Bui, P., Levy, S. S., Kowalski, M., Baigelenov, A., & Parsons, P. (2025). De-skilling, cognitive offloading, and misplaced responsibilities: Potential ironies of AI-assisted design. *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3706599.3719931

Sidji, M., Smith, W., & Rogerson, M. J. (2024). Human-AI collaboration in cooperative games: A study of playing codenames with an LLM assistant. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CHI PLAY), 1–25. https://doi.org/10.1145/3677081

Siu, H. C., Peña, J. D., Chen, E., Zhou, Y., Lopez, V. J., Palko, K., Chang, K. C., & Allen, R. E. (2021). Evaluation of human-AI teams for learned and rule-based agents in Hanabi. https://arxiv.org/abs/2107.07630

Siu, H. C., Peña, J. D., Zhou, Y., & Allen, R. E. (2025). In pursuit of predictive models of human preferences toward AI teammates. https://doi.org/10.48550/arXiv.2503.15516

Stattkus-Fortange, B., Attig, C., Wiebel-Herboth, C., & Franke, T. (2026). Understanding the charging behavior of electric vehicle drivers on long-distance trips – The roles of range regulation and human-automation cooperation. *Transportation Research Part F: Traffic Psychology and Behaviour*, *120*, 103610. https://doi.org/10.1016/j.trf.2026.103610

Stein, J.-P., Messingschlager, T., Gnambs, T., Hutmacher, F., & Appel, M. (2024). Attitudes towards AI: Measurement and associations with personality. *Scientific Reports*, *14*(1), 2909. https://doi.org/10.1038/s41598-024-53335-2

Tang, P. M., Koopman, J., Mai, K. M., De Cremer, D., Zhang, J. H., Reynders, P., Ng, C. T. S., & Chen, I.-H. (2023). No person is an island: Unpacking the work and after-work consequences of interacting with artificial intelligence. *Journal of Applied Psychology*, *108*(11), 1766–1789. https://doi.org/10.1037/apl0001103

Tey, K. S., Mazar, A., Tomaino, G., Duckworth, A. L., & Ungar, L. H. (2024). People judge others more harshly after talking to bots. *PNAS Nexus*, *3*(9), pgae397. https://doi.org/10.1093/pnasnexus/pgae397

Tomasello, M., Melis, A. P., Tennie, C., Wyman, E., & Herrmann, E. (2012). Two key steps in the evolution of human cooperation: The interdependence hypothesis. *Current Anthropology*, *53*(6), 673–692. https://doi.org/10.1086/668207

Tuckman, B. W. (1965). Developmental sequence in small groups. *Psychological Bulletin*, *63*(6), 384–399. https://doi.org/10.1037/h0022100

Van Der Laan, J. D., Heino, A., & De Waard, D. (1997). A simple procedure for the assessment of acceptance of advanced transport telematics. *Transportation Research Part C: Emerging Technologies*, *5*(1), 1–10. https://doi.org/10.1016/S0968-090X(96)00025-3

Walliser, J. C., Mead, P. R., & Shaw, T. H. (2017). The perception of teamwork with an autonomous agent enhances affect and performance outcomes. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *61*(1), 231–235. https://doi.org/10.1177/1541931213601541

Walton-Rivers, J., Williams, P. R., Bartle, R., Perez-Liebana, D., & Lucas, S. M. (2017). Evaluating and modelling Hanabi-playing agents. *arXiv preprint arXiv:1704.07069*. https://doi.org/10.48550/ARXIV.1704.07069

West, S. A., Griffin, A. S., & Gardner, A. (2007). Evolutionary explanations for cooperation. *Current Biology*, *17*(16), R661–R672. https://doi.org/10.1016/j.cub.2007.06.004

Wolf, V., & Maier, C. (2024). ChatGPT usage in everyday life: A motivation-theoretic mixed-methods study. *International Journal of Information Management*, *79*, 102821. https://doi.org/10.1016/j.ijinfomgt.2024.102821

Wu, S., Liu, Y., Ruan, M., Chen, S., & Xie, X.-Y. (2025). Human-generative AI collaboration enhances task performance but undermines human's intrinsic motivation. *Scientific Reports*, *15*(1), 15105. https://doi.org/10.1038/s41598-025-98385-2

Yu, Y., Yetim, M. A., & Cocieru, O. (2026). Whose voice is it anyway? Understanding AI customization and responsibility attribution in human-AI collaboration. *International Journal of Human–Computer Interaction*, *42*(14), 11015–11044. https://doi.org/10.1080/10447318.2025.2588652

Zakaria, N. J., Shapiai, M. I., Ghani, R. A., Yassin, M. N. M., Ibrahim, M. Z., & Wahid, N. (2023). Lane detection in autonomous vehicles: A systematic review. *IEEE Access*, *11*, 3729–3765. https://doi.org/10.1109/ACCESS.2023.3234442

Zhang, X., Zhang, Z., Liu, Y., Xu, Z., & Qu, X. (2024). A review of machine learning approaches for electric vehicle energy consumption modelling in urban transportation. *Renewable Energy*, *234*, 121243. https://doi.org/10.1016/j.renene.2024.121243