Human-AI Collaboration in Cooperative Games: A Study of Playing Codenames with an LLM Assistant

MATTHEW SIDJI, University of Melbourne, Australia
WALLY SMITH, University of Melbourne, Australia
MELISSA J. ROGERSON, University of Melbourne, Melbourne

Playing partial information, restricted communication, cooperative (PIRCC) games with humans have proven challenging for AI, due to our reliance on social dynamics and sophisticated cognitive techniques. Yet, recent advances in generative AI may change this situation through new forms of human-AI collaboration. This paper investigates how teams of players interact with an AI assistant in the PIRCC game *Codenames* and the impact this has on cognition, social dynamics, and player experience. We observed gameplay and conducted post-game focus groups with 54 participants across ten groups. Each group played three rounds of *Codenames*, with an AI assistant supporting Cluegivers. We found the AI assistant enhanced players' convergent and divergent thinking, but interfered with formation of team mental models, highlighting a tension in the use of AI in creative team scenarios. The presence of the AI challenged many players' understanding of the "spirit of the game". Furthermore, the presence of the AI assistants weakened social connections between human teammates, but strengthened connections across teams. This paper provides an empirical account of an AI assistant's effect on cognition, social dynamics, and player experience in *Codenames*. We highlight the opportunities and challenges that arise when designing hybrid digital boardgames that include AI assistants.

CCS Concepts: • **Human-centered computing** → *Collaborative interaction*.

Additional Key Words and Phrases: human-AI Interaction, boardgames, Human-Computer Interaction, teaming, human-AI Teaming, theory of mind, boardgames

ACM Reference Format:
Matthew Sidji, Wally Smith, and Melissa J. Rogerson. 2024. Human-AI Collaboration in Cooperative Games: A Study of Playing Codenames with an LLM Assistant. *Proc. ACM Hum.-Comput. Interact.* 8, CHI PLAY, Article 316 (October 2024), 25 pages. https://doi.org/10.1145/3677081

1 Introduction

Artificial Intelligence (AI) has long been used in gaming. AI agents can take on many roles such as non-player characters (NPCs) or adversaries, and advancements in their capabilities have seen them defeat professional players in boards games like Chess, Go and Poker and digital games like Starcraft 2 and DOTA 2 [11, 12, 16, 22, 53]. Despite achieving superhuman performance in competitive games, AI still struggles to achieve near human levels of performance in cooperative games. Particularly difficult for AI are the highly situated and often idiosyncratic communications that characterise games like Hanabi and Codenames, what we call PIRCC games, standing for Partial Information (PI) and Restricted Communication (RC) Cooperative (C) games [18, 31, 38]. Yet, recent advances in the capabilities of Large Language Models (LLM) may enable generative AI to participate successfully in PIRCC games as collaborative assistants to human players [61].

Authors' Contact Information: Matthew Sidji, matthew.sidji@unimelb.edu.au, University of Melbourne, Melbourne, Victoria, Australia; Wally Smith, University of Melbourne, Melbourne, Victoria, Australia, ; Melissa J. Rogerson, University of Melbourne, Melbourne, Victoria, Melbourne.

ACM 2573-0142/2024/10-ART316
https://doi.org/10.1145/3677081

AI in cooperative games is of interest to many fields including AI, HCI, and games research communities. The human-AI teaming and collaboration literature has been concerned with AI's effect on cognition including mental models, creative thinking and theory of mind [2, 17, 60]. Yet, in the context of games, studies often involve human-AI dyads which forgo group dynamics [23]. Often the games used are made purely for research purposes which prioritise controlled environments over naturalistic settings, or do not involve actual human-AI interaction in games at all, instead asking participants about a hypothetical game AI [30, 57, 78]. Work on hybrid digital play has raised concerns around the social effect of technology within physical games, highlighting a danger of disrupting the typical social dynamics [48]. Investigations into social practices within PIRCC games have shown the importance of social roles for humans performing well in these settings, which raises the question of how introducing an AI agent might affect social dynamics [54]. Games research has a wide body of work on AI teammates and NPCs, but this has typically focused on purely digital games [1, 17, 64, 72]. With the rise of hybrid digital games, there is a clear gap in our understanding of how AI agents will affect player experience and social dynamics in hybrid digital board games.

In order to address these gaps our study aims to investigate three research questions: Firstly, how does an assistive AI agent affect players' cognition within a cooperative game, including their creative thinking, theory of mind reasoning, and the formation of mental models? Secondly, what effect does the agent have on social interactions among players? Do players ascribe certain social roles and identities to the agent? Thirdly, how does the presence of an AI agent impact each player's experience of the game?

As we elaborate in section 2.1 we chose to investigate this within the context of the PIRCC game *Codenames*, as it is currently a test bed for LLMs and has been used as the basis for previous studies on human AI teaming [5, 18].

In the present study, participants (N = 54) were split into ten play groups ranging from four to six players. Participants either played with others they already knew or with strangers. Participants played three rounds of *Codenames* where the human Cluegiver had access to an assistive AI agent (referred to in this paper as the Agent) that could help them generate clues. Our findings reveal that the Agent had notable effects on the players' performance during the game as well as the team dynamics through: 1) Assisting Cluegivers' convergent and divergent thinking while interfering with their theory of mind reasoning, 2) Challenging what players view as within the spirit of the game, 3) Altering the player experience of the game through humor, cognitive housekeeping, and distractions, and 4) Reconfiguring team dynamics within and between teams. This study's contributions are as follows. Firstly, it provides an empirical account of player experience while cooperating with AI in a hybrid boardgame. Secondly, the paper outlines the benefits and challenges faced when incorporating AI into physical boardgames, which could help game designers understand the opportunities and drawbacks of incorporating AI agents into their games. Thirdly, we show how players augment their thinking in order to collaborate with AI agents in games. Finally, the paper describes AI's effect on existing team dynamics as well as the formation of new ones.

2 Related research

In this section, we review the current literature on AI in PIRCC games and human-AI teaming and collaboration. We are not only interested in what we can learn about human-AI teaming through investigating games of *Codenames* with AI, but also what unique game experiences and design opportunities AI brings to PIRCC games such as *Codenames*. As such, we also review recent work in the field of hybrid play, the study of physical boardgames with digital components.

2.1 AI in PIRCC games

Games have always been test beds for exploring AI abilities. From traditional boardgames like Backgammon, Chess and Go to more recent digital games such as Starcraft 2 and DOTA 2, defeating professional human players in these games has been a goal for many AI researchers [14, 16, 44, 55, 63]. While there is controversy on whether these examples constitute fair comparisons between Human and AI intelligence [15] many view these games as "solved" with AI having defeated top professional players. Accordingly, research focus has shifted to a different class of game.

PIRCC games have recently gained interest as a new challenge space for AI [5, 8, 75, 77]. This is not only because AI reasoning is notoriously difficult in these environments [42] but also because games are an ideal setting to investigate human-AI interaction, particularly with neural networks like LLMs [79]. This is because games provide a low-stakes environment, leading to more player experimentation [66], while also providing a system and narrative context giving players a local set of social norms and feedback [26]. PIRCC games are also a widely popular genre of boardgame with 450 entries on Boardgamegeek.com classified under "Communication Limits" and "Cooperative" or "Team based" [21]. Well-known examples include: *Hanabi, Pictionary, Codenames, The Crew: The Quest for Planet Nine, The Mind, Decrypto, Letter Jam, Just one, and Taboo* [3, 6, 9, 18, 20, 29, 56, 59, 71].

HCI researchers have been using game AI to investigate aspects of player-AI interaction. Villareale et al. created a framework for how players form mental models of AI agents based on observations of players in an AI-based game [65]. While this was done in an adversarial context, mental model formation is also important for understanding human-AI cooperation [7]. Game AI has been used to simulate player experience and game difficulty, a task typically done by humans [25, 50]. In another study, Ashktorab et al. investigated human-AI collaboration in a language-based PIRCC game. They found that when players perceive their teammates to be humans they find them more likeable, intelligent and creative regardless of any difference in their overall game performance [5]. Yet, it is not clear how their results might generalise to a hybrid human-AI team where a human has to team with another human using an AI agent.

The PIRCC game Hanabi [9] has seen particular interest from researchers due to it requiring sophisticated levels of theory of mind reasoning - reasoning about another agents beliefs, intentions and desires - an important aspect of many well-functioning teams [8, 9]. Unfortunately, Hanabi agents still perform extremely poorly when paired with human teammates [31]. This is perhaps due to humans' utilisation of externalised cognition, social roles and rules flexibility [54]. Yet, recent advances in the capabilities of LLMs show promising performance in language-based PIRCC games. LLMs such as ChatGPT are showing breakthrough performance in the popular language-based PIRCC game *Codenames* [61]. This is a marked improvement over the Reinforcement Learning (RL) based models which, when paired with humans, do not show performance above that of random guessing [36, 58]. This breakthrough performance of LLMs, along with showing some signs of rudimentary theory of mind [37] reveals an opportunity for investigating human-AI teaming in *Codenames*.

2.2 Hybrid play

Hybrid play is an emerging area of study of particular interest to the CHI PLAY community. Hybrid play is the study of physical boardgames with digital components [34]. In investigating the effects on player experience of different levels of automation in tabletop games, Wallace et al. found that while it provided advantages to players, it detracted from enjoyment, game state awareness, and flexibility [68]. There is a significant risk of digital components detracting from the social and material experience of the game: game designers have expressed concern that digital components may reduce the enjoyable social dynamics [48]. The importance of maintaining sociability is highlighted by Kankainen & Paavilainen through their hybrid play design guidelines which also emphasize designing hybrid play to support accessibility [34].

Rogerson et al. present a model for understanding hybrid boardgame play by categorizing the functions of the digital elements in each game [49]. In this model, digital components perform the functions of either: Teaching, Timing, Calculating, Randomising, Remembering, Housekeeping, Storytelling or Informing. The authors call for more research into the assessment of player experience in hybrid games, as well as whether and how hybridity contributes to accessibility.

While hybrid play is a burgeoning area of research, current hybrid games tend to limit the focus of their digital components to board set or narration [39, 40]. Furthermore, although the modification of existing boardgames to include digital components to study hybrid play is well supported in the literature [40, 41] there is currently little investigation into the introduction of AI as a form of hybrid play. From human-AI collaboration literature, we know that AI can have wide-ranging effects on social dynamics within a group [62], yet it is unclear what impact this will have on the player experience and social dynamics in a hybrid digital game.

2.3 Human-AI teaming and collaboration

Both human-AI teaming and collaboration fall under the broader research area of human-AI cooperation, which aims to help humans and machine agents find ways to improve joint welfare [19]. Human-AI teaming emphasizes close cooperation between human and machine agents. In this setting, the human and AI share a goal and equally use each other in order to achieve it [78]. Much of the human-AI teaming literature has focused on aspects of cognition; in particular, how we form and improve our mental models of AI teammates [2, 7, 10, 35, 66], as an accurate mental model is crucial to high team performance [51]. Using a think-aloud study in a word association game, Gero et al. measured players' mental model of AI teammates, and showed a positive correlation between accuracy of the mental model and game score [23]. Formation of these mental models has been shown to be easier when AI agents "fail in predictable ways" [7]. Humans not only form more accurate mental models when AI behaves in a consistent way, they also expect AI agents to be more consistent and perform to a higher quality than humans in certain tasks [35]. Siu et al. provide further support for humans' preference for consistent AI teammates in their investigation of human teaming with either a rule or learning based AI agent. They find that humans generally prefer the more predictable rule based agent over the learning based one, regardless of any difference in performance [57].

Human-AI collaboration is often studied using digital assistants, where one or more humans are working on a task using an AI assistant [70]. The creative aspects of cognition have been a focus for human-AI collaboration research with studies in creative domains ranging from music composition, creative writing, visual design and level design for games [28, 62, 67, 69]. In tests with practising videogame designers, Guzdial et al. found that designers varied in their desired levels of interaction with the collaborative AI agent, yet most found it valuable to their practise [28]. During multi-person creativity tasks, AI has been shown to have a positive social effect. In a music composition task, Suh et al. found that AI seeded common ground between human collaborators, acted as a psychological safety net, mitigated stalling and was a force for forward progress. Interestingly, poor technical performance from the AI seemed to strengthen the social relationship between the participants [62]. This finding was supported in an investigation of human-AI collaboration in prewriting with LLMs [69]. This study found that in the ideation phase LLMs tended to take the initiative, and that unexpected results from the AI were good sources of inspiration and divergent thinking. The researchers actively discourage designing for reduced randomness in AI responses to promote creativity. Another study by Hwang and Won. investigated whether perceived identity and conversational style of the AI would moderate the creative outcome of human-AI collaborators in a creative writing task [32]. They found that participants contributed more ideas and ideas of higher quality when thinking their teammate was an AI and that participants with high anxiety in group communication reported greater creative self-efficacy in task performance. This body of work seems to indicate that AI collaboration in creative domains has strong positive social, creative, and psychological consequences, and that unexpected results from AI agents benefit divergent thinking [24]. Language games, such as *Codenames*, have been presented as the ideal setting for studying creativity and provide further opportunity for investigating human-AI collaboration [18, 60, 80].

This literature suggests a tension between the encouragement of randomness and unpredictability from human-AI collaboration research on creativity, and the need for predictability to form accurate team mental models in the human-AI teaming literature. To our knowledge, there is little research on what happens when you combine these settings, where humans must work together in a team with some humans simultaneously collaborating with an AI agent. This configuration of a human-AI cooperative group is not unlikely to occur in the future and hence warrants further investigation [19].

3 Method

In this study, we aimed to understand how player experience, team dynamics and reasoning were affected by the introduction of an AI clue-giving tool. To do this we used qualitative focus groups as well as observations of game sessions to explore both individual participants and the group's attitudes and opinions towards the AI agent within the game. Conducting data collection immediately after game sessions allowed participants to reflect on experiences with higher accuracy and while they were still in the context of the study.

3.1 Codenames game rules

Codenames is a game for four to eight players in which two teams compete to be the first to guess all their team's secret words [18]. In front of all players is a five by five grid of 25 words. These words are secretly assigned as either the red or blue team's words, with the remainder assigned as neutral words and a single "assassin" word. Each team comprises of one Cluegiver and one or more Guessers. The Cluegivers can see which words belong to their team and give their Guessers one word clues to help them identify their words. Clues are accompanied by a number, which tells the Guessers how many of the words the clue relates to (for example: "space, two"). Cluegivers take turns giving clues and having their team guess. The game ends when all of one team's words have been guessed, in which case they win, or the assassin is picked, resulting in a loss.

3.2 Study set up

From September to December 2023, we conducted ten in-person *Codenames* sessions with a total of 54 participants. Each session contained two teams of two to three players for a total of four to six participants each session. A session consisted of three games of *Codenames* followed by a 30 minute focus group. In each game, the Cluegiver had access to the AI clue-giving agent on a laptop in front of them. In accordance with how Codenames is typically played, Cluegivers sat across from Guessers to ensure the secret key and the suggested AI clues were private. Cluegivers were asked to use the AI to generate clue suggestions for at least the first three turns in each game but were not required to give those suggestions as clues. After the third turn, Cluegivers were free to stop using the AI if they wished. It should be noted that all but one participant chose to continually engage with the AI after the initial 3 rounds. After each round, the Cluegiver was asked to swap with another member of their team so all participants had at least one game as Cluegiver with the AI agent. A diagram of the set-up can be seen in Figure 1.

Fig. 1. Cluegivers (foreground) formulating clues for their Guessers across the table.

Following the game session, a researcher facilitated a focus group by asking participants open-ended questions. The questions ranged from players' previous experience with AI, how they used the agent in this game, how it affected their self-perceived clue-giving skill, and how it affected how they worked together as a team. These questions were developed based on discussions after pilot testing and aimed to elicit responses from the participants that related to the research questions. Focus groups were recorded and then transcribed.

Since we were interested in comparing how the AI would affect new versus existing teams, we split participants into two conditions, strangers and friends. In the groups of strangers, none of the participants in a particular team had played with or met each other before the session. In the groups of friends, all members of a team knew each other, with most having played *Codenames* together multiple times. To preserve as natural a setting as possible, participants were encouraged to play with any house rules they would normally employ. These included using pen-and-paper to write down previous clues, utilising the timer, or not allowing reminders of previous clues.

Players were recruited through local boardgame groups, as well as advertising on Facebook and X (formerly Twitter). All participants were required to have some level of experience with *Codenames* and be over 18 years old. Ethics approval for the study was obtained from the University of Melbourne. Participants were remunerated with $20 gift vouchers.

3.3 Analysis

4 AI agent design

In this section, we detail the design of our *Codenames* Cluegiver agent, which can be seen in Figure 2. The Cluegiver agent generates possible candidate clues given a set of target-words and avoid-words typed in by the human player. It attempts to generate a clue that connects to some or all the target-words while avoiding associations with the avoid-words.

The AI model used is the Generative Pre-Trained Transformer (GPT) which is a class of attention-based language model, specifically, OpenAI's GPT-3.5 model [45]. The choice of an LLM was made due to their superior performance over reinforcement learning (RL) agents[33, 61]. The model was trained using few shot Chain of Thought prompting, which has been shown to increase the model's reasoning capabilities in common sense and symbolic reasoning tasks [13, 73]. This involves providing the model with example prompts and answers to train it for new prompts from participants. This technique encourages the model to write out each step of its reasoning. In the case of formulating a *Codenames* clue, this involved first coming up with a candidate clue, then

Fig. 2. Screenshot of the Agent interface

going through and listing how it relates to each target word while avoiding each unwanted word. We chose to show this explanation step to users as players from our pilot testing found it helpful for clue generation. For example:

> **User:** *"Give me some clues that relate the following target words, but avoids the unwanted words. Target words: tree, page, book. Unwanted words: material."*
> **Bot:** *"Here are some clues I would give.*
> *1. 'LEAF 3'. Leaf relates to tree as some trees have leaves on their branches. Leaf also relates to page and book as a reader can leaf through the pages of a book. The clue LEAF avoids relation to the word material, as even though leaves can be used as a type of material the clue more strongly relates to the target words tree, page and book. I avoided suggesting a word like PAPER as that would relate to the concept of material.*
> 2. 'READ 2'. Read refers to page and book as one can read both a page and a book. It avoids association with the word material."

The temperature of the model was set to 1.0 in order to encourage divergent results. Temperature is a measure of the amount of randomness in the result [46]. This allowed participants to regenerate clues given the same input and acquire a potentially different result.

5 Findings

Through analysis and coding of the focus group transcripts along with observations during game sessions, we generated four broad themes concerning how the Agent affected players' cognitive abilities, the "spirit of the game", how play experience was altered, and the effects on social dynamics.

5.1 AI augments Cluegivers' creative thinking but interferes with team cognition

One of the most striking findings was that although only the Cluegivers had access to the Agent, it affected the reasoning of all players throughout the games. Cluegivers utilised the AI to augment their creativity by enhancing their divergent and convergent thinking capabilities. Despite the Agent's beneficial effect on cluegiving it seemed to have an overall negative effect on both Cluegivers' and Guessers' ability to do theory of mind reasoning and form mental models of teammates.

Explanations provided by the Agent persuaded Cluegivers to give tricky clues against their better judgement, and Guessers needed to consider if a clue was majority AI or human-generated, leading to frustration.

5.1.1 The Agent enhanced players' convergent and divergent thinking abilities.

> *"A lot of the answers were absolutely horrendous, but I feel like it got my brain thinking in a different way...it's like a plus one to what your brain is already"* (P3 G1)

Clue generation is a creative multi-step process with different stages involving convergent and divergent thinking. Cluegivers' divergent thinking abilities were aided through using suggestions from the Agent as starting points to further iterate upon. While it was rare that Cluegivers simply accepted the Agent's suggestions, most used the suggestions as starting points to generate improved clues; *"I was like, "that doesn't work" And then I think, "why doesn't that work?" And then I'd have an answer for a clue."* (P3 G4)

Beyond generating a clue given a set of target words, Cluegivers also needed to use convergent and divergent thinking when trying to group a certain subset of target words that are likely to relate to a good clue. While some players used the Agent to find a clue given a connection that they had already identified – *"If I saw words that were semi-related I would put those in and ask it to relate those for me."* (P5 G1) – others refined clues suggested by the Agent based on how that participant thought their team would make connections.

> *"Sometimes I picked basically what the AI said... like in the first round it gave me NATURE for three, then I thought, "they might not get that for scale" which was connected to that, so I changed it to TREE, for just bark and trunk."* (P4 G1)

Here we have the Agent assisting in the process of coming up with a range of clues given a set of target words (divergent thinking) and also Cluegivers refining a clue given by the Agent for a smaller set of words (convergent thinking).

Importantly the suggestions from the Agent didn't have to be good for them to be useful to many Cluegivers. In this example, the poor reasoning presented by the Agent was questioned by a player which in turn helped them formulate a clue;*"It did help me because sometimes I would see it and think "that's terrible" and then I would think of something good. That's how it helped."* (P3 G4). While in this case the process of interrogating the Agent's reasoning helped Cluegivers brainstorm, we will see later that this overwhelmed and frustrated others. At times, Cluegivers drew inspiration not from the suggested clue but from the accompanying explanation.

> *"My clue didn't come from the suggestion it came from it talking about the suggestion. It (the Agent) said something about performances happening on a stage, but it wasn't trying to say stage. But then I used stage."* (P3 G4)

The clue generation process is highly creative and iterative, involving both divergent and convergent thinking. Considering all the different possible combinations of words to clue and then selecting candidate words to associate relies highly on divergent thinking, which in *Codenames* can help to alleviate fixation on specific combinations of words. Once a player has selected a set of words to convey through a clue, converging on the specific clue with just the right level of association between targets and semantic distance from unwanted words requires players to utilise convergent thinking. Here we can see that both types of thinking have been aided in some way by the presence of the Agent. Yet, generating good clues is not enough to have a *Codenames* team perform effectively. Forming mental models of teammates is crucial for knowing whether a given clue will be guessed correctly. In the next section we will detail how the Agent affected players mental model formation.

5.1.2 The Agent interfered with the formation of Cluegivers' team mental models. The presence of the Agent drew Cluegivers' attention away from forming mental models of their teammates, instead focussing them on forming mental models of how the Agent would give clues.

> *"I already knew how your mind worked making connections, where you mentioned something and the inverse of it...so yeah very quick to learn as a team without the AI."* (P3 G5).
>
> *"It's more to think about and maybe less enjoyable because you don't get to know as much how the other person thinks"* (P1 G4)

Interrupting players' formation of mental models of their teammates detracted from the enjoyment of the game. Many participants stated that this was one of their favourite parts of playing *Codenames*. Even players in the friends setting, who presumably already had mental models of their teammates, found it difficult to give clues tailored to their teammates because of the presence of the Agent.

> *"I feel I usually do give clues tailored to my teammates, but because I was focused on using the Agent, I wasn't thinking about that side of the game as much."* (P3 G1)

Having accurate mental models of teammates is an important aspect of the game for both Cluegivers and Guessers. Having to also form mental models of the Agent seems to increase the overall cognitive work in the game.

5.1.3 The Agent's explanations increased likelihood of tenuous clues. At times, the Agent's explanations increased the confidence with which a Cluegiver gave difficult or tenuous clues. In doing this, the Agent swayed Cluegivers to give clues in line with what it had suggested, but not necessarily with what Guessers might be able to interpret easily.

> *"The AI said GAME for the word tag. I didn't take the suggestion. But it did make me think that everyone knows the game tag. I feel I couldn't figure out how to say it in one word. So I came up with IT, trusting that people would know."* (P3 G4)

Due to the Agent's suggestion, this player was primed to interpret TAG as the children's chasing game, and incorrectly asserted that their Guessers would as well. In another game, this convincing quality of the explanation led to frustration between two teammates who had played many games together before.

> *"Just general frustration that I could obviously see an explanation of someone giving me the answer on the association and (P1) couldn't see it. That was really frustrating because I was like 'that's so obvious' "* (P2 G1)

P1 responded to this with frustration, stating that *"P2 you can see the link because you've read it."* (P1 G1). This interaction went on to trigger a larger discussion in the group about the merits of explanations for clues. Some participants still felt explanations were helpful in coming up with their own clues.

> *"Their [the Agent's] justification didn't make a lot of sense, but at least I was like 'OK I can sort of see your train of thought'...I feel like it helped me understand the ways other people would think."* (P2 G1).

Others felt that the explanations could sometimes convince Cluegivers to give clues that would be difficult for Guessers to decipher without the provided explanation.

> *"P2 shouldn't need to be told why ANCIENT 2 is a good clue. Codenames should be an intuitive game, so [the Agent] trying to convince you that the clue is intuitive, I don't think is intuitive."* (P1 G1)

Players pointed out that Guessers lack access to explanations when given clues so letting the explanation determine whether Cluegivers give a clue is a path to failure.

> *"The clue should be enough to connect it to the words. Otherwise, how can you expect them to use that word to make the connection?" (P5 G1)*

While explanations provided by the Agent served as useful starting points for many players, they also increased Cluegivers' confidence in their Guessers' ability to decipher difficult clues. This sometimes led to Cluegivers giving overly complex clues, which they otherwise might not have said.

5.1.4 The Agent affected Guessers' theory of mind. Surprisingly, the presence of the Agent affected the theory of mind reasoning abilities of Guessers. Some reported having to perform a filtering step when receiving a clue to determine whether the clue originated from the Agent or the human.

> *"I found myself almost guessing how much of that (clue) was my partner and how much of it was the Agent."* (P4 G5)

Once Guessers formed a belief as to where the majority of the clue was formulated, they applied their own mental model of how the Agent gives clues. Yet, while participants had a strong sense of the types of clues the Agent tended to give, determining whether or not that clue was from an AI was often completely random. For some Guessers the Agent added a level of uncertainty when guessing which led to an unenjoyable experience:

> *"I would enjoy playing it where you were forced to completely use it or not use it at all. For example, the players are just using the Agent and then we're just guessing what the Agent is saying [. . .] it makes it an equal game, I guess this is sort of a weird mix of like you can use it or not use it."* (P4 G4)

As we have shown in this theme, the Agent had wide-reaching effects on how all players, Cluegivers and Guessers, reasoned throughout the game. The Agent aided Cluegivers' convergent and divergent thinking, assisting them in generating clues. Yet the presence of the Agent also weakened players' mental models of their teammates. The explanations provided by the Agent also increased Cluegivers' confidence in tricky clues, which otherwise might not have been given. Finally, Guessers now had to consider how much of a clue was AI generated which led to frustration.

5.2 How the Agent affected the Spirit of the Game

Participants often spoke about how the Agent affected what they viewed of as *"The spirit of the game"*. This spanned from how to balance AI competence to maintain challenge, players' opinions on whether using the Agent was cheating, what players most enjoy about playing *Codenames*, and feelings of psychological safety in play.

5.2.1 Desired level of AI competence. Participants felt there was a fine line that the Agent had to toe between being so good that it defeats the challenge of the game and so poor that it no longer contributes anything to play. For some, the performance of the Agent was poor enough to cause frustration - *"Yeah, it was helpful early like just because there's so many things but as I went on, it was more frustrating,"* (P6 G6) - and even resulted in some participants choosing not to use it at all; *"I just stopped using it at some point, because [the Agent's] answers are not where my mind would go."* (P4 G4). Yet, for others this poor performance contributed heavily to their enjoyment:

> *"I feel like if it was any better it would have been less fun. Because then you can't come up with fun little relations and stuff like that. I think the reason it made it more fun was because the Agent was quite bad."* (P6 G2)

Equally, many other participants spoke about how improving the Agent's performance would detract from the game, because it would ruin the challenge; *"If it was too good, then it would almost take away from the role of the Codemaster [Cluegiver]. I would be like, why is the Codemaster there?"*

(P2 G2). Likening an improved AI to a “Chess engine,” one participant stated that, "*I think if the engine (AI) was working very, very well, it would probably destroy the game*" (P4 G5).

Participants also raised concerns about balancing the level of agency the Agent possesses. For some, an AI with full control over what clues to give would have added to the hilarity of the game; *"I think it'd be hilarious [...] it would be the best way to lose"* (P3 G1). For others, it would have been frustrating, inspiring a sense of dissatisfaction with winning; *"You wouldn't feel like you'd won"* (P2 G1) and *"It would be frustrating to play with because there's so many clues where it's just a leap of logic."* (P4 G1). Overall, participants had diverse views about the level of competence and agency they desired in the Agent.

5.2.2 Whether the Agent was considered cheating. Participants had varying strong opinions on whether the use of the Agent should be considered a form of cheating. Some participants did not consider it cheating as its effect is symmetrical across the teams (i.e., both teams have access to the Agent):

> *"I felt like the tool was good for that common knowledge [...] I Google things sometimes because I don't know what they are and like I know sometimes that feels like cheating, but I like to put everyone on an equal footing"* (P3 G4)

Many participants were in support of technology used as an equalizing tool, often comparing the Agent to the use of Google; "*I feel like it was an aid for the game, less of a cheating mechanic. It's almost like if you don't know what a word meant and you had to Google it.*" (P2 G5). While most participants were *"totally okay"* with using Google - *"I think you're allowed to Google the definitions of words"* (P1 G1) - one in particular preferred the Agent over Google. This was because they felt the Agent was more appropriate for the task as well as having the sole purpose of wanting you to give a good clue:

> *“I am more anti-Googling and more pro AI. Maybe it's because I use AI a lot in my work. I find the AI is far more effective at delivering what I need it to deliver. [...] it also doesn't lead you astray. It is very much trying to give you (a good clue)"* (P5 G3)

Another participant strongly considered the use of the Agent to be cheating, as it changed what they consider the core challenge of *Codenames*: making connections between words. For this participant, the quality of the clues given had no impact on whether they considered the use of the Agent cheating:

> *"To actually have a device that connects it for you. That's the whole point of the game [...] it's taking away the point of the game, the challenge. [...] Even if it were better or worse, using a tool to come up with words to connect the things, that's the point of the game"* (P3 G5)

Yet, for another participant, whether the Agent was considered cheating heavily depended on how good the Agent was at giving clues. They suggested that if it was more *"chess-engine like"* i.e. more effective; *"there's almost no point to playing"*. Yet, because of its low-quality clues it *"added a different level of hilarity to it."* (P6 G5)

5.2.3 The Agent's effect on why people play Codenames. Participants had various reasons they enjoyed playing *Codenames*, which were reflected in their style of play. Consequently, they had differing opinions on the appropriateness of an AI tool in the game. For some participants, the act of forming, delivering and then having their teammates correctly guess complex clues was what they enjoyed most:

> *"I'm not sure if I believe in its ability to make such random lengthy links, which is how I play with my friends. That's the real core of the game in my opinion... even with little phrases or inside jokes or references."* (P1 G1)

Because this player aimed to give "random lengthy links" they did not believe that the play style of the Agent was aligned with their own and therefore felt it detracted from the game; *"I don't believe in an AI's ability to do that. So I'd argue that it takes away from that aspect of the game."* (P1 G1). Another player would tailor their clues to their own guessing style; *"I was tailoring it to me. I was sitting there with the Agent being like "yeah we're on the same wavelength"* (P2 G3). Players that used this complex play style were not always concerned with giving the clearest clues. Rather, they enjoyed giving deep multilayered clues that, if guessed, would show a strong connection between the Cluegiver and Guesser. Surprisingly, even players who shared this complex play style had differing opinions on whether the Agent supported that play style. (P1 G1) said it took away from the aspect they enjoyed, while (P2 G3) felt the Agent supported their play style, with P3 in the same game stating *"You and the AI as a team were killer".*

In contrast to this play style, many others shared what one player termed an *"empathetic"* (P5 G3) play style. These players enjoyed learning how their teammates thought, and tailoring clues specifically for them; *"I just think the game is about playing the people [. . .] for me, it's about playing each other."* (P5 G3). Learning how other players thought was, for many players, what they found most fun; *"what I really like about the game is the insight into personal psychology"* (P4 G6) and *"part of the fun is like seeing how other people think you know, how they make connections"* (P6 G6). These participants all shared in the sentiment that the Agent detracted from this aspect. For those that preferred this aspect of *Codenames* they *"just don't see the value in it"* (P5 G3), saying it *"takes away that wealth of enjoyment"* (P6 G6). The participant who enjoyed *Codenames*'s *"insight into personal psychology"* noted that the Agent reduces the ability to give personalised clues because the Agent has the same clue suggestion style regardless of who is guessing; *"having the Agent kind of regulates that, but in a way it depersonalises it."* (P4 G6).

5.2.4 Psychological Safety. While all participants enjoyed playing the base version of *Codenames*, many shared the sentiment that giving clues was a solitary experience. For many, this was an aspect they did not enjoy about the base game as it led to feelings of isolation, and increased pressure to perform, which for some participants is a very stressful experience.

Following game 1, one participant declared that "*I can find Codenames a little stressful [. . .] because often I get a bit overwhelmed with how many choices there are.*" (P4 G1). Being in the role of Cluegiver, especially in the beginning where there are many words to consider, can often be an overwhelming experience, particularly because the Cluegiver is left alone to formulate a clue. This participant appreciated the Agent's presence in alleviating some of that initial stress; *"even if I could see a couple of the Agent options are wrong, it's easier to pick say 'out of these three this one is the best' and I can probably modify it [. . .] so it made it a little bit less stressful as the spymaster."* (P4 G1). Notably, just the presence of the Agent had this stress-relieving effect, regardless of the quality of its clues. This sentiment was shared by many, as they thought of the Agent as a sort of backup option; *"(it) was kind of a safety net, I guess for me having that assistant to do the thinking for a little bit"* (P4 G6).

Along with reducing stress, participants noted feelings of isolation as Cluegiver. Although some participants described playing the game with multiple Cluegivers in one team, the rules specify that this is a job for one player and prohibit any communication with Guessers outside of clue-giving:

> *"as the Cluegiver it is a very solitary thing, and that's not my favourite way to play a game. I much prefer a collaborative experience, so even if I'm doing it with something that I know is not a real human, I'm just like, yes, "we're in it together" "* (P3 G3)

Another participant in the same game shared this sentiment; *"As the Cluegiver, you are playing a solitary game and if you're not able to share that with someone, being able to share it with some half-dumb robot is kind of fun."* (P2 G3). As well as reducing isolation, some participants felt comforted by the presence of the Agent and the confirmation it gave when its answers were similar to their own; *"It just felt reassuring, that I'm on the right track [. . .] For example, if (the Agent) gives a very similar answer I would think 'this is backed up' ."* (P4 G2). The Agent also served a scapegoat role for some participants where they could blame it for poorly received clues or mistakes they made; *"I'm going to blame the AI for my blunders"* (P1 G2).

5.3 AI changed the play experience

All groups reported that the Agent transformed the typical *Codenames* player experience. The Agent provided a lot of opportunity for humour, especially in groups of strangers. This often came about through ribbing the Agent for giving poor clues. The Agent also fulfilled many of the typical housekeeping actions such as keeping track of words, and raised the overall performance for most Cluegivers. Distraction and perception of extended waiting time was a common experience brought about by the Agent, specifically for Guessers.

5.3.1 The Agent invited humour and ribbing behaviour. The overall poor performance of the Agent was a humorous experience for most participants. One participant noted, *"as the game went on I think we all noticed that there was this of potential for humour in terms of things not making sense"* (P4 G6). Others pointed out that it *"added a different level of hilarity"* (P4 G5) than would otherwise be experienced in a typical game of *Codenames*.

This was potentially due to the mismatched expectations of AI's competence and its actual performance - *"It was actually below my expectation, that it did as poorly as it did. For something like word association, it makes sense for an AI to be able to nail that."* (P6 G2) - as well as the confidence with which the agent delivers its suggestions; *"even if it's nowhere near the right thing. It's always very confident and you can always feel its confidence."* (P2 G3). At the conclusion of most games, participants would go through some of the more ridiculous explanations of clues the Agent presented in the previous game. For example, (P4 G2) was particularly amused by the following explanation; *"I had Dragon and March and it said 'you should say FIRE because dragons breathe fire and the image of an army marching into battle will conjure fire.'"* (P1 G2)

Much of the humour came in the form of ribbing or making fun of the Agent; *"We were laughing at it."* (P2 G2) and *"I actually really enjoyed laughing at what it told me"* (P3 G1). One player pointed out that this ribbing behaviour is one aspect they enjoy about *Codenames* in particular; *"Half the fun of the game is making fun of people for their terrible clues."* (P4 G5). Being in the stranger setting they pointed out that they *"played the game more straight than I would with my friends. . . I mean we're all strangers, we wouldn't (rib each other), but with friends absolutely."* (P4 G5). Yet, when asked if the Agent introduced more opportunities to make fun of someone's clue they replied, *"I mean, there was plenty"* (P4 G5). This shows that the Agent has the potential to introduce an enjoyable ribbing aspect, which may be lacking when playing with strangers.

5.3.2 The Agent fulfilled cognitive housekeeping actions. We found that several of the 'friend' groups played *Codenames* with house rules. Several, in particular, used pen and paper to note down words and clues, allowing them to keep track of the 25 cards on the table. Managing words and keeping track of clues and guesses became a form of in-game housekeeping (see [49]) and replaced many of the typical housekeeping actions done by Cluegivers, in particular memorising your words. This work was eased by the presence of the Agent. Yet, depending on the stage of the game this had both positive and detrimental effects on cognitive load for Cluegivers.

Many players reported struggling with the beginning stages of the game due to the high volume of words a player had to internalize before they could make a clue. Simply the act of typing the words into the text box and having them present aided participants; *“Even having the words in the little box here helps a lot"* (P5 G1). This became especially helpful when participants were also trying to avoid associations with the opposing team’s words; *“One thing I struggle with is, I can hold all of our words in my head but I can’t hold their (other team) words in my head. So it’s really handy for that."* (P2 G3). Even the explanation text reminded participants that their clues needed to avoid unwanted associations; *"Then also in its explanation it’s like “ohh and it avoids these words by the way, just letting you know I’m avoiding these words” which reminds me, Oh yeah I’ve gotta be doing that"* (P2 G3). Many of these housekeeping actions were particularly useful in the early stages of the game, with participants identifying the Agent being most useful in that stage; *“[It’s] probably more useful at the start, honestly.”* (P1 G1).

Yet, despite the benefits the Agent provided through performing these housekeeping actions, it also had unintended drawbacks in terms of error recovery and acting as a crutch. Since it was up to participants to type their target words and avoid words, any typos would lead to clues given for incorrect words:

> *"Yeah, I realised I made a typo so I wasn’t actually excluding that word...But I guess that was me looking at it wrong and that would have happened in person too, but it would have been corrected straight away in person. But here it wasn’t corrected because I didn’t think to change it [in the Agent].”* (P3 G4)

In another game, the participants had accidentally swapped words, with each player typing the opposing team’s words into their Agent. Recovering from this error significantly slowed the pace of the game, as participants had to delete and retype each word. Interestingly, participants in this game pointed out that using the Agent in the early stages of the game made the later stages more difficult:

> *"in a sense like I was being lazy the first round. Whereas I find without the AI, first round’s really hard but it gets easier as you become familiar with the words [...] in [a typical Codenames] game, I struggle a bit initially, but I found that struggle and that acclimatisation meant it’s easier to find connections later on, whereas with this, you know, I just sort of turned my brain off"* (P6 G6).

In offloading the memorization of the words to the Agent, by the time it came to the last few words, Cluegivers had not internalized their words; *“you’re not reinforcing your own words”* (P3 G6). This led to the final turns of this game taking far longer than initial turns.

5.3.3 Raising the floor on clue quality. Participants spoke about how the Agent made the game feel easier by raising the minimum quality of clues given, but not necessarily helping them give their best clues. In other words, the Agent was most helpful for those who struggled with the game, rather than those who were already confident and looking to perform even better.

In discussing who the Agent might be most beneficial for, one participant said, *“I feel like [this would be useful for] people who can’t get started, it would be useful even if it gives you bad ideas [...] And if English is a second language or if you can’t think of those subtle connections."* (P3 G6). Another participant supported this sentiment, saying the Agent is most helpful when one is not performing at 100%; *"I think it’s great if you’re kind of not fully awake or if you’re not in the perfect mindset where everyone’s in the pocket, and ready to play this game. It’s better when you’re at a random time of the day you want a study break."* (P1 G1).

For Cluegivers, the beginning of the game is often one of the most difficult points; *“I think it was a good starting place for a train of thought, which was good because otherwise if I have to decide I get* *too frazzled"* (P5 G2). Due to the agent's usefulness at this stage many participants reported the game felt easier; *"I'm not sure how much better (I did) but I definitely found it kind of easier"* (P4 G1). This was a typical response when participants were asked if they felt the Agent *"made them better at the game"*. This indicates that the Agent helped participants perform poorly less often but had little effect on how often they perform above average, as shown by this participant in game six; *"If you're wanting to play your best game of Codenames it's not useful enough to be worth the trouble."* (P3 G6).

5.3.4 Increased perception of downtime can lead to distractions. Many participants mentioned feeling like there was more downtime and waiting for Cluegivers to use the Agent, even though Cluegivers generally felt that they were giving clues faster.

Guessers in game 5 all agreed that when the Cluegiver was using the Agent, the pace of the game slowed, with P1 saying that the *"(Cluegivers) were looking at it for a couple minutes and we both started to get very fidgety"* (P1 G5) and P2 *"The typing, that's the only thing that slowed it down a lot"* (P2 G5). This sentiment was shared in other games; *"On this end it's a little bit more waiting and it probably made it a little less enjoyable"* (P5 G4).

Many Cluegivers mentioned the Agent expediting their clue-giving ability. One participant, also from game 4, commented on how they were usually a slow Cluegiver; *"I feel like it helped me not be that person (the slow person) as much as I usually am"* (P6 G4). We can also see evidence of the Agent speeding up cluegiving in other sessions; *"I think it made me a faster Cluegiver not necessarily a better one. It allowed me to make connections between words faster."* (P3 G6). Yet, while Cluegivers generally felt they gave clues faster, one participant in game 7 pointed out that the game felt longer because there was more time reading rather than thinking:

> *"It felt slower, because even though I think I spent less time being stuck [...] it felt like you just spend a lot of time looking for an answer rather than thinking about an answer. I'll have an idea, I'll type it in and then I'm reading it. While I'm reading I'm not necessarily thinking [...] So I felt like not only the game took longer, but that I spent longer reading than actually playing the game."* (P1 G7)

This feeling of not being fully present in the game was spoken about in many other focus groups. Many participants, as Cluegivers or Guessers, reported feeling distracted by the agent:

> *"Waiting for people to type, it took me out of the game. I could watch you think and I'd happily be into that. But you just hearing you typing, I'm gone"* (P3 G5)

As a Guesser this player became distracted while hearing Cluegivers type their words. Interestingly it doesn't seem to be the actual time waiting for a clue that creates this feeling of distraction, as many Cluegivers were ultimately faster. Rather guessers felt their teammate's attention was away from the game as they typed, which led to a feeling of disconnection. This was further backed up by a participant from game 7. When speaking about their experience as Cluegiver they say:

> *"When it was my turn and when I was listening to my teammates' reasoning, I was back in and it was like a normal game of Codenames, but anything outside of that time it was me and the computer [...] yeah totally distracted."* (P5 G7)

Here this participant comments that the only time they felt connected to the game was when their Guessers were actively reasoning through a clue that had just been given, otherwise they felt distracted from the wider game due to the Agent. This was further evidenced when (P5), while typing into the AI, was completely distracted from a conversation going on between their two Guessers who were clarifying the word Bugle wasn't the word Beagle. After this (P5) gave a clue for the word Beagle because they hadn't paid attention to the discussion that was happening at the table; *"And actually to add about what (P5) said, that he thought it was Beagle. P4 and I spoke about that word 20,000 times. So I think it also relates back to how distracted he was with the AI"* (P6 G7).

5.4 The Agent reconfigured social dynamics

Participants frequently reported on the Agent's effect on how they performed as a team, and its role within the team. Connections within teams seemed to become weaker, whereas the cross-team connection between Cluegivers seemed to become stronger. The Agent was often thought of as an extra player, separate from the role of Guessers and Cluegivers. Finally, the Agent led to a reduction in body language and nonverbal communication within teams.

5.4.1 Re-formed relationships within and across teams. In almost all games, Guessers and Cluegivers felt less connected to each other because of the Agent; *"it seemed to be quite detrimental and quite negative affecting as a team"* (P1 G5). Even participants who often play together in the past felt the Agent actively worsened how they performed as a team:

> *"We've played this in the past and I've found me and (P2) are typically quite good with each other" (P1) , "This was not good" (P2), "We did not have any synergy and I could just tell she was using AI"* (P1 G1)

For strangers to perform well in a team, it was important for Cluegivers to learn how their Guessers made connections between words. In game 5, participant two compares their experience with the Agent to a warm-up game they played; *"I already knew how your mind worked making connections, where you mentioned something and the inverse of it...so yeah very quick to learn as a team without the AI."* (P2 G5). Some participants indicated that as Cluegivers, the Agent split their focus between thinking about their teammates and thinking about the Agent. This was mentioned in both strangers - *"We're split between trying to get to know each other and trying to figure out how the AI is kind of doing things."* (P5 G6) - and friends settings; *"I feel I usually do those (style of clues), but because I was focused on using the AI, I wasn't thinking about that side of the game as much."* (P3 G1). Notably, this disconnection appeared much more frequently in strangers than with friends, as friends already had a shared connection with each other. We observed Guessers in the friends settings chatting to other Guessers, on either team, which explains why friends might not have felt as disconnected when Cluegivers were typing.

Interestingly, while the Agent seemed to weaken the connection between Cluegiver and Guesser within a team, it strengthened (if not created) the relationship between Cluegivers on different teams. This was mainly due to the shared novel experience of using the Agent. Cluegivers could often be seen looking at each other's screens, and laughing at the suggestions presented by their AI's; *"I was tempted to show you (other Cluegiver) my screen like "hey look at this" "* (P2 G5). This connection was explicitly pointed out by one participant:

> *"So we were looking at each other during the game. Funnily enough as spy masters, we made connection there rather than with the team almost."* (P4 G5)
>
> *"Yeah it shifted from Guesser and Cluemaster to Cluemaster and Cluemaster"* (P2 G5)

Overall, the Agent seemed to shift the social dynamics of the game, strengthening the connection between Cluegivers, while weakening the connection between a Cluegiver and their Guessers.

5.4.2 AI as an additional player. Although it was introduced as an assistive tool for Cluegivers, many participants reported feeling as though the Agent became another player on their team.

After having used the Agent and then swapping back to the Guesser role, one participant noted how they tried to discern how much of an influence the Agent had on a clue given; *"I found myself almost guessing how much of that (clue) was my partner and how much of it was the AI... because having read the AI and trying to balance that, it almost felt like there was a third player there in a* *sense."* (P4 G5) They likened the experience to playing with two Cluegivers who collaborate on a clue. This was also supported by participants on the Cluegiver side. P2 in game three felt the agent served "as a third party" which can provide new information; *"If I feel like I've done enough thinking, I'll see what this thing thinks. It's like a third party that's doing some dumb thinking, but at the very least it's new information that I get to look at."* (P2 G3).

A few participants chose to name their agent; *"let's see what CC says" and then I will be like "oh that's total bullshit never mind. I won't ask CC again on this one".* (P2 G3) and *"So P2 and I are on a team and say BAZZ the robot is going to blurt out a clue [...] I say "BAZZ Carpet for 2" ."* (P1 G1).

Participants sometimes reported that as Cluegivers, they felt their role shift to an intermediary between the Agent and their Guessers:

> *"The resource makes you think more."* (P5 G4)
>
> *"Yeah! Because you're trying to know how your friends think as well and whether they'll get it."* (P2 G4)
>
> *"Whether your friends think like AI?"* (P4 G4)

Normally Cluegivers double check whether their Guessers will understand the reasoning behind their own clue. With the Agent present, Cluegivers now needed to consider whether Guessers will understand the reasoning of the Agent's clue. In this way Cluegivers acted as a mediator between what the Agent suggested and what they thought their teammates might understand.

5.4.3 Interrupting body language. The presence of the laptop and typing interfered with body language by acting as a physical barrier between Cluegiver and Guessers. While body language could be seen as part of the player experience, participants often spoke about it in the context of their team.

Participants spoke about how the Agent interfered with the communicative aspects of body language. P1 in game 5 found that they felt left out of the shared communal activity of the game because their Cluegiver's body language was being blocked:

> *"There are some aspects that are body language so having the AI here in the way of the table or what we're doing, puts that barrier between us. Here's an activity we're all taking part in and I put a small barrier in front of me for it [...] You kind of get that sense of here's what we are all doing, and if people are using a different device your body language is directed towards something else"* (P1 G5)

Participants in game 1 spoke about the importance of body language for gathering non-verbal information about clues; *"Even just where your eyes dart to. If someone is looking at a certain spot, you kind of notice that"* (P1 G1). Since Cluegivers were putting their words into the laptop they were no longer looking at the words on the table. Many players would consider reading facial expression and gaze tracking as cheating; *"Spy Masters should not say anything. I think there was some reactions every now and then. I think hiding your face almost (is good)."* (P1 G1). Because of this, one participant found that the Agent's interference with body language and facial cues was helpful in not giving away information:

> *"I always find it really hard to hide what I'm feeling. So it did kind of help. I was able to look at the computer instead of looking at the board and giving away what it was"* (P3 G4)

6 Discussion

Our findings showed that the effect of an assistive AI agent goes beyond simply assisting Cluegivers cognitively in playing a typical game of *Codenames* to influencing all players' reasoning, communication and social dynamics. This had wide reaching effects on player experience, fundamentally changing the nature of the game for both Cluegivers and Guessers.

6.1 AI as an equalizer

The AI acted as an equalizer between novice and more experienced players. It did this by making aspects of the game easier for players who were newer or struggling, and by making it more difficult for players looking to excel. Players' struggles were lifted through the enhancement of convergent and divergent thinking, allowing Cluegivers to give clues faster and with less mental blocks. The Agent also alleviated some of the major psychological struggles faced by players such as feelings of stress, anxiety and isolation. Yet, the Agent constrained players looking to excel. It diverted Cluegivers' attention away from giving tailored clues or learning about teammates, instead leading them to fixate on the Agent. The Agent also affected Guessers' ability to interpret Cluegivers' intentions, as many Guessers struggled to determine how much of a clue was influenced by the AI. In our findings around why people play, many participants specified these theory of mind related activities or "insights into personal psychology" were the main attractive feature of *Codenames*.

While the Agent shows the potential for making *Codenames* fairer by levelling the playing field between players of different experience levels, this does seem to come at the cost of reducing some aspects of what people find most enjoyable about the game. Yet, this maintenance of fair playing fields has been shown to be of high importance to many boardgame players [47]. By equalizing the competition between experienced and inexperienced players, AI shows potential for promoting accessibility in hybrid boardgames, a challenge documented by the literature [49]. To further characterise the role played by the AI in the context of hybrid play, we draw on the Hybrid Digital Boardgame Model [49]. From this perspective, the Agent can be seen as taking on a housekeeping and teaching role through its ability to hold Cluegivers' target words and suggest ways of thinking about clues that help inexperienced players formulate their own clues.

Xu et al. warn against automating housekeeping or chore-like activities as these aid in social interaction, reflection on gameplay, and strategy development [76]. Yet interestingly, we found that through this housekeeping role, the agent often aided participants in their creative thinking, led to discussions between participants as to the nature of good clues, and brought Cluegivers closer together. This highlights the potential for AI to enhance the hybrid gaming experience and also the need for future work to interrogate the unique differences AI brings to the hybrid play experience compared to other digital technologies.

6.2 AI changes the distribution of cognitive labour

Our design and intention for the Agent was to assist Cluegivers in generating clues, which we expected to make the game easier overall. While many Cluegivers reported the Agent made some aspects of the game easier, including the creative thinking associated with clue generation, we were surprised to see the Agent increasing the complexity and cognitive labour in other parts of the game for both Cluegivers and Guessers.

For Cluegivers this came about through the Agent shifting most of the cognitive labour to the end of the game, as opposed to how it typically distributed throughout. In an ordinary game of *Codenames* Cluegivers are typically running through possible clues for all their words, not just the words they give their first few clues for. In doing this, they memorise and familiarise themselves with their target words and think of potential ideas for future clues. This can be seen as "precomputation" where the cognitive labour of cluing the final words, which is typically harder than initial words, is distributed throughout the course of the game [43, 52]. In our setup, the Agent allowed players to offload this cognitive work by simply typing their words onto the laptop. Cluegivers would quickly input different combinations of words until the Agent suggested something the players were happy with or could alter easily. In doing this Cluegivers were forgoing the process of internalizing their words or possible clue ideas. While Cluegivers reported this made the start of the game easier, they felt it made the end of the game much more difficult. Typically as the game progresses, finding associations between words becomes more difficult, requiring nuanced or creative clues, something the Agent struggled with. This meant that players could no longer rely on the Agent as they did in the beginning, and had to switch back to coming up with their clues alone, something that was frustrating to most Cluegivers. Ultimately Cluegiver's over-reliance on the Agent at the beginning of the game led to a frustrating increase in cognitive labour at the end.

For Guessers, increased cognitive labour came in the form of having to develop more complex mental models of their Cluegivers, which in turn increased the difficulty of interpreting clues. In a typical game of *Codenames*, Guessers might use a mental model of their Cluegivers to perform theory of mind reasoning to decipher the clue given. This explains why players who know each other well typically find it easier to guess clues compared to strangers, as they have existing mental models. Yet, in our version of *Codenames*, Guessers were not aware of whether a clue was AI-generated, human-generated or a mix of both. Therefore, they could not simply apply their mental model of the Cluegiver to decipher the clue. To perform any theory of mind reasoning, Guessers had to form mental models of the Agent, which we saw evidence of when they referred to the Agent as "another player" in the game and when they compared how Cluegivers think to how the Agent thinks. On top of this, Guessers also needed to know what their Cluegiver thought of the Agent as this helped them determine how much of the clue was influenced by the Agent. This amounted to Guessers having to do extra cognitive work to decipher a clue, which some participants referred to as applying an "AI filter".

What we see here, therefore is a critical tension between the AI's role as a collaborator and its role as a teammate. For collaboration, an unpredictable agent results in enhancing player creativity in clue generation, which is in line with the current literature [24, 32, 67]. Yet, as previously stated, this unpredictability led to difficulty in team mental model formation and theory of mind reasoning for teaming, again in line with the teaming literature [7, 57]. This highlights cooperative creative domains as a unique and challenging settings for investigating human-AI cooperation, where tension arises from the needs to balance teaming and creativity. Further research should investigate how other aspects of human-AI teaming and collaboration, such as trust and perceived identity, apply to cooperative creative domains. In the context of hybrid game design, having players perform theory of mind reasoning over a hybrid human-AI mind presents a challenging yet intriguing design space which may provide a unique experience for players.

6.3 AI as clown and social disruptor

Maintaining sociality is one of the top priorities for designers of hybrid games, with designers showing concern that digital elements risk detracting from the sociality of play [48]. Therefore, we were interested to see how strongly the Agent contributed to and disrupted the social dynamics within the game.

The Agent typically was given a social role akin to that of a clown or jester, where its suggestions added an extra layer of hilarity to the game. These suggestions invited ribbing from participants, a sort of friendly banter, directed towards the Agent. This was observed in the groups of strangers, who might otherwise be more reserved and polite. This is inline with Suh et al's concept of "AI as social glue" where the AI acts to mediate social dynamics, especially in settings with implicit tensions such as strangers working together on a creative project [62]. Our Agent thereby enhanced the social experience of players by acting as a psychological safety net. This builds upon results from Hwang and Wong, who find that AI can improve the creative experience for participants with high anxiety, extending this finding to the setting of games [32]. This psychological safety net supports the notion that technology can support accessibility in hybrid games [49]. The social presence goes beyond simply that of a tool, as shown when participants referred to the Agent as a teammate or give it a name.

While the Agent took on a strong social presence within the game, it also disrupted the typical social dynamics between players in *Codenames*, weakening social ties between Cluegiver and Guesser while strengthening those between Cluegiver and Cluegiver. This extends findings presented by Suh et al. who show that AI assistants "seed common ground between human collaborators" [62]. Our *Codenames* set up highlights the importance of who has access to the assistive technology. Since Cluegivers but not Guessers had access, this inequality led to a divide between teammates. In contrast, for Cluegivers on different teams who have no incentive to collaborate on clues, the assistive agent actually led to new collaborative interactions. This shows that AI assistants can even seed common ground and enhance collaboration between participants who have no incentive to collaborate in the first place.

Given the concerns raised by designers regarding the impacts of technology on social dynamics in hybrid games, it is important to consider that AI can have unexpected yet positive effects on the social dynamics within a social game. In particular, careful consideration should be placed on which players have access to the technology, as equal access to assistive AI technology can lead to increased collaborative interaction.

6.4 AI reduces externalised cognition and adds hidden information

In boardgames, player cognition is typically externalised through the pieces and the board [47, 74]. Research has shown how this externalisation of cognition is critical to how effectively humans cooperate, especially in PIRCC games, where it acts as a side channel of communication, aiding players in performing theory of mind reasoning [54]. In a typical game of *Codenames* players are sharing a space over the the game pieces and focusing their attention over this space. This aspect of the game was something that players felt connected them as a team and has also been shown to improve non-verbal communication [4]. In our set up of *Codenames*, with the AI located on a laptop, this connection was disrupted in two ways. Firstly, the shared space was disrupted due to the laptop acting as a physical barrier between Cluegiver and Guesser. Indications of the Cluegivers attention, such as eye gaze, are now more strongly focused on a screen and away from a shared space, which is important for establishing a mutual understanding of the situation through 'common ground' [27]. In this way, the Agent violated guiding principles on the use of digital technology in hybrid game design by drawing players' attention away from the board [48]. Secondly, the suggestions given by the Agent added another layer of hidden information for Cluegivers. This made it harder for Guessers to decipher clues as there were more unknown factors that might influence the Cluegiver's clue. While the first issue of drawing attention away from the board could potentially be avoided through having the AI deployed in something like a head mounted display or augmented reality, the second issue of added hidden information will be present no matter how the technology is designed and deployed.

This highlights an important consideration when including digital technology in physical games. Designers should consider not only the effects that the physical technology will have on the game, but also what effect it will have on players' cognition. Technology (including AI) may alter the cognitive experience of the game by adding hidden information, providing psychological support, or adding an additional mind with which players must reason over. In our study these have all been shown to alter the player experience in drastic ways. We therefore suggest that designers looking to include AI agents in their game carefully consider how this might alter the experience of the game.

6.5 Limitations and Future Work

In this study, we chose to focus on a single PIRCC game, namely *Codenames*. This may limit the generalisability of our findings to other PIRCC games, which warrant their own investigation. We believe that many of our insights could apply to other PIRCC games and non-gaming PIRCC environments as many of our findings relate to how AI affects general team communication and reasoning in these hidden information environments. We see great potential in future work exploring AI's effect on players reasoning and communication in other PIRCC games. We also acknowledge that while our study design splits participants into friends and strangers groups, our results tended not to draw direct comparisons between the two groups. While this could have been achieved by constructing a more controlled experimental setup, we decided to prioritise maintaining a natural environment for players, which we believe led to richer results that better capture how players might behave in a non-lab setting. A promising avenue for future work would be the effects of AI as an assistant to Guessers rather than, or as well as, Cluegivers. We also see potential in exploring different implementations of an assistive AI agent for Cluegivers, either through a shared device for both Cluegivers or exploring the effect of other devices such as mobile or AR systems.

7 Conclusion

This paper investigates how an assistive AI affects the cognition, social dynamics, and player experience of teams of *Codenames* players. Our results show that our AI agent enhances convergent and divergent thinking, challenges players' notion of "the spirit of the game", alters typical team social connection, and adds to the play experience by introducing hilarity and psychological safety. We address what ways AI can improve or detract from the play experience in hybrid PIRCC games. AI improves accessibility of hybrid games by acting as equalising tool between novice and experienced players. We identify how AI alters the distribution of cognitive labour throughout the game, concentrating it around the end of the game. We also highlight that AI lessens elements of externalised cognition and adds additional hidden information to the game. Finally, our work identifies an inherent tension between the recommendations of predictability from human-AI teaming, and randomness from human-AI collaboration. This work informs the design and study of human-AI cooperation in PIRCC settings as well as the trade offs that are introduced when designing hybrid games with AI.

Acknowledgments

Matthew Sidji's work is supported by the Australian Government Research Training Program Scholarship, by the University of Melbourne Faculty of Engineering and Information Technology Ingenium Scholarship Award, and by the LMA STELaRLab top up Scholarship. We would also like to acknowledge the work done by Samangi Wadinambiarachchi in illustrating some of the figures in this paper.

References

[1] Safinah Ali, Hae Won Park, and Cynthia Breazeal. 2020. Can Children Emulate a Robotic Non-Player Character's Figural Creativity?. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '20)*. Association for Computing Machinery, New York, NY, USA, 499–509. https://doi.org/10.1145/3410404.3414251

[2] Andrew Anderson, Jonathan Dodge, Amrita Sadarangani, Zoe Juozapaitis, Evan Newman, Jed Irvine, Souti Chattopadhyay, Matthew Olson, Alan Fern, and Margaret Burnett. 2020. Mental Models of Mere Mortals with Explanations of Reinforcement Learning. *ACM Transactions on Interactive Intelligent Systems* 10 (5 2020), 1–37. https://doi.org/10.1145/3366485

[3] Roberts Angel. 1981. *Pictionary*. Angel Games.

[4] Michael Argyle and Jean Ann Graham. 1976. The Central Europe Experiment: Looking at Persons and Looking at Objects. *Environmental psychology and nonverbal behavior* 1, 1 (Sept. 1976), 6–16. https://doi.org/10.1007/BF01115461
[5] Zahra Ashktorab, Q. Vera Liao, Casey Dugan, James Johnson, Qian Pan, Wei Zhang, Sadhana Kumaravel, and Murray Campbell. 2020. Human-AI Collaboration in a Cooperative Game Setting: Measuring Social Perception and Outcomes. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2 (2020), Article 96. https://doi.org/10.1145/3415167
[6] Eric Azagury and Bruno Poullet. 2018. *Just One.* Repos Production.
[7] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7, 1 (Oct. 2019), 2–11.
[8] Nolan Bard, Jakob N. Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H. Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, Iain Dunning, Shibl Mourad, Hugo Larochelle, Marc G. Bellemare, and Michael Bowling. 2020. The Hanabi challenge: A new frontier for AI research. *Artificial Intelligence* 280 (1 3 2020), 103216. https://doi.org/10.1016/j.artint.2019.103216
[9] Antoine Bauza. 2010. *Hanabi.* R&R Games.
[10] Rhyse Bendell, Jessica Williams, Stephen M. Fiore, and Florian Jentsch. 2021. Towards Artificial Social Intelligence: Inherent Features, Individual Differences, Mental Models, and Theory of Mind, Hasan Ayaz, Umer Asgher, and Lucas Paletta (Eds.). *Advances in Neuroergonomics and Cognitive Engineering*, 20–28. https://doi.org/10.1007/978-3-030-80285-1_3
[11] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, and Chris Hesse. 2019. Dota 2 with Large Scale Deep Reinforcement Learning. *arXiv preprint arXiv:1912.06680* (2019). arXiv:1912.06680
[12] Noam Brown and Tuomas Sandholm. 2019. Superhuman AI for Multiplayer Poker. *Science* 365, 6456 (2019), 885–890. https://doi.org/10.1126/science.aay2400
[13] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askell. 2020. Language Models Are Few-Shot Learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
[14] Murray Campbell, A. Joseph Hoane, and Feng-hsiung Hsu. 2002. Deep Blue. *Artificial Intelligence* 134, 1 (2002), 57–83. https://doi.org/10.1016/S0004-3702(01)00129-1
[15] Rodrigo Canaan, Christoph Salge, Julian Togelius, and Andy Nealen. 2019. Leveling the playing field: fairness in AI versus human game benchmarks. *Proceedings of the 14th international conference on the foundations of digital games.* https://doi.org/10.1145/3337722.3337750 Citation Key: 10.1145/3337722.3337750 number-of-pages: 8 publisher-place: San Luis Obispo, California, USA tex.articleno: 37.
[16] Michal Čertický and David Churchill. 2017. The Current State of StarCraft AI Competitions and Bots. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 13. 2–7.
[17] Prithvijit Chattopadhyay, Deshraj Yadav, Viraj Prabhu, Arjun Chandrasekaran, Abhishek Das, Stefan Lee, Dhruv Batra, and Devi Parikh. 2017. Evaluating Visual Conversational Agents via Cooperative Human-AI Games. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 5 (Sept. 2017), 2–10. https://doi.org/10.1609/hcomp.v5i1.13312
[18] Vlaada Chvátil. 2015. *Codenames.* Czech Games Edition.
[19] Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R McKee, Joel Z Leibo, Kate Larson, and Thore Graepel. 2020. Open problems in cooperative AI. *arXiv preprint arXiv:2012.08630* (2020).
[20] Thomas Dagenais-Lesperance. 2018. *Decrypto.* Scorpion Masque.
[21] Emile de Maat. 2024. Communication Limits | Board Game Mechanic | BoardGameGeek. https://boardgamegeek.com/boardgamemechanic/2893/communication-limits/linkeditems/boardgamemechanic?pageid=1&mechanicfilter=2023 Accessed : 2024-05-28.
[22] Attila Egri-Nagy and Antti Törmänen. 2020. The Game Is Not over Yet—Go in the Post-AlphaGo Era. *Philosophies* 5, 4 (2020), 37–0. https://doi.org/10.3390/philosophies5040037
[23] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R. Millen, Murray Campbell, Sadhana Kumaravel, and Wei Zhang. 2020. Mental Models of AI Agents in a Cooperative Game Setting. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* Association for Computing Machinery, 1–12. https://doi.org/10.1145/3313831.3376316 DOI: 10.1145/3313831.3376316.
[24] Katy Ilonka Gero and Lydia B. Chilton. 2019. Metaphoria: An algorithmic companion for metaphor creation. *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–12. https://doi.org/10.1145/3290605.3300526 Citation Key: 10.1145/3290605.3300526 number-of-pages: 12 publisher-place: Glasgow, Scotland Uk.
[25] Christian Guckelsberger, Christoph Salge, Jeremy Gow, and Paul Cairns. 2017. Predicting player experience without the player.: An exploratory study. *Proceedings of the annual symposium on computer-human interaction in play*, 305–315. https://doi.org/10.1145/3116595.3116631 Citation Key: 10.1145/3116595.3116631 number-of-pages: 11 publisher-place:

Amsterdam, The Netherlands.

[26] Rotem D Guttman, Jessica Hammer, Erik Harpstead, and Carol J Smith. 2021. Play for Real (ism)-Using Games to Predict Human-AI interactions in the Real World. *Proceedings of the ACM on Human-Computer Interaction* 5, CHI PLAY (2021), 1–17. publisher: ACM New York, NY, USA.

[27] Carl Gutwin and Saul Greenberg. 2002. A Descriptive Framework of Workspace Awareness for Real-Time Groupware. *Computer Supported Cooperative Work (CSCW)* 11, 3 (Sept. 2002), 411–446. https://doi.org/10.1023/A:1021271517844

[28] Matthew Guzdial, Nicholas Liao, Jonathan Chen, Shao-Yu Chen, Shukan Shah, Vishwa Shah, Joshua Reno, Gillian Smith, and Mark O. Riedl. 2019. Friend, Collaborator, Student, Manager: How Design of an AI-Driven Game Level Editor Affects Creators. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300854

[29] Brian Hersch. 1989. *Taboo*. Editions ATLAS.

[30] Keke Hou, Tingting Hou, and Lili Cai. 2023. Exploring Trust in Human–AI Collaboration in the Context of Multiplayer Online Games. *Systems* 11, 5 (May 2023), 217. https://doi.org/10.3390/systems11050217

[31] Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. 2020. "Other-Play" for Zero-Shot Coordination. In *Proceedings of the 37th International Conference on Machine Learning*, Daumé Hal, III and Singh Aarti (Eds.), Vol. 119. PMLR, Proceedings of Machine Learning Research, 4399–4410.

[32] Angel Hsing-Chi Hwang and Andrea Stevenson Won. 2021. IdeaBot: Investigating Social Facilitation in Human-Machine Team Creativity. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, Yokohama, Japan, Article 712. https://doi.org/10.1145/3411764.3445270

[33] Catalina Jaramillo, Megan Charity, Rodrigo Canaan, and Julian Togelius. 2020. Word Autobots: Using Transformers for Word Association in the Game Codenames. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 16. 231–237.

[34] Ville Kankainen and Janne Paavilainen. 2019. Hybrid Board Game Design Guidelines.

[35] Markelle Kelly, Aakriti Kumar, Padhraic Smyth, and Mark Steyvers. 2023. Capturing Humans' Mental Models of AI: An Item Response Theory Approach. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1723–1734. https://doi.org/10.1145/3593013.3594111 event-place: Chicago, IL, USA.

[36] Andrew Kim, Maxim Ruzmaykin, Aaron Truong, and Adam Summerville. 2019. Cooperation and Codenames: Understanding Natural Language Processing via Codenames. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* 15, 1 (8 10 2019), 160–166. section: Poster Papers.

[37] Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083* (2023).

[38] Adam Lerer, Hengyuan Hu, Jakob Foerster, and Noam Brown. 2020. Search in Cooperative Partially-Observable Games. In *AAAI Conference on Artificial Intelligence*.

[39] Bernhard Maurer and Verena Fuchsberger. 2019. Dislocated boardgames: Design potentials for remote tangible play. *Multimodal technologies and interaction* 3, 4 (2019), 72. publisher: MDPI.

[40] Chelsea Mills, Denise Y. Geiskkovitch, Carman Neustaedter, William Odom, and Benett Axtell. 2023. Remote wavelength: Design and evaluation of a system for social connectedness through distributed tabletop gameplay. *Proceedings of the 2023 CHI conference on human factors in computing systems*. https://doi.org/10.1145/3544548.3581142 Citation Key: 10.1145/3544548.3581142 number-of-pages: 19 publisher-place:Hamburg, Germany tex.articleno: 680.

[41] Chelsea Mills, Carman Neustaedter, and William Odom. 2021. Distributed letter jam: Designing distributed board game play for social connection. *Companion publication of the 2021 conference on computer supported cooperative work and social computing*, 135–139. https://doi.org/10.1145/3462204.3481753 Citation Key: 10.1145/3462204.3481753 number-of-pages: 5 publisher-place: Virtual Event, USA.

[42] Edi Mukardin, Martin Tappler, Bernhard K Aichernig, and Ingo Pill. 2023. Reinforcement Learning under Partial Observability Guided by Learned Environment Models. In *International Conference on Integrated Formal Methods*. Springer, 257–276.

[43] Donald A Norman. 1992. Design Principles for Cognitive Artifacts. *Research in Engineering Design* 4, 1 (1992), 43–50.

[44] OpenAI, :, Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemys law Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, Rafal Józefowicz, Scott Gray, Catherine Olsson, Jakub Pachocki, Michael Petrov, Henrique P. d. O. Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, Jie Tang, Filip Wolski, and Susan Zhang. 2019. Dota 2 with Large Scale Deep Reinforcement Learning. *arXiv e-prints* (12 2019), arXiv:1912.06680. eprint: 1912.06680.

[45] OpenAI. 2022. ChatGPT: Optimizing Language Models for Dialogue.

[46] OpenAI. 2023. OpenAI API.

[47] Melissa J Rogerson, Martin R Gibbs, and Wally Smith. 2018. Cooperating to Compete: The Mutuality of Cooperation and Competition in Boardgame Play. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[48] Melissa J. Rogerson, Lucy A. Sparrow, and Martin R. Gibbs. 2021. More than a gimmick - digital tools for boardgame play. *Proc. ACM Hum.-Comput. Interact.* 5, CHI PLAY (10 2021). https://doi.org/10.1145/3474688 Citation Key: 10.1145/3474688 number-of-pages: 23 publisher-place: New York, NY, USA publisher: Association for Computing Machinery.

[49] Melissa J. Rogerson, Lucy A. Sparrow, and Martin R. Gibbs. 2021. Unpacking "Boardgames with apps": The hybrid digital boardgame model. *Proceedings of the 2021 CHI conference on human factors in computing systems.* https://doi.org/10.1145/3411764.3445077 Citation Key: 10.1145/3411764.3445077 number-of-pages: 17 publisher-place: Yokohama, Japan tex.articleno: 111.

[50] Shaghayegh Roohi, Asko Relas, Jari Takatalo, Henri Heiskanen, and Perttu Hämäläinen. 2020. Predicting game difficulty and churn without players. *Proceedings of the annual symposium on computer-human interaction in play*, 585–593. https://doi.org/10.1145/3410404.3414235 Citation Key: 10.1145/3410404.3414235 number-of-pages: 9 publisher-place: Virtual Event, Canada.

[51] William B Rouse, Janis A Cannon-Bowers, and Eduardo Salas. 1992. The role of mental models in team performance in complex systems. *IEEE transactions on systems, man, and cybernetics* 22, 6 (1992), 1296–1308. publisher: IEEE.

[52] Kjeld Schmidt and Carla Simonee. 1996. Coordination Mechanisms: Towards a Conceptual Foundation of CSCW Systems Design. *Computer Supported Cooperative Work (CSCW)* 5, 2 (1996), 155–200.

[53] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. 2020. Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model. *Nature* 588, 7839 (Dec. 2020), 604–609. https://doi.org/10.1038/s41586-020-03051-4

[54] Matthew Sidji, Wally Smith, and Melissa J. Rogerson. 2023. The Hidden Rules of Hanabi: How Humans Outperform AI Agents. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3544548.3581550

[55] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, and Adrian Bolton. 2017. Mastering the game of go without human knowledge. *nature* 550, 7676 (2017), 354–359.

[56] Thomas Sing. 2019. *The Crew: The Quest for Planet Nine.* Kosmos.

[57] Ho Chit Siu, Jaime Peña, Edenna Chen, Yutai Zhou, Victor Lopez, Kyle Palko, Kimberlee Chang, and Ross Allen. 2021. Evaluation of Human-AI Teams for Learned and Rule-Based Agents in Hanabi. *Advances in Neural Information Processing Systems* 34 (2021).

[58] Sherman Siu. 2022. Towards automating Codenames spymasters with deep reinforcement learning. *arXiv preprint arXiv:2212.14104* (2022).

[59] Ondra Skoupy. 2019. *Letter Jam.* Czech Games Edition.

[60] Brad Spendlove and Dan Ventura. 2022. Competitive Language Games as Creative Tasks with Well-Defined Goals.. In *ICCC*. 291–299.

[61] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, and Adrià Garriga-Alonso. 2022. Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models. *arXiv preprint arXiv:2206.04615* (2022). arXiv:2206.04615

[62] Minhyang (Mia) Suh, Emily Youngblom, Michael Terry, and Carrie J Cai. 2021. AI as Social Glue: Uncovering the Roles of Deep Generative AI during Social Music Composition. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* https://doi.org/10.1145/3411764.3445219 event-place: Yokohama, Japan.

[63] Gerald Tesauro. 1995. Temporal difference learning and TD-Gammon. *Commun. ACM* 38, 3 (1995), 58–68.

[64] Mike Treanor, Alexander Zook, Mirjam P Eladhari, Julian Togelius, Gillian Smith, Michael Cook, Tommy Thompson, Brian Magerko, John Levine, and Adam Smith. 2015. AI-based Game Design Patterns. In *Proceedings of the 10th International Conference on the Foundations of Digital Games 2015 (FDG 2015).*

[65] Jennifer Villareale, Casper Harteveld, and Jichen Zhu. 2022. "I Want To See How Smart This AI Really Is": Player mental model development of an adversarial AI player. *Proc. ACM Hum. Comput. Interact.* 6, CHI PLAY (10 2022). https://doi.org/10.1145/3549482 Citation Key: 10.1145/3549482 number-of-pages: 26 publisher-place: New York, NY, USA publisher: Association for Computing Machinery tex.articleno: 219 tex.issue date: October 2022.

[66] Jennifer Villareale and Jichen Zhu. 2021. Understanding Mental Models of AI through Player-AI Interaction. *arXiv preprint arXiv:2103.16168* (2021).

[67] Samangi Wadinambiarachchi, Ryan M. Kelly, Saumya Pareek, Qiushi Zhou, and Eduardo Velloso. 2024. The Effects of Generative AI on Design Fixation and Divergent Thinking. In *Proceedings of the CHI Conference on Human Factors in Computing Systems.* ACM, Honolulu HI USA, 1–18. https://doi.org/10.1145/3613904.3642919

[68] James R. Wallace, Joseph Pape, Yu-Ling Betty Chang, Phillip J. McClelland, T.C. Nicholas Graham, Stacey D. Scott, and Mark Hancock. 2012. Exploring automation in digital tabletop board game. *Proceedings of the ACM 2012 conference on computer supported cooperative work companion*, 231–234. https://doi.org/10.1145/2141512.2141585 Citation Key:

10.1145/2141512.2141585 number-of-pages: 4 publisher-place: Seattle, Washington, USA.
[69] Qian Wan, Siying Hu, Yu Zhang, Piaohong Wang, Bo Wen, and Zhicong Lu. 2023. " It Felt Like Having a Second Mind": Investigating Human-AI Co-creativity in Prewriting with Large Language Models. *arXiv preprint arXiv:2307.10811* (2023).
[70] Dakuo Wang, Elizabeth Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. 2020. From human-human collaboration to Human-AI collaboration: Designing AI systems that can work together with people. 1–6.
[71] Wolfgang Warsch. 2018. *The Mind.* Nürnberger-Spielkarten-Verlag.
[72] Rina R. Wehbe, Edward Lank, and Lennart E. Nacke. 2017. Left Them 4 Dead: Perception of Humans versus Non-Player Character Teammates in Cooperative Gameplay. In *Proceedings of the 2017 Conference on Designing Interactive Systems (DIS '17)*. Association for Computing Machinery, New York, NY, USA, 403–415. https://doi.org/10.1145/3064663.3064712
[73] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. *arXiv preprint arXiv:2201.11903* (2022). arXiv:2201.11903
[74] Margaret Wilson. 2002. Six Views of Embodied Cognition. *Psychonomic bulletin & review* 9 (2002), 625–636.
[75] Maximilian Wittmann and Benedikt Morschheuser. 2022. What do games teach us about designing effective human-AI cooperation?-A systematic literature review and thematic synthesis on design patterns of non-player characters. *Player Characters* (2022), 95–104.
[76] Yan Xu, Evan Barba, Iulian Radu, Maribeth Gandy, and Blair MacIntyre. 2011. Chores Are Fun: Understanding Social Play in Board Games for Digital Tabletop Game Design.. In *Digra Conference*.
[77] Qiao Zhang. 2023. Understanding human-AI teaming dynamics through gaming environments, Vol. 19. 440–443. issue: 1.
[78] Rui Zhang, Nathan J McNeese, Guo Freeman, and Geoff Musick. 2021. " An Ideal Human" Expectations of AI Teammates in Human-AI Teaming. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–25. publisher: ACM New York, NY, USA.
[79] Jichen Zhu, Jennifer Villareale, Nithesh Javvaji, Sebastian Risi, Mathias Löwe, Rush Weigelt, and Casper Harteveld. 2021. Player-AI Interaction: What Neural Network Games Reveal About AI as Play. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, Yokohama, Japan, Article 77. https://doi.org/10.1145/3411764.3445307
[80] Faheem Hassan Zunjani and Ana-Maria Olteteanu. 2019. Towards Reframing Codenames for Computational Modelling and Creativity Support Using Associative Creativity Principles. *Proceedings of the 2019 on Creativity and Cognition*, 407–413. https://doi.org/10.1145/3325480.3325510 event-place: San Diego, CA, USA.

Received February 2024; revised June 2024; accepted July 2024