"I Want To See How Smart This AI Really Is": Player Mental Model Development of an Adversarial AI Player

JENNIFER VILLAREALE, Drexel University, USA
CASPER HARTEVELD, Northeastern University, USA
JICHEN ZHU, IT University of Copenhagen, Denmark

Understanding players' mental models are crucial for game designers who wish to successfully integrate player-AI interactions into their game. However, game designers face the difficult challenge of anticipating how players model these AI agents during gameplay and how they may change their mental models with experience. In this work, we conduct a qualitative study to examine how a pair of players develop mental models of an adversarial AI player during gameplay in the multiplayer drawing game *iNNk*. We conducted ten gameplay sessions in which two players (n = 20, 10 pairs) worked together to defeat an AI player. As a result of our analysis, we uncovered two dominant dimensions that describe players' mental model development (i.e., focus and style). The first dimension describes the focus of development which refers to what players pay attention to for the development of their mental model (i.e., top-down vs. bottom-up focus). The second dimension describes the differences in the style of development, which refers to how players integrate new information into their mental model (i.e., systematic vs. reactive style). In our preliminary framework, we further note how players process a change when a discrepancy occurs, which we observed occur through comparisons (i.e., compare to other systems, compare to gameplay, compare to self). We offer these results as a preliminary framework for player mental model development to help game designers anticipate how different players may model adversarial AI players during gameplay.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; **Human computer interaction (HCI)**.

Additional Key Words and Phrases: Human-AI Interaction; Mental Models; Game Design; User Experience

ACM Reference Format:
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, Article 219 (October 2022), 25 pages. https://doi.org/10.1145/3549482

1 INTRODUCTION

With the recent boom in Artificial Intelligence (AI) applications, game designers have been increasingly exploring a variety of AI approaches in computer games [72, 79], from procedural content generation (PCG) [10, 56, 63] to intelligent non-player characters [24, 72]. As a result, designers are producing novel gameplay experiences by foregrounding these complex systems in the user interface (UI) and enabling players to directly interact with the AI as part of the core gameplay experience [8, 21, 29, 54, 64, 79]. With this development, players are now becoming aware of and playfully interacting with a growing number of these complex systems, and in turn, constructing mental models that may be

Authors' addresses: Jennifer Villareale, jmv85@drexel.edu, Drexel University, USA; Casper Harteveld, c.harteveld@northeastern.edu, Northeastern University, USA; Jichen Zhu, jichen.zhu@gmail.com, IT University of Copenhagen, Denmark.

more or less complete and accurate. Understanding players' mental models and designing the gameplay respectively can be instrumental to a positive player experience (PX) [17, 21, 23]. However, game designers face the difficult challenge of anticipating how players model these AI agents during gameplay and how they may change their mental models with experience.

HCI researchers have long used the mental model construct to understand how a user thinks a system works or behaves [35, 49]. Recently, a growing body of work in HCI has used this construct to study how people model AI systems [4, 21, 37]. However, current research on mental models of AI is relatively limited. Most existing work studies users' mental models *after* interacting with an AI [4, 37] and primarily focuses on what factors increase the accuracy of people's mental models. For instance, previous work has studied the effect of tutorials [37], explanations [7], and different kinds of AI errors [4]. Yet, an essential part of understanding mental models is to examine how users develop these models based on the response of the system [36, 68, 76]. Unfortunately, we have little knowledge of how people develop their mental model of AI as they interact with it.

Recently, a small group of work has studied how mental models of AI develop in games, primarily exploring how gameplay and surface features of the AI system impact mental model accuracy [21, 23]. For instance, Gero et al. [21] found that playing more games did not increase the accuracy of players' mental models but that players who won more often had more accurate mental models. However, how and why people form different mental models of AI systems and why some models shift with experience while others stay stable remains unclear. Insights into how players develop their mental models are crucial for game designers who wish to successfully integrate player-AI interactions into their games. However, we have little understanding of how this happens, especially over time.

To increase our understanding of player mental model development, we conducted a qualitative study to examine how players (n = 20, 10 pairs) develop mental models of AI during gameplay. Specifically, we examine how a *pair of players* makes sense of an *adversarial* AI player. Here, we examine the development process as a whole by utilizing the player-AI interaction [79] component and think-aloud data. We define player-AI interaction as the cycle of 1) player input (i.e., what the player does with the AI player) and the AI output (i.e., the feedback the player receives). We leverage these components to observe the differences in how players develop their mental models. Toward this end, we focus on novices of AI systems because they will likely engage in more mental model development [68] as opposed to experts who already have an established mental model to draw on. We also focus on adversarial AI games because such games require recognizing the AI's limitations as a critical component of gameplay. For example, an experienced *StarCraft* player that competes with an AI opponent uses their understanding of how the AI works and what its limitations are to win the match. Studying how players develop their mental model of adversarial AI is relevant for the game design community, as AI-based player experience has become prevalent in computer games [72, 79] and recently has been used as a productive domain of research for mental models [17, 23]. Furthermore, conducting a study with a pair of players leverages the social convention of discussing a shared task, a strategy utilized in similar work [53]. As such, we pose the following research question: How does a pair of players develop mental models of an adversarial AI during gameplay?

For our study, we utilize the AI-based game *iNNk* [64], a web-based multiplayer drawing game inspired by the well-known *Pictionary* game. With *iNNk*, however, two people play together against a Neural Network (NN). To win the game, the Guesser player has to correctly type the secret codeword based on the drawing provided by the Sketcher player before the NN guesses it. This game is well suited for our study because the success of the human team hinges on players' recognizing the AI's limitations, making the AI's error boundaries a critical component of gameplay.

Based on our study, we uncovered two dominant dimensions that describe how players' are developing their mental model (i.e., focus and style). The first dimension describes the focus of development which refers to what players pay

attention to for the development of their mental model, either attending to gameplay observations (i.e., bottom-up focus) or utilizing prior knowledge or experiences (i.e., top-down focus) to develop their mental model. The second dimension describes the differences in the style of development, which refers to how players integrate new information into their mental model, either by building on previous observations that were integrated into their model (i.e., systematic style) or by attending to the moment and integrating new information that is independent of what was previously observed (i.e., reactive style). In our preliminary framework, we further note how players process a change when a discrepancy occurs in their mental model, which we observed occur through comparisons (i.e., compare to other systems, compare to gameplay, compare to self).

We offer these results as a preliminary framework of player mental model development to help game designers anticipate how different players may model adversarial AI players during gameplay. This paper makes the following contributions:

- A qualitative analysis of ten gameplay sessions (n = 20, 10 pairs) in which participants played *iNNk* against an adversarial AI player, illustrating how players developed their mental model of the AI player during gameplay.
- We present a preliminary framework for describing how players' mental models may develop during gameplay. We discuss how this framework can be further developed and can be generalized to evaluate or design other adversarial AI-based games using *iNNk* as an example.
- We provide several design implications from our study for game designers and HCI researchers to improve current challenges in facilitating mental models of AI systems.

2 RELATED WORK

This section summarizes mental model theory from cognitive psychology and HCI literature. Then, we situate our work in the existing literature on the use of the construct of mental models to study AI systems in general and in the context of games specifically.

2.1 Mental Model Theory

A variety of constructs were proposed in cognitive psychology to explain knowledge representation and information processing [47, 58, 61, 70]. Among them, the most frequently used are mental models and schemata. While no clear lines have been drawn between these concepts [31, 58], it is widely acknowledged that mental models are their own concept and are not redundant with similar constructs like schemata [31, 50, 70, 76]. Schemata are defined as building blocks of cognition and are regarded as the unit by which people process information and make sense of new situations [58]. Theorists have suggested that it is helpful to think of a schema as a kind of interpretation of an event, object, or situation that is an informal, private, unarticulated theory about reality [51].

Mental models are closely related to schemata, and most researchers believe that mental models arise from and are the *running* mode of schemata [59]. As such, mental models are defined as internal representations of external reality that people use to interact with the world around them [11, 31, 49]. For example, when a person turns on their computer or interacts with a new device, people use mental models to instantiate a schema and simulate what actions to perform to complete the task at hand. Researchers have suggested that mental models evoke mental simulations [16, 36, 40] and are spatially arrayed corresponding to their real-life counterparts [13]. Simulating a model allows people to envision and predict future states of an environment or object, which involves completing a series of actions internally in a visual

format [40]. These models have been described as being more functional than accurate [49], allowing people to quickly respond to a changing environment [14, 19, 36] or produce explanations for events that have occurred previously [69].

Mental model studies have led cognitive theorists to assert that these models are dynamic and heavily dependent on personal experiences in the real world [9, 31, 49], thus making them unique to every person. According to Collins and Gentner [9], people develop mental models through analogical thinking. When a person experiences an unfamiliar domain, they draw on familiar experiences they perceive as similar. This process involves pulling from prior knowledge in a related schema and importing its relational structure to the domain in question [20, 35]. For example, a mental model of how water flows may be used to explain electrical current; despite its incorrectness, familiar concepts and relations are mapped onto the new model [9]. However, researchers have suggested that the ability to construct appropriate mental models depends on how well the new environment relates to prior knowledge [36, 67].

A general belief in HCI is that people learn to use and understand complex systems by developing mental models [31, 32, 36, 49, 68]. HCI researchers have long used this construct in understanding the behavior of humans interacting with systems [20] and how individuals think a system works or behaves [49]. In addition, it is often used as a thinking tool to design technologies in a way that fits human capabilities and helps consider the different ways users come to understand or misunderstand the systems they interact with [52].

Studies of people developing mental models with computer systems have allowed researchers to begin to understand how these models develop [32, 36, 49] and what activities they engage in [35, 76]. Researchers have asserted that mental model construction involves multiple stages, highlighting the generative and dynamic quality of these models [35, 45, 76]. Other work has found that people instinctively recognize when they need to adapt their mental model [49, 60, 75]. For instance, Yan Zhang [76] found that three mental activities occur when the current model is insufficient: assimilating new information into an existing model, eliminating old information, and adjusting existing information.

A common approach in HCI is to examine how users develop their models based on the response of the system [35, 60, 68, 75]. Several useful theories have been proposed regarding users' approaches when constructing their mental model [35, 45, 68, 75]. For instance, Waern [68] suggested that a *top-down* approach is typically used by experienced users, in which users tap their existing knowledge and modify it based on new information. On the other hand, novice users use a *bottom-up* approach in which information is gradually utilized during interaction to construct a mental model. Savage-Knepshield [60] echoed this finding and found that users form their mental models of familiar computer systems using a top-down process. In our work, we focus on novices of AI systems because they will likely engage in more mental model development than experts or experienced users who already have an established mental model [68].

Despite the fact that HCI has a long tradition of understanding users through the construct of mental models, this approach is still an underutilized area with regard to AI systems, perhaps because the interest in understanding human-AI interaction is only recently emerging [2, 4, 37, 71]. In this paper, we seek to contribute to a better understanding of mental model development in the context of video games. Here, we leverage existing cognitive psychology and HCI theories to understand the differences in how players develop their mental models.

2.2 Mental Models of AI Systems

Although the existing literature is limited, various work has tackled how users model AI systems to develop more human-centered approaches to explainable AI (XAI) [21, 39, 39, 62], human-AI interaction (HAI) [4, 37], and game progression [17, 23]. Studies have primarily focused on how to influence mental model accuracy [7, 37, 38]. For instance, Kulesza et al. [37] studied the effect of accurate mental models of an intelligent music recommender system. They measure users' mental models with a survey and find that a 15-minute tutorial increases model accuracy before using

the system. Their results suggest that scaffolding instruction positively impacts user satisfaction and the usability of debugging the AI system. Other work has examined the impact of AI errors on performance over time. Bansal et al. [4] look at the effect of different kinds of AI errors on user's mental models, using performance (i.e., if the model's decision was accepted or declined by the user) as an indicator of a user's mental model. They found that performance will improve if error boundaries are represented straightforwardly, and users can easily distinguish successes from errors.

For AI systems in games, studies focused on understanding when players' mental models revise and what factors impact model development. Graham et al. [23] performed a pilot study to explore how players develop their mental models of game-embedded AI agents in the game *Command & Conquer Generals* over five days. They quantified players' mental models using a Likert-scaled dissimilarity questionnaire at the beginning and end of the study. They found that only some of the initial models were based on the available surface features of the AI agent in the game, and only some moved away from these surface features to more functional features with additional experience. More recently, Gero et al. [21] studied how people develop their mental model of an AI agent in a cooperative word guessing game. They found that players tended to revise their mental models in the face of anomalies. They also explored how gameplay impacts model accuracy and found that playing more games did not increase the accuracy of the mental model but that players who won more often had more accurate mental models. However, how and why people develop a more accurate mental model of the AI agent remains unclear.

In all these studies, whether AI systems are studied in the context of games or not, mental models are described as dynamic constructs developed and modified over time with experience. However, most existing work studies mental models after interacting with an AI [4, 37, 62] and uses the mental model construct to understand how to influence model accuracy [4, 21, 23, 37]. As such, most existing work does not explore mental model development in response to the AI system over time, which has shown to be crucial in developing human-centered approaches to system design [45, 57] and a better understanding of how mental models develop [35, 68, 75]. In this paper, we extend existing work by examining how users develop their models based on the response of the AI player over time. We study this with the AI-based game *iNNk*, which we describe in the next section.

3 METHODS

This work examines how a pair of players develop mental models of an adversarial AI during gameplay. Towards this end, we conducted a qualitative study to examine how players (n = 20, 10 pairs) develop mental models of AI during gameplay. In this section, we discuss the design of the study and our analysis.

3.1 Participants

For our study, we required participants to be 18 years of age or older, have access to a computer or laptop to play the game, and be able to communicate in written and spoken English. In addition, we screened eligible participants via a survey about their knowledge of Artificial Intelligence (AI), including machine learning and data science. Specifically, we asked participants to self-report their AI experience using a five-point Likert scale ranging from "No knowledge" to "A lot of knowledge," which we adapted from prior work [6]. We excluded any individuals who self-reported having some knowledge (i.e., I have used AI algorithms in my work, or I have taken 1-2 courses on AI) or a lot of knowledge (i.e., I have used AI algorithms frequently in my work, or I am pursuing/have a degree in AI) because the focus of this study is to examine how novices make sense of AI players. As stated earlier, novices will likely engage with more mental model development, which is of key interest to our study, compared to experts.

In total, we recruited 20 students, with 18 reporting having little knowledge of AI (i.e., I know basic concepts about how AI works) and two participants reporting no knowledge of AI. The average age of the participants was 22.8 years (*SD* = 3.9). The gender ratio was 13 female, five male, and two non-binary. In terms of knowledge of coding programs and coding experience, ten participants reported some knowledge and experience and eight participants reported little knowledge (i.e., basic concepts in programming) and limited coding experience. Two participants had no knowledge or coding experience at all. We matched players into pairs based on their self-reported AI experience and availability.

3.2 Adversarial AI Game

Here, we detail the adversarial AI-based game *iNNk* that we used for our study and the Neural Network (NN) model, which is the adversarial AI player in this game.

3.2.1 iNNk. *iNNk* is a web-based multiplayer drawing game where two or more people play together against a Neural Network (NN) [64]. To win the game, the players must successfully communicate a secret codeword to each other through drawings without being deciphered by the NN. Players are assigned one of the two roles during the game: the Sketcher and the Guesser (See Figure 1). The Sketcher is tasked with drawing something based on the codeword assigned by the game. The goal is to draw the codeword so that the human Guesser can interpret the codeword accurately before the NN. The Guessers are tasked with entering their guess of the codeword based on the Sketcher's drawing before the NN guesses correctly. The NN always plays the role of a Guesser, and its goal is to decipher correctly first.

The game is structured around six 30-second rounds. If either the human team or the NN guesses the codeword correctly within 30 seconds, the respective side wins the round. Otherwise, the round will be marked as a tie if neither side guesses correctly within the time limit. A tied round does not count toward the six rounds. The team (i.e., humans or AI) that has the most points out of the six rounds wins the match. The success of the human team hinges on players' recognizing the AI's limitations, making error boundaries a critical component of gameplay. *iNNk* is particularly useful for this study because it pushes players to develop their mental model of the AI player through trial-and-error gameplay, which allows us to gain insight into the mental model process.

For the purposes of our study, players played three matches (i.e., at least 18 rounds in total), with each player playing at least three rounds as the Sketcher and Guesser roles in each match. The generation of codewords was randomized for each group so the researchers could holistically examine mental model development in an unpredictable environment that mimics actual gameplay as close as possible. Our intention was not to compare mental model accuracy. The majority of the words (i.e., 345 words) are of similar difficulty (i.e., bus, cat, donut); however, we deemed 5% of these words of a higher level of difficulty (i.e., animal migration, camouflage). In line with this percentage, we observed that 4.8% of 247 rounds played in this study had more difficult words. Due to their low occurrence, we did not consider the difficulty level in our analysis.

3.2.2 Neural Network. *iNNk* uses deep learning as the framework for its AI system. Specifically, it was built using Google's *Quick Draw!* [22] NN architecture. The model was trained on hand-labeled sketch data from a canvas similar to the one used in the game. This data was taken from Google's publicly available *Quick Draw!* [25] dataset and includes 40 million drawings across 345 categories (i.e., 345 supported codewords) of example sketches. For more detailed information on the NN's architecture, see [43].

The model starts to make predictions (i.e., internal guesses) from the moment when the Sketcher makes the first stroke on the canvas. The categorization label with the highest predicted confidence constitutes the guess of the NN. The NN continues to generate guesses; however, they are only presented to the players once a guess is above a certain

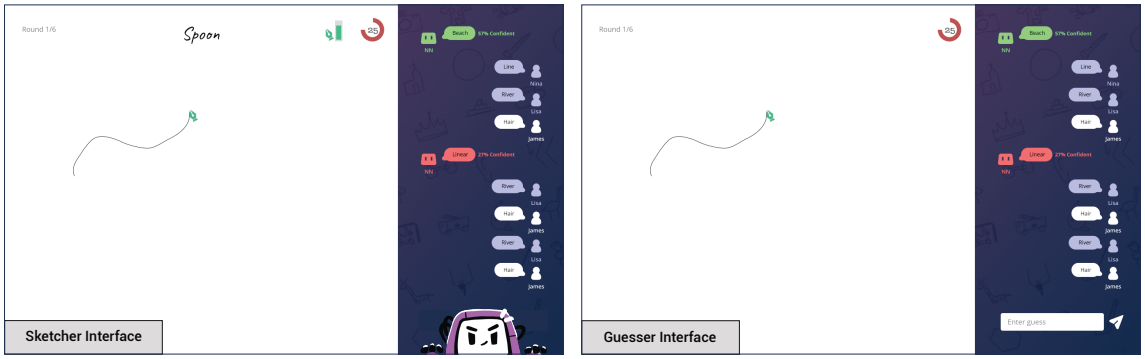

Fig. 1. The image on the left shows the interface for the Sketcher. The Sketcher is tasked with drawing the codeword on the white canvas. In the Sketcher interface, the player has access to the codeword, the guesses made by both the human and AI player. The image on the right shows the interface for the Guesser. The Guesser is tasked with correctly typing in the codeword from the Sketcher's drawing before the NN. In the Guesser interface, the player has access to a guess input box and the guesses made by both the human and AI player. Both interfaces display the AI player's confidence percentage next to its guess.

confidence value. The NN will note previous incorrect guesses by both the NN and the human players and are not used in future guesses during that round. In this way, the NN is able to participate in a way that mimics the other human Guessers.

Important in the context of interpreting our study results is to emphasize the following about the NN-based adversarial AI player: (1) it only makes guesses based on the training data and, thus, does not use online learning in which it will consider new drawings from players in real-time; (2) the sketches in its dataset are human-made digital drawings, (3) its guessing strategy does not change or adapt; instead, it tracks incorrect guesses made by both the NN and the human players and will not use these words in future guesses for the same round; and (4) it generates a prediction based on all the strokes created on the entire canvas every 2.5 seconds, and the guess with the highest predicted confidence is sent to the game every 5 seconds.

3.3 Procedure

Due to the COVID-19 pandemic, we conducted this study remotely over a university zoom account. The sessions included one researcher and two participants. Sessions took 60 to 90 minutes, which included a pre-gameplay survey, three matches of the game, and a short semi-structured interview after each match. University IRB approved the study protocol beforehand.

Before gameplay, participants completed a pre-gameplay survey which gathered demographic information, technical literacy, and self-reported understanding of how they think Neural Networks (NNs) recognize images. To evaluate technical literacy, we asked participants to self-report their programming experience using a 5-point Likert scale ranging from "No knowledge" to "A lot of knowledge," which is similar to the AI screening question adapted from [6]. Then, they were asked to briefly elaborate on their current knowledge of how NN recognizes images in an open text field.

Participants were given instructions for playing the game and instructed to think aloud and find different ways to beat the AI. Each group then played three matches (at least 18 rounds total). The total number of rounds would vary as some sessions would tie more than others, and a tied match did not count toward a score. We recorded the audio and video of this part of the session with the participants' permission.

After playing a match, participants engaged in a debriefing interview with the researcher. Participants were asked a series of questions designed to elicit information about how they thought the AI worked, what the AI reminded them of, if anything, and elaborate on a failed and successful round against the AI player for when they were the Sketcher. If the group did not lose or win a round, they were asked to elaborate on a tied match. The decision to ask these questions was to provide another opportunity to elicit their mental model.

At the end of the interview, participants were asked to describe their strategies during the match and if they thought it was effective at beating the AI. If they believed it was effective, the researcher asked the group not to use this strategy in the next match. This choice was to ensure groups did not stall the development of their mental models once they found a working strategy and to increase the number of observable behaviors to analyze. If participants were unsure if their strategy was effective or if there was disagreement (i.e., one player thought it was effective and the other did not), the researcher did not ask the group to exclude the strategy for the next match.

3.4 Data Analysis

Here, we describe our data preparation regarding how the data was transcribed and visualized for analysis. Next, we break down our analysis into two phases. First, inspired by HCI theories, we analyze the mental model activities and AI concepts across each round. Second, we analyze the players' development for each session as a whole. As a reminder, a *session* consists of three matches of the game. A *match* consists of at least six rounds (more if there are ties). A *round* is one instance in the game where a secret codeword needs to be drawn by the Sketcher and guessed correctly by either the human Guesser or the AI player.

We further note that the coding process described below involved all authors and was led by the researcher who conducted and transcribed all sessions. We coded primarily by reviewing the data visualization discussed below but referenced the raw data when clarification or confirmation was needed. Outcomes of the analysis were discussed in order to reach consensus [26, 55]. We refer to a player from a specific session by a label that first states if it is Player 1 or 2 and then the session number. For example, "P2-3" is Player 2 from Session 3.

3.4.1 Data Preparation. To conduct our analysis, one researcher transcribed all ten session's audio recordings and transferred the gameplay data into a diagram (see Figure 2) using Miro, a collaborative digital whiteboard tool. The diagram visualizes the player-AI interaction of a session, which includes the cycle of each round of (1) player input (i.e., drawing), (2) the AI output (i.e., the AI guesses), (3) the game outcome (i.e., win/loss), and (4) the utterances that corresponded to expressions of a participant's mental model of the AI agent, expressed while playing (i.e., think aloud) or after the match completed during the semi-structured interview. This data visualization facilitated analyzing how a pair of players developed their mental model over time.

To build this data visualization, first, screenshots were taken of each drawing from the video file and added to the diagram, organized by match and the corresponding player when they played the Sketcher role. Second, codewords, the AI's guesses, and the round outcome (i.e., humans win or AI wins) were taken from the video file and marked on each drawing. Third, the researcher read all transcriptions at least twice. On the second reading, the researcher took notes of pertinent utterances that corresponded to expressions of a participant's mental model and referenced the AI player (i.e., "I want to see how smart this AI really is" (P2-1). These utterances were then added to the diagram, distinguishing if these were made during a match (via think aloud) or the interview. Building the data visualization was an iterative process involving two other researchers who reviewed the transcripts and the evolving diagrams to determine if the data visualization comprehensively captured the development of mental models for all ten sessions.

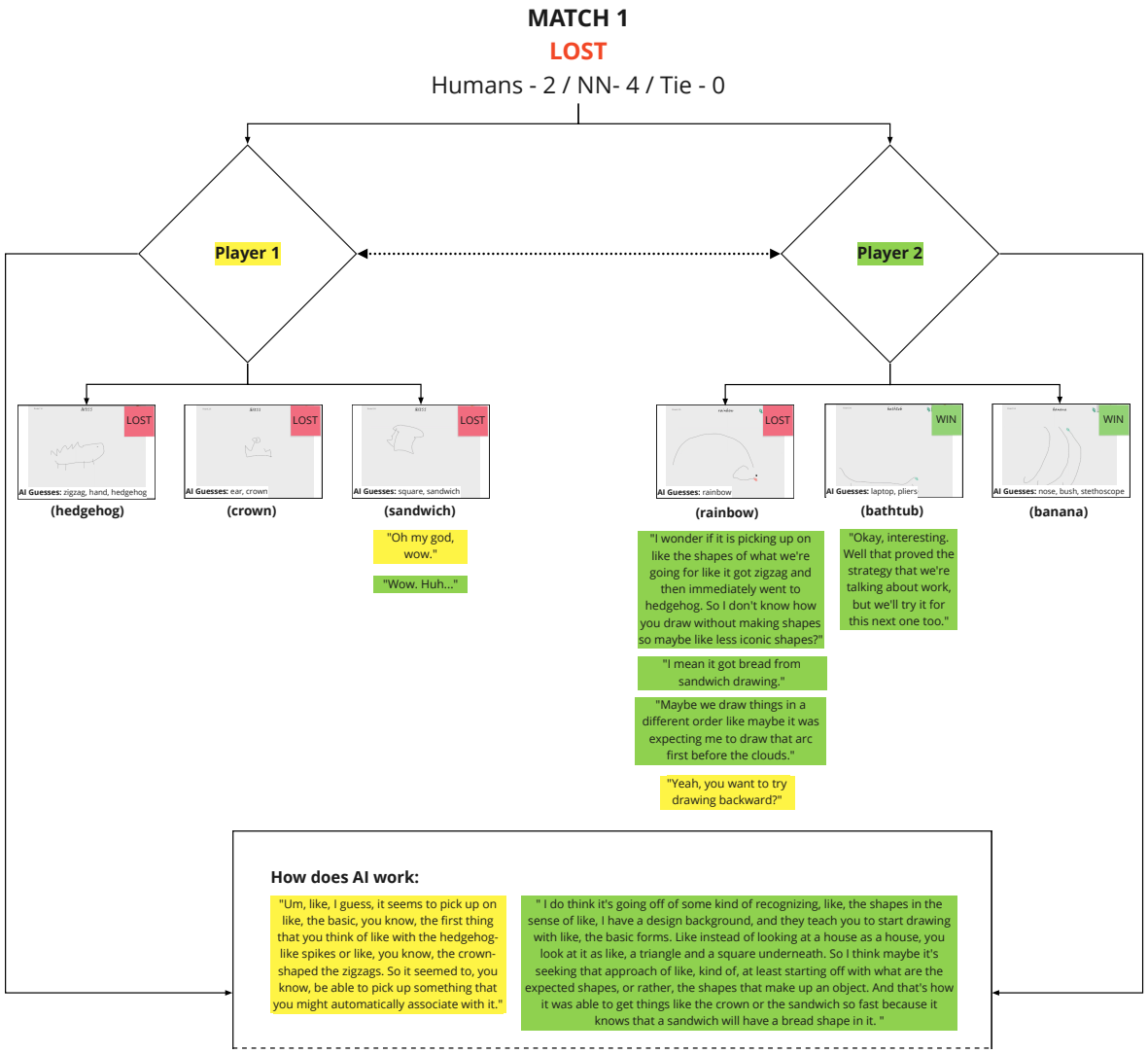

Fig. 2. This screenshot is an example diagram that visualizes the player-AI interaction and part of the interview for Session 2's first match. The top of the diagram describes the match number (1 of 3), the outcome (lost), and the score, indicating how many rounds the human team and the NN won and any tied matches. The diagram is split in half, visualizing player 1's (highlighted in yellow) drawings on the left and player 2's (highlighted in green) on the right. Each drawing includes if the human team won or lost the round, which is indicated by green or red squares in the upper right corner of the drawing and the AI's guesses in the order that the AI sent them in the bottom left corner of the drawing. In addition, under each drawing, the diagram includes the provided codeword. Below each round, we display the utterances made by both players (highlighted in their corresponding player color) that are expressions of the player's mental model of the AI player during gameplay. Finally, in the bottom box below all the individual rounds, we include the utterances from the interview. Note that this figure only displays the utterances in response to the first interview question. For all interview questions, see Section 3.3 or the entire diagram can be reviewed in the supplementary material.

3.4.2 Phase 1: Round Analysis. Once the data visualization was completed and to begin our mental model analysis, we referenced existing cognitive development and HCI literature [66, 68] to create an initial codebook for mental model

development (i.e., focus, assimilation, and accommodation). The codes, definitions, and examples can be found in the following section. Using each round of gameplay as the unit of analysis, we then collectively coded the focus of the development (i.e., top-down vs. bottom-up) and the mental model activities, which refer to if and when the group changed their mental model (i.e., assimilation vs. accommodation) to map the progression of development across each match.

In addition to the initial codebook, we coded *AI concepts*. We refer to an AI concept as a verbalized notion about the AI that encapsulates hypotheses or conclusions regarding how the AI works. For example, Session 2 began their second match by investigating how the AI uses the canvas: "I don't know if that is at all part of the Neural Network. But what if we tried drawing on a different part of the canvas . . . like the upper corners or something?" (P2-2). The corresponding concept we applied for this is "AI uses the center of the canvas." An AI concept can be elaborated upon or associated with similar AI concepts. In the case of Session 2, the players continued to investigate ideas with the aforementioned concept (i.e., drawing in the corners, drawing tall, drawing wide).

3.4.3 Initial Codebook. In her article *On The Dynamics of Mental Models*, Waern [68] emphasizes that when examining mental models, an important aspect concerns how users select elements of the observed situation to begin construction. While she identified seven events that occur in the construction process of a mental model, we focus on the first event, *Intention and Attention*, as other events in the framework (i.e., evocation of prior knowledge, memorization) are beyond the scope of this work. We use Waern's *top-down* and *bottom-up* distinction as a lens to understand the focus of mental model development. We refer to focus as to what players pay attention to for the development of their mental model, either attending to gameplay observations (i.e., bottom-up) or utilizing prior knowledge or experiences (i.e., top-down).

We adapt Waern's [68] definitions to our context by emphasizing whether players generate a concept about the AI based on gameplay observations (i.e., win/loss, AI guesses) or by evoking prior knowledge (i.e., databases) to generate a concept about the AI player. For example, Session 2 constructed a concept about the AI from observing gameplay "[I] don't know if that is at all part of the neural network. But what if we tried drawing on a different part of the canvas...in the upper corners or something?" (P2-2). The corresponding AI concept is: AI uses the center of the canvas. The original and adapted definitions can be seen in Table 1.

An essential aspect of understanding how mental models develop is examining how they change [35, 51, 68, 76]. One approach to understanding changes in mental models is examining if and when new information is assimilated or accommodated. These terms—assimilation and accommodation—were originally proposed by Piaget [66] who is well-known for his observations and intellectual contributions regarding the cognitive development of children. In both cognitive development and HCI literature, it is widely recognized that people can assimilate new information into an existing schema [59, 66, 68, 75]. Or, when no appropriate schemata exist, new schemata can be created to accommodate the new information by modifying an existing schema or creating a new one [59, 66].

As a first step in understanding how change happens across our data, we utilize Piaget's [66] theory of assimilation and accommodation. These constructs have been shown to be useful for understanding a user's mental model with a new system and have been applied to work in HCI previously [1, 5, 15]. We adapt these definitions to our context by emphasizing whether players integrated and maintained a concept about the AI player they were addressing in gameplay (i.e., assimilation) or changed the concept due to a discrepancy (i.e., accommodation). The original and adapted definitions can be seen in Table 1.

3.4.4 Phase 2: Session Analysis. After applying the initial codebook across each session, we coded the pair of players *mental model development* by reviewing the progression of mental model activities and concepts for each session

Table 1. This table includes the initial codes for the focus of the development (i.e., top-down vs. bottom-up) and the mental model activities, which refer to if and when the group changed their mental model (i.e., assimilation vs. accommodation) to map the progression of development across each match. We include the original and adapted definition and citations.

Code	Original & Adapted Definition	Example
Focus	Original definitions: • "A learner who builds a conceptual model of the system solely on basis on the experiences of interactions with the system can be regarded to use a bottom-up learning approach." (pp. 74-75). • "A learner who builds a conceptual model of the system on basis on his expectations of the system, derived from his prior knowledge of similar tasks or systems, can be regarded to use a top-down learning approach." (p. 75). Adapted definitions: • *Bottom-up*: Players build a mental model of the AI player from gameplay observations: "I wonder if it is picking up on like the shapes . . . it got zigzag and then immediately went to hedgehog" (P2-2). • *Top-down*: Players tap into their existing knowledge to develop their mental model of the AI player and modify it based on gameplay observations. "Definitely uses trends recognition. I'm assuming . . . the database is built up from playing against other humans" (P2-9).	Session 2 constructed a concept about the AI from observing gameplay (i.e., bottom-up) "[I] don't know if that is at all part of the neural network. But what if we tried drawing on a different part of the canvas . . . in the upper corners or something?" (P2-2) The corresponding AI concept is: AI uses the center of the canvas.
Assimilation	Original definition[66]: • "Assimilation is the cognitive process by which a person integrates new perceptual, motor, or conceptual matter into an existing schemata or patterns of behavior. " (p. 14). Adapted definition: • Players assimilate by adding a concept about the AI agent into their mental model on success or failure. Players may also assimilate by expanding on a concept by adjusting it or revising it on failure. If a player maintains the concept they are addressing and does not change it; we consider it assimilation.	Session 2 assimilated in response to a failure that the "AI uses the corners of the canvas" after investigating multiple ideas i.e., drawing in the corners, drawing wide, drawing tall) within the original constructed concept: AI uses the center of the canvas.
Accommodation	Original definition [66]: • "Accommodation is the creation of new schemata or the modification of old schemata. Both actions result in a change in, or development of, cognitive structures. " (p. 15). Adapted definition: • Players accommodate by changing their mental model on success or failure to restore balance when a discrepancy occurs. Players change their mental model to account for new information.	Session 2 accommodated that the AI can instead predict. P1-2 "[It] was just the beginning of the image, and yet it was able to fill it in." P2-2 "I feel like it's also able to somehow predict."

as a whole. Here, we found that the AI concepts that participants expressed during the post-match interview were particularly helpful in understanding their development and used these as a basis to code the differences (see Table 2). In one instance, with Session 1, we observed clearly that the two players have different mental model development and reported them separately.

All these steps, first coding the individual rounds in Phase 1 (i.e., top-down vs. bottom-up, assimilation vs. accommodation and AI concepts) and then the entire session in Phase 2 (i.e., development), led to establishing our codebook of mental model development, see Section 4.1. Following this, we extracted key patterns of the observed

mental model development to answer our research question. This process involved leveraging the codebook to examine what dimensions played a role in the development and interrelating the different codes. The result of this phase is our framework of mental development, see Section 4.2.

4 RESULTS

In this section, we first present our codebook of mental development, which describes how we observed mental development in the context of a pair of players facing an adversarial AI player (using the focus and the mental model activities as perspectives on how to interpret mental model development). Following this, we present our framework based on examining how the codes from the codebook interrelate and what codes emerge to be most dominant in understanding mental model development.

4.1 Codebook of Mental Model Development

For establishing our codebook, we first considered the focus of development and the mental model activities based on existing literature (see Table 1). Then, based on the results of our Phase 1 analysis, we looked at the overall development. We discuss the codes for each below. Table 2 provides an overview of the results applied to the different sessions.

4.1.1 Mental Model Focus. As described in Table 1, we distinguish the focus of the development into top-down and bottom-up. The original two codes derived from Waern's [68] work were sufficient to describe what players pay attention to for the development, either attending to gameplay observations (i.e., bottom-up) or utilizing prior knowledge (i.e., top-down) to develop their mental model.

4.1.2 Mental Model Style. As described in Table 1, we determine if and when the group changed their mental model by distinguishing mental model activities into assimilation and accommodation. For assimilation, two main codes were derived from examining how groups assimilated new concepts, particularly if the groups built on and utilized previously assimilated concepts (i.e., systematic) or not (i.e., reactive). We refer to these codes as *Styles* to describe how development differs in terms of integrating new information.

- *Systematic*: Players are systematic in their mental model development and build on previous observations that were integrated into their mental model. For example, Session 3 constructed from gameplay that the "AI has trouble with multiple drawings" and "looks at the overall shape". Then, they iterated on this understanding by hypothesizing different ways to disrupt the overall shape (i.e., upside down, drawing stylized, adding different shapes) and assimilated new information in relation to how the AI looks at the overall shape (i.e., "AI interprets the drawing right-side-up").
- *Reactive*: Players are reactive in their mental model development and do not build on previous observations. Instead, they attend to the moment and integrate new information independent of what was previously observed. For example, Session 2 constructed new AI concepts and did not build on what they previously assimilated from prior matches. In the first match, they constructed that the "AI picks up on iconic/basic shapes" and did not utilize this in the following match. Instead, they constructed a new AI concept from gameplay observations (i.e., "AI uses the center of the canvas", transitioning to "AI uses the corners of the canvas", and ending with "AI can predict").

Table 2. This table is an overview of the ten sessions with regards to the number of wins out of three matches and the key codes that describe a session's mental model development. This includes the focus of the development (i.e., top-down vs. bottom-up focus) and the development style in terms of how development differs when players integrate new information into their mental model (i.e., systematic vs. reactive style) and the development in general. For the latter, we carefully considered the AI concepts that the players expressed during the post-match debriefings and noted these for transparency. We also indicate how players process a change when a discrepancy occurs, which we observed occur through comparisons (i.e., compare to other systems, compare to gameplay, compare to self). The supplementary material contains a larger table with additional participant characteristics.

Session	# of Wins	Focus	Style	Comparisons	Development AI Concepts from Debriefing
P1-1	2/3	Top-Down	Reactive	Compare to Self	**AI functions differently than originally attributed** AI Concepts: (1) AI uses pattern recognition, (2) AI narrows guesses via Confidence meter
P2-1					**AI functions as expected** AI Concepts: (1) AI uses deductive reasoning and learns off previous images
P1-2	0/3 (1 tie)	Bottom-Up	Reactive	Compare to Other System	**AI is more powerful than initially attributed** AI Concepts: (1) AI picks up basic shapes, (2) AI can predict, (3) AI learns strategies
P2-2					
P1-3	3/3	Bottom-Up	Systematic	Compare to Self	**AI interprets the drawing differently than initially attributed** AI Concepts: (1) AI looks at the overall shape, (2) AI examines the image right-side up, (3) AI has trouble with stylistic drawings, (4) Scale does not impact the AI
P2-3					
P1-4	2/3 (1 tie)	Top-Down	Systematic	Compare to Gameplay	**AI references a specific type of data than initially attributed** AI Concepts: (1) AI recognizes patterns and compares to image references, (2) AI uses the whole drawing, (3) AI references photorealistic images and more complete drawings
P2-4					
P1-5	0/3 (3 ties)	Top-Down	Systematic	Compare to Gameplay	**AI references a specific type of data than initially attributed** AI Concepts: (1) AI recognizes patterns using the drawing as a whole and compares to database of images, (2) AI's database is made up of other human drawings and more complete drawings
P2-5					
P1-6	2/3	Bottom-Up	Reactive	Compare to Other System	**AI is more powerful than initially attributed** AI Concepts: (1) AI makes its guess on outer shape of the drawing, (2) AI gets faster, (3) AI can predict
P2-6					
P1-7	1/3 (2 tie)	Bottom-Up	Systematic	Compare to Gameplay	**AI interprets the drawing differently than initially attributed** AI Concepts: (1) AI recognizes patterns by tracing the image to make shapes and compares to other images, (2) AI uses the whole drawing, (3) AI deems the largest part of the drawing as most important and focuses on this area of the drawing
P2-7					
P1-8	2/3	Top-Down	Systematic	Compare to Gameplay	**AI functions differently than initially attributed** AI Concepts: (1) AI recognizes shapes in the drawing by looking for edges and positive/negative space, (2) AI picks up patterns and then works off a category of words associated with that pattern, (3) AI looks at the whole canvas
P2-8					
P1-9	3/3	Top-Down	Systematic	Compare to Gameplay	**AI interprets the drawing and functions differently than initially attributed** AI Concepts: (1) AI recognizes trends and compares it to a bank of images and learns over time, (2) AI's database is made up of other human drawings, (3) AI deems the largest part of the drawing as most important and focuses on this area of the drawing, (4) AI does not learn over time, (5) AI uses a word bank
P2-9					
P1-10	1/3 (1 Tie)	Top-Down	Reactive	Compare to Self	**AI functions differently than originally attributed** AI Concepts: (1) AI recognizes the general shape and compares the drawing to a library of images on the internet, (2) AI uses a catalog of word prompts that are organized by shapes, (3) AI does not use images from the internet
P2-10					

Finally, for accommodation, codes were derived from examining what players said and did in response to a discrepancy in their mental model and how they processed this change to restore balance, which we observe occur through comparisons. We identified the following three main ways this activity took place:

- *Compare to other systems*: Players accommodate new information by comparing gameplay observations to another system or functionality. For example, P1-2: "I feel like it . . . learned our strategy . . . [Like] an automatic automated chess player . . . it's able to learn your moves and able to defeat you by seeing your patterns."
- *Compare to gameplay*: Players accommodate new information based on comparing gameplay from earlier matches in relation to the current results: "The person in that one is smaller than the postcard, which is different than round 1 and 2 where the person is big. So it could be something to do with what the AI determines as important." (P2-9).
- *Compare to self* : Players accommodate new information by comparing how humans and AI players perceive gameplay based on the similarities or dissimilarities identified. For example, consider P2-3: "The scale of the [drawing] . . . it doesn't work like human perception, right? . . . [It's] not thinking in inches or meters or any specific unit."

4.1.3 Development. Development codes were derived from examining the pair's debriefing responses across each match to capture their development experience. We specifically considered here how the AI concepts progressed over time. We intended to capture the pair's development experience with a single code. For example, in Session 2, the players started with the AI concept "AI picks up iconic/basic shapes" and progressed from "AI can predict" to "AI learns strategies." As such, this pair of players attributed more capabilities to the AI over time, and we described this as "AI is more powerful than initially attributed." We note that in Session 1, the two players seemed to think differently about the AI, and we assigned a different development code for each player in that session. In total, we identified five different ways the players' development was experienced (i.e., "AI is more powerful than initially attributed.")

- *AI functions differently than initially attributed*: For example, Session 1's Player 2 started with the AI concept "AI uses pattern recognition" and progressed to "AI narrows guesses via confidence meter." These groups attribute different functionalities to the AI player over time.
- *AI functions as expected*: This code relates only to Session 1. Session 1's Player 1 started with the AI concept "AI uses deductive reasoning and learns off previous images" and maintained this throughout the matches.
- *AI is more powerful than initially attributed*: For example, Session 2 started with the AI concept "AI picks up iconic/basic shapes," and progressed to "AI can predict," ending on "AI learns strategies." These groups attribute more capabilities to the AI player over time.
- *AI interprets the drawing differently than initially attributed*: For example, Session 7 started with the AI concept "AI recognizes patterns by tracing the image to make shapes and compares to other images," then progressed to "AI uses the whole drawing," ending with "AI deems the largest part of the drawing as most important and focuses on this area of the drawing." These groups attribute specific ways the AI deciphers the drawing over time.
- *AI references a specific type of data than initially attributed*: For example, Session 5 started with the AI concept "AI recognizes patterns using the drawing as a whole and compares to a database of images" and progressed to "AI's database is made up of other human drawings and more complete drawings." These groups attribute specific characteristics to the data the AI player references over time.

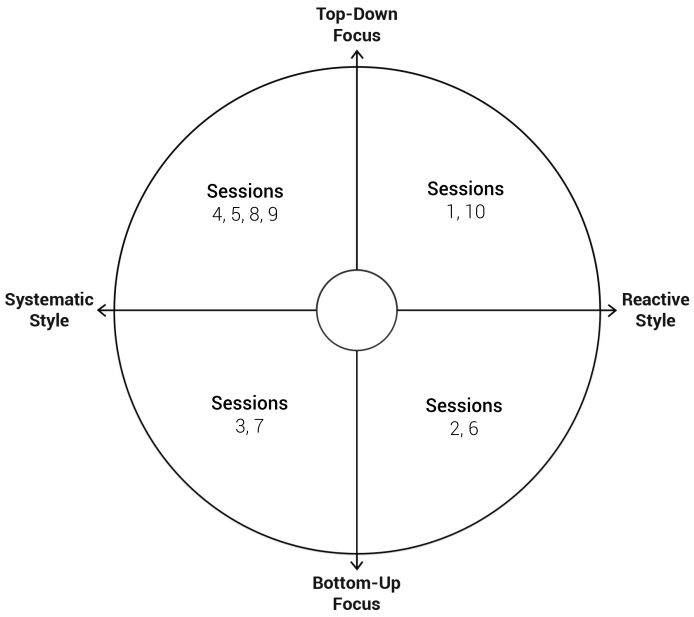

Fig. 3. Distribution of ten gameplay sessions (n = 20, 10 pairs) categorized by the focus of development (i.e., top-down vs. bottom-up) and the style of development (i.e., systematic vs. reactive).

4.2 A Framework of Mental Model Development

After establishing our codebook and reviewing the development of the ten sessions, we find that the codes for the focus and the style of development best describe how players' are developing their mental model. As such, we recognize the first dominant dimension as focus, where players either try to beat the AI player based on an existing model (i.e., top-down focus) throughout their experience or attend to the game experience and base construction on what they observe during gameplay (i.e., bottom-up focus). The second dominant dimension we observed is that players integrate new information in a *systematic* or *reactive* style. We further describe this dimension with how players process a change when a discrepancy occurs, which we observe occur through comparisons (i.e., compare to other systems, compare to gameplay, compare to self). Finally, we also find that the development experience is related to the focus dimension of the model development. Below, we describe the two development dimensions that form our framework and illustrate each with particular sessions. For an overview of each session with the applied key codes, see Figure 3.

Unfortunately

4.2.1 Top-Down Focus. Our analysis found that six groups developed their mental model of AI by evoking prior knowledge and assimilating new information from gameplay observations (i.e., win/loss, AI guesses). For example, Session 5 initially described that the AI works by "recognizing patterns and compares the drawing to what it has already seen." Then, using this mental model, they assimilated new information about the AI over time from gameplay observations. For instance, they assimilated that the AI's database contains "other human drawings" after experiencing a failed round. Finally, they reasoned about the AI using this failure in the context of their existing mental model: "I'm starting to think that the data is collected off of like a similar game sort of thing . . . it's based off of drawings because

I'm looking back at the motorcycle [round] and it looks nothing at all like a motorcycle . . . I feel like that's how most people would start drawing one." (P1-5).

Of the groups coded with a *top-down focus*, we found three types of development: (1) AI references a specific type of data than initially attributed, (2) AI functions differently than initially attributed, and (3) AI functions as expected. Groups in this category would get more specific within the context of the existing mental model. For example, Sessions 5 and 9 had a prior understanding of databases and, over time, assimilated that the AI player's database contains "human drawings". In contrast, Session 4 concluded that the database contains more "photorealistic images" after experiencing consecutive wins by drawing simplistically. Other groups would change their assumption that the AI player functions differently than initially attributed. For example, Session 1 had prior knowledge of pattern recognition and accommodated on failure at the end of the final match that the AI narrows guesses with a confidence meter: "I think what it's doing is it's getting a level of confidence . . . somehow using that percentage to get closer to answers . . ." (P2-1). However, P1-1's mental model remained consistent despite their teammate changing their mental model: "Same answer as last time, I think just take images and learns from them."

4.2.2 Bottom-Up Focus. Our analysis found that four groups developed their mental model of AI by attending to the game experience and assimilating new information from gameplay observations (i.e., win/loss, AI guesses). For example, Session 2 constructed the concept "AI picks up on iconic shapes" from experiencing consecutive failures: "I wonder if it is picking up on like the shapes . . . it got zigzag and then immediately went to hedgehog" (P2-2).

Of the groups coded as using a *bottom-up focus*, we found two types of development: (1) AI is more powerful than initially attributed, and (2) AI interprets the drawing differently than initially attributed. Groups coded as "AI is more powerful than initially attributed" would draw more general conclusions about the AI player over time. For example, Session 2 constructed the concept "AI uses the center of the canvas," assimilated in response to failure that the "AI uses the corners of the canvas," and accommodated after another failure by *compare to other systems*: ". . . I feel like it's also able to somehow predict" (P2-2). P1-2 responded by saying: "I totally see that . . . it was just the beginning of the image, and yet it was able to, to fill it in."

In contrast, groups coded as "AI interprets the drawing differently than initially attributed" would get more specific regarding how the AI deciphers the drawing to make its guess. For example, Session 3 constructed a mental model that the "AI looks at the overall shape." Over time, they assimilated that the AI interprets the drawing right-side-up in response to consecutive wins by drawing the codeword upside down: "Um, surprisingly, I think the AI is only evaluating images right-side-up . . . ." (P2-3). P1-3 confirmed this AI concept: "Being able to draw upside down and beating the AI every time shows that . . . it does not take into consideration orientation."

4.2.3 Systematic Style. Six groups were coded as systematically developing their mental model. Of these six groups, two were coded as using *bottom-up focus*, and four were coded as using *top-down focus*. We use Session 9 (i.e., top-down focus, systematic style) as a detailed example. Session 9 started the match by using their existing mental model to collaboratively generate ways to defeat the AI (i.e., draw half the codeword, draw the environment/context of the codeword). Next, they constructed the AI concept "AI is good at obvious drawings of the codeword" after a failed round. Finally, they assimilated this information into an existing mental model. They described in the debriefing that the "AI recognizes trends and compares it to a bank of images" and that the AI might be "learning over time."

They built on the previously assimilated concepts in the second match by not drawing the codeword in an obvious way and elaborating on this by drawing additional shapes that represent the context or environment of the codeword. They assimilated the AI concept "AI has trouble with multiple drawings" after winning and on failure that the "AI's

database consists of other human drawings." In the debriefing, they concluded "[The] database is built up from playing against other humans" (P2-9).

In the third match, they continue to build on the previous mental model of "AI has trouble with multiple drawings" by hypothesizing ways to disrupt the overall shape. After experiencing a failed round, they accommodated by referencing earlier matches: "The person in that one is smaller than the postcard, which is different than round 1 and 2 where the person is big. So it could be something to do with what the AI determines as important" (P2-9). Based on this, they accommodated to include "AI deems the largest part of the drawing as most important and focuses on this area of the drawing" to make its guess. They elaborate on this in their gameplay by drawing larger distractions. At the end, they assimilated the AI concepts "AI does not learn over time" and "AI uses a word bank." Session 9 built on and utilized previous observations that were integrated into their model.

Groups coded as *top-down focus* and using a *systematic style* processed a change in their mental model by comparing to gameplay. When a conflict occurred, they would use *compare to gameplay* by comparing earlier rounds with the current results. In contrast, groups coded as *bottom-up focus* and using a *systematic style* were observed using *compare to gameplay* or *compare to self*. For example, Session 3 used *compare to self* after experiencing a failure: "The scale of the [drawing] . . . it doesn't work like human perception, right? . . . A hot air balloon is huge. If you draw it really small, you would never consider a huge thing to be drawn that small. Where the AI, I think in this way it has an advantage because it's still looking at the exact same patterns . . .[It's] not thinking in inches or meters or any specific unit."

4.2.4 Reactive Style. Four groups were coded as reactive in developing their mental model. Of these four groups, two were coded as *bottom-up focus*, and two were coded as *top-down focus*. We use Session 2 (i.e., bottom-up focus, reactive style) as a detailed example. After consecutive failures, Session 2 first constructed a concept from gameplay observations about the AI player that "AI picks up on iconic shapes." As P2-2 stated: "I wonder if it is picking up on like the shapes . . . it got zigzag and then immediately went to hedgehog." Next, they elaborate on this concept in their gameplay by changing the order they drew, saving the most notable parts of a codeword for last. Finally, they assimilate that the "AI picks up on iconic/basic shapes" based on consecutive wins. They concluded in the debriefing: "[It] seems to pick up on like, the basic [shapes] . . . So it seemed to, you know, be able to pick up something that you might automatically associate with it" (P1-2).

Starting match two, they construct a new AI concept from gameplay observations that the "AI uses the center of the canvas." Here, P2-2 mentions: "I don't know if that is at all part of the Neural Network. But what if we tried drawing on a different part of the canvas . . . like the upper corners or something?" Next, they assimilated in response to a failure that the "AI uses the corners of the canvas" and then accommodated after another failure by matching to another system by concluding in the debriefing that the "AI can predict." P2-2 comments: ". . . I feel like it's also able to somehow predict." P1-2 responded by saying: "I totally see that . . . it was just the beginning of the image, and yet it was able to, to fill it in." They end the match by constructing another new concept about the AI "Drawing upside down tricks the AI" and assimilated this concept on success.

Starting the third match, they continued to draw upside down and accommodated after a single failure by matching to another system: "Oh, okay. So maybe it learned?" (P2-2) changing the AI concept to include "AI learns strategies." They continued by constructing a new concept from previous gameplay, "Drawing in one line stumps the AI" and iterated on this concept until constructing a new concept, "Drawing related symbols," which they continued to elaborate on until the end of the match. They concluded in the debriefing that the AI could also learn: "Um, so I think what it's going through, at least what it's cycling through, is not only that predictive thing we were talking about before. But

like P1 says I think it is kind of learning in the sense" (P2-2). P1-2 then responds with: "I feel like it . . . learned our strategy . . . [Like] an automatic automated chess player . . . it's able to learn your moves and able to defeat you by seeing your patterns." Session 2 attended to the moment and integrated new information independent of what was previously observed.

Groups coded as *top-down focus* and using a *reactive style* processed a change in their mental model by comparing to self. For example, Session 1 processed a change after a failure by reasoning about the AI using their own perception: "I was thinking if I separated the images that it wouldn't get it, but it must have been like, oh, I see stripes I see a horse I see . . ." (P2-1). P1-1 continued this comparison: "Like, oh, if I see a plus sign, maybe that means like, oh, I'm adding this thing on the left with this thing on the right . . ." In contrast, groups coded as *bottom-up focus* and using a *reactive style* used *compare to other systems*. For example, Session 6 changed their mental model in response to consecutive failures that the AI could predict: "I think it just got faster. . . because like up until now we could at least like to draw a part of it and it wouldn't guess it but now even before we draw a part of it . . . I guess it is kind of predicted how we're drawing" (P2-6).

5 DISCUSSION

In this section, we present how our framework can be generalized to the design of other adversarial AI-based games through the evaluation or design of such games. We use *iNNk* as an example. Finally, we describe design implications for game designers and HCI researchers interested in how we may better examine and design for mental models of AI.

5.1 Applying the Framework

Anticipating how different players may model AI players during gameplay and how they may change their models over time is a difficult challenge [17, 21, 23]. However, our framework can be used as a thinking tool, either starting from an existing AI implementation to evaluate the impact on players' mental model development or using it for design thinking on how the game's design can facilitate model development. Below, we offer some initial guidance on these efforts for other adversarial AI-based games.

5.1.1 Evaluation. It is common in traditional UX design to evaluate designs without involving users in which designers have to imagine or model how a design is likely to be used [30, 57]. A common UX method, cognitive walkthrough, allows designers to step through a prototype and answer a set of questions to understand the design's impact on a user's learning of the system [57]. We suggest using this method to evaluate how an existing AI-based game or prototype may impact players' mental models. Researchers have suggested adapting UX methods to accommodate AI [48, 71]. However, to what extent traditional UX methods need to be adapted for AI-based games remains unclear. Therefore, we suggest that game designers use this method as a starting point.

Game designers conducting a cognitive walkthrough of their AI-based game can simulate a player's mental model at each step in the player-AI interaction (i.e., player action and AI response) from the point of view of our framework. Specifically, designers can examine the potential impact on how the player's mental model progresses by answering a set of questions. For example, in the context of *iNNk*, designers may stop after each round and ask: (1) What concept(s) would players with a bottom-up focus pick up about the AI player from gameplay observations? (2) What other systems or prior experiences may players with a top-down focus reference to construct a concept about the AI player? (3) How will players using a systematic style build on previous observations in their gameplay? (3) How will players using a reactive style attend to the current moment and construct new observations from gameplay? By answering these

questions, designers may better anticipate changes in mental model development throughout the player experience to identify potential misconceptions and better tailor design decisions.

5.1.2 Design. Researchers have suggested that mental model development can inform game progression [17, 23]. For example, Graham et al. [23] and echoed by Furlough et al. [17] suggest that instead of designing a set of difficulty levels for all players, progression should be designed around requiring the player to adapt their current mental model by designing for mental model accommodation and assimilation. However, how to begin the design process remains unclear, which is a challenge since designers often begin the design process by understanding player behavior or their motivations for play [73, 74] to provide a lens for design. Our framework can be used as a tool to start design thinking.

For example, in the game *iNNk*, the designer may want players that use a *reactive style* to maintain and build off a previously integrated concept, such as the "AI picks up on iconic/basic shapes" to deter new construction. In doing so, designers may consider redesigning the conditions to draw the codeword to help counter new construction and reinforce this observation. For instance, possible conditions might be to *fill in the blank* where part of a drawing is already present on the board. This example condition could help encourage reactive players' to focus on the key features and shapes of the codeword.

In contrast, for players that use a *systematic style*, designers may want to design progression to expand mental model development to other features of the AI or facilitate more accurate conclusions. For example, to facilitate more directed conclusions about the AI player, designers may consider incorporating conditions such as *draw realistically* or *draw simplistically* to help redirect conclusions about what type of data the AI uses. Or, to expand mental model development to other features of the AI, such as the AI's word bank, the game could prompt players with a reflective question before or after gameplay. For instance, asking what type of words the AI knows to help encourage systematic players' to consider a different feature of the AI in their mental model.

When it comes to *top-down focus* versus *bottom-up focus*, designers can consider what existing mental models might be triggered or what models may emerge based on how the game is initially presented to the players. Prior work has suggested that surface features of the AI agent may impact the initial construction of mental models [23]. As such, designers should consider how to represent the AI player in the game carefully. This representation may impact how bottom-up players select elements of gameplay for model construction or what existing mental model top-down players may evoke to begin construction. In the context of *iNNk*, the AI is presented as another player playing the same role as the human Guesser. This presentation could be why we observed some players attribute more capabilities to the AI, such as learning and prediction. This phenomenon has been recently explored by Hwang et al. [28] who revealed that users possess their own "baseline" mind perception toward AI entities. They found that adding human touches, such as visual or audio features, to the representation of an AI can cause some users to mistakenly treat AI-mediated agents as overly human. Therefore, game designers should consider how representations of AI players impact the focus dimension.

In addition, designers can consider what additional information is provided about the AI player and how it functions in the player experience throughout the game. These details could also encourage players to either abandon an existing model or consider a new model. For example, *iNNk* displays a confidence percentage alongside the AI's guesses (see Figure 1). Initially, some players did not notice this percentage until later matches. Not emphasizing this detail earlier could be why we observed some sessions attribute a different functionality to the AI over time, like P2-1, who changed their mental model of how the AI makes its guess after considering this detail in the final match. Designers should consider how accompanying information about the AI player influences the style dimension over time.

5.2 Extending the Framework

To increase our understanding of player mental model development, we conducted a qualitative study to examine how players make sense of an *adversarial* AI player. Here, we examined the development process as a whole and uncovered two dominant dimensions that describe how players develop their mental model (i.e., focus and style). However, mental model development is a much more complex and nuanced process [68, 76]. Therefore, more work is needed to further unpack development on a more granular level. Future work can examine these dimensions on a spectrum to further describe the nuanced changes over time and individual differences regarding developmental activities.

One approach to extending the framework is to utilize existing HCI theories on mental model construction and development [17, 68, 76]. These rich and complex perspectives may further shed light on the nuances of development. For example, Waern [68] outlines seven theoretical concepts which researchers can use to understand the dynamics of users' mental models. While her framework is intended to examine the model development of a single computer task, this series of events can be used as a starting point to evaluate the granularities of change. For instance, examining how players progress through these events and how this interrelates with our framework dimensions could extend how we describe development in games.

While our preliminary framework describes how players develop their mental models, there remains an open question regarding how these differences relate to the accuracy of players' mental models concerning the different components of the AI system. Recently, Gero et al. [21] proposed a conceptual model of an AI agent consisting of three components: global behavior, knowledge distribution, and local behavior. This conceptual model could evaluate players using different focus and development styles. Perhaps, through this lens, we may better explain how and why certain players develop more accurate mental models than others.

Finally, to fully understand the different ways players develop their mental models and the potential impact of different games on how players make sense of these systems, more work is needed to examine development in other AI-based games. For our study, we utilized an adversarial game in which humans and AI compete against each other. It is unclear if the framework derived from this game can be generalized to other AI-based games or if players develop their mental models differently in other types of games, such as games that foster a sense of cooperation with AI. We hypothesize that *iNNk's* competitive interaction pushes players to reason and strategize more about the AI's limitations, thus providing more opportunities for model development. However, other interactions may very well support more model development. Therefore, we encourage the community to investigate the impact of different AI-based games on model development.

5.3 Design Implications

For game designers, UX designers, and HCI researchers interested in mental models of AI, we propose the following implications for future work in this area.

5.3.1 Fostering a Particular Development Style. From our results, we found that development differs in terms of integrating new information (i.e., systematic vs. reactive style). However, what remains unclear is if games should foster a particular style and when they should do so. For example, a game about learning AI and its limitations may want to foster a systematic style, so the players may gradually build a more specific mental model of the AI's limitations. On the other hand, a game that aims to provide players with an increasingly competitive challenge may want to foster a reactive style so that the players integrate more diverse information about the AI to maintain excitement or engagement. Balancing these two types remains an open question for further investigation. A common approach to

balancing gameplay and player engagement is the concept of flow [12], which can be a helpful perspective for designing mental model progression in games, perhaps supporting both systematic and reactive styles to facilitate and guide development in a particular way.

5.3.2 Guiding Mental Model Development. Our study observed numerous creative ways that players made sense of the AI player and its limitations, perhaps facilitated by our research design of requesting players to consider a different strategy for the next match. As a result, players generated many AI concepts in which some sessions constructed and elaborated on more than others. We noticed that generating more AI concepts does not necessarily indicate a better mental model. For instance, Session 2 had 11 AI concepts, while Session 7 had 4 AI concepts in total. Session 2, however, attributed more capabilities to the AI, while Session 7 attributed more specifics in terms of how the AI deciphers the drawing. This raises questions on how mental model development should be guided. For example, designers could consider fostering reflection on AI players by incorporating reflective moments in their gameplay (see [44, 46]). Future research should consider what form of reflections should be provided to encourage or inhibit the construction and elaboration of AI concepts in their game.

Designers may also consider establishing appropriate bounds for players that generate more AI concepts; as observed in our study, more is not necessarily better. Designers may facilitate these bounds with an appropriate frame. Prior research shows that how an activity is *framed* strongly influences the experience (see [41]). In the context of mental model development, we suggest that the concept of *framing* should be explored for eliciting and guiding mental models, with open questions on when and where such framing needs to happen.

5.3.3 Detecting Mental Models of AI. While games are a relatively new domain for personalization [78], an open problem is properly detecting the appropriate behaviors for adaptation. Researchers have explored experience managers or another AI system that oversees how players interact with the AI player in a game [77]. However, we have yet to explore how these systems can personalize interaction based on the state of a player's mental model. Game designers or researchers may consider utilizing the framework to indicate the state and approach of a player's mental model development. For example, we noticed that players who generate more AI concepts during gameplay than others could indicate a player using a *reactive style*. However, more work is needed to validate these findings in a larger study.

Constructing such a system or adapting an NN to account for this can be challenging [43]. Game designers could also detect mental model development explicitly by asking players to describe the state of their mental model through in-game prompts, which is a common approach to facilitate reflection in learning games [65]. For example, in *iNNk*, the game could promote moments of explicit reflection after a match by asking "How do you think the AI works?" Such reflective prompts may lead to more appreciation for the game and the AI player.

5.3.4 Using Failure to Encourage Change. Failure is an important step in the process of acquiring accurate mental models as it is known to be good for reflection and learning [18, 33, 34]. It is also a fundamental element in games and is often used by designers to improve players' knowledge of the game itself [3, 18, 33] by allowing players to succeed by repeatedly failing [33]. As failure is crucial to mental model development, we noticed that players tend to change their mental model after failure. This coincides with Gero et al.'s [21] finding that people are most open to revision when an anomaly occurs. Our findings can extend how players process revisions which we observed occur through comparisons. Future work may want to explore further the role of failure on mental model development. More research on how failure is interpreted and influences mental model development may assist in describing the nuances of this process and how to best design failure to guide mental model development.

A potential starting place for future work is considering the concept of productive failure proposed by Kapur [34] who explored the benefits of failure on students' problem-solving ability by having students struggle through ill-structured problems. He found that those who struggle early on and explore the problem space more fully come to have a deeper understanding than those with more direct guidance. Perhaps, designers could consider incorporating moments of productive failure to facilitate mental model development.

5.4 Limitations

There are several limitations to this study. First, to observe how players develop their mental model, we acknowledge that parts of our protocol may have affected its development. Specifically, (1) asking players to find different ways to defeat the AI, and (2) asking players not to continue using a strategy they believed effective. This choice was to ensure groups did not stall the development of their mental models once they found a working strategy, which would have limited the observable behaviors for the researchers to analyze.

Second, we acknowledge the analysis's subjective and interpretive nature. An explicit limitation is that we infer mental models' from the data gathered; thus, we cannot guarantee this is representative of the players' actual mental model or if players share the same mental model. Although there is no agreement on the methodological approach to studying and measuring mental models [16, 27], we utilized common approaches in mental model studies. We used think aloud, interview methods, and gameplay data to capture, as best as possible, opportunities to glimpse inside players' mental model development across the entire session.

Third, we encouraged players to verbalize how they thought the AI worked, specifically in the debriefing interview, which is a form of reflection [42]. This verbalization could assist in the development of players' mental models. However, it is likely that this outward reflection may not happen outside of a study context.

Finally, this study focused on a specific type of AI system in a particular adversarial game, specifically an image recognition game. To fully understand the different ways mental models of AI develop during gameplay and the generalizability of our results, more work is needed to compare these results with other types of AI systems and genres of games and to extend this work to a larger sample size.

6 CONCLUSION

This paper examines how a pair of players develop mental models of an adversarial AI player during gameplay in the multiplayer drawing game *iNNk*. We conducted ten gameplay sessions in which two players (n = 20, 10 pairs) worked together to defeat the AI agent in this game. Our results uncovered two dominant dimensions that describe players' mental model development (i.e., focus and style). We present this as a preliminary framework for game designers and researchers interested in mental models of AI. We discuss how designers can utilize this framework to examine and design for mental model development in adversarial AI-based games. Further, we provide several design implications from our study for game designers and HCI researchers to improve current challenges in facilitating mental models of AI systems.

ACKNOWLEDGMENTS

This work is partially supported by the National Science Foundation (NSF) under Grant Numbers IIS-1917855 and IIS-1816470 as well as the Novo Nordisk Foundation Grant under the Grant Number NNF20OC0066119. The authors would like to thank Amanda Jørgensen for her assistance in collecting literature in cognitive psychology and providing valuable feedback. Finally, we thank all past and current members of the project.

REFERENCES

[1] David Ackermann, Michael J Tauber, and Michael J Tauber. 1990. *Mental models and human-computer interaction 1.* Number 3. North Holland.

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[3] Craig G Anderson, Jen Dalsen, Vishesh Kumar, Matthew Berland, and Constance Steinkuehler. 2018. Failing up: How failure in a game environment promotes learning through discourse. *Thinking Skills and Creativity* 30 (2018), 135–144.

[4] 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. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 2–11.

[5] Pascal Béguin and Pierre Rabardel. 2000. Designing for instrument-mediated activity. *Scandinavian Journal of Information Systems* 12, 1 (2000), 1.

[6] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

[7] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. In *26th International Conference on Intelligent User Interfaces*. 307–317.

[8] Gabriele Cimolino, Sam Lee, Quentin Petraroia, and TC Nicholas Graham. 2019. Oui, Chef!!: Supervised Learning for Novel Gameplay with Believable AI. In *Extended Abstracts of the Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*. 241–246.

[9] Allan Collins and Dedre Gentner. 1987. How people construct mental models. *Cultural models in language and thought* 243 (1987), 243–265.

[10] Michael Cook, Mirjam Eladhari, Andy Nealen, Mike Treanor, Eddy Boxerman, Alex Jaffe, Paul Sottosanti, and Steve Swink. 2016. PCG-Based Game Design Patterns. *arXiv preprint arXiv:1610.03138* (2016).

[11] Kenneth James Williams Craik. 1943. *The nature of explanation*. Vol. 445. CUP Archive.

[12] Mihaly Csikszentmihalyi. 1990. *Flow: The psychology of optimal experience*. Vol. 1990. Harper & Row New York.

[13] James K Doyle and David N Ford. 1998. Mental models concepts for system dynamics research. *System dynamics review: the journal of the System Dynamics Society* 14, 1 (1998), 3–29.

[14] Kate Ehrlich. 1996. Applied mental models in human–computer interaction. *Mental models in cognitive science: Essays in honour of Phil Johnson-Laird* (1996), 223.

[15] Katrin Etzrodt and Sven Engesser. 2021. Voice-based agents as personified things: Assimilation and accommodation as equilibration of doubt. *Human-Machine Communication* 2 (2021), 57–76.

[16] Creso Franco and Dominique Colinvaux. 2000. Grasping mental models. In *Developing models in science education*. Springer, 93–118.

[17] Caleb S Furlough and Douglas J Gillan. 2018. Mental models: structural differences and the role of experience. *Journal of Cognitive Engineering and Decision Making* 12, 4 (2018), 269–287.

[18] James Paul Gee. 2005. Learning by design: Good video games as learning machines. *E-learning and Digital Media* 2, 1 (2005), 5–16.

[19] Dedre Gentner. 2002. Mental models, psychology of. In *International encyclopedia of the social and behavioral sciences*. Elsevier Science, 9683–9687.

[20] Dedre Gentner and Albert L. Stevens. 1983. *Mental models*. Lawrence Erlbaum Associates.

[21] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R Millen, Murray Campbell, et al. 2020. Mental models of AI agents in a cooperative game setting. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[22] Google. 2016. Quick, Draw! https://quickdraw.withgoogle.com/

[23] John Graham, Liya Zheng, and Cleotilde Gonzalez. 2006. A Cognitive Approach to Game Usability and Design: Mental Model Development in Novice Real-Time Strategy Gamers. *CyberPsychology & Behavior* 9, 3 (2006), 361–366.

[24] Robert C Gray, Jichen Zhu, and Santiago Ontañón. 2021. Multiplayer Modeling via Multi-Armed Bandits. In *2021 IEEE Conference on Games (CoG)*. IEEE, 01–08.

[25] David Ha and Douglas Eck. 2017. A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477* (2017).

[26] Clara E Hill, Sarah Knox, Barbara J Thompson, Elizabeth Nutt Williams, Shirley A Hess, and Nicholas Ladany. 2005. Consensual qualitative research: An update. *Journal of counseling psychology* 52, 2 (2005), 196.

[27] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[28] Angel Hsing-Chi Hwang and Andrea Stevenson Won. 2022. AI in Your Mind: Counterbalancing Perceived Agency and Experience in Human-AI Interaction. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–10.

[29] Daniel Jallov, Sebastian Risi, and Julian Togelius. 2016. EvoCommander: A novel game based on evolving and switching between artificial brains. *IEEE Transactions on Computational Intelligence and AI in Games* 9, 2 (2016), 181–191.

[30] Bernard Jansen, Joni Salminen, Soon-gyo Jung, and Kathleen Guan. 2021. Data-driven personas. *Synthesis Lectures on Human-Centered Informatics* 14, 1 (2021), i–317.

[31] Philip Nicholas Johnson-Laird. 1983. *Mental models: Towards a cognitive science of language, inference, and consciousness*. Number 6. Harvard University Press.

[32] Natalie A Jones, Helen Ross, Timothy Lynam, Pascal Perez, and Anne Leitch. 2011. Mental models: an interdisciplinary synthesis of theory and methods. *Ecology and Society* 16, 1 (2011).

[33] Jesper Juul. 2013. *The art of failure: An essay on the pain of playing video games.* MIT press.

[34] Manu Kapur and Nikol Rummel. 2012. Productive failure in learning from generation and invention activities. *Instructional Science* 40, 4 (2012), 645–650.

[35] Cecilia Katzeff. 1988. The effect of different conceptual models upon reasoning in a database query writing task. *International Journal of Man-Machine Studies* 29, 1 (1988), 37–62.

[36] Cecilia Katzeff. 1990. System demands on mental models for a fulltext database. *International Journal of Man-Machine Studies* 32, 5 (1990), 483–509.

[37] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more? The effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1–10.

[38] Todd Kulesza, Simone Stumpf, Margaret Burnett, Weng-Keen Wong, Yann Riche, Travis Moore, Ian Oberst, Amber Shinsel, and Kevin McIntosh. 2010. Explanatory debugging: Supporting end-user debugging of machine-learned programs. In *2010 IEEE Symposium on Visual Languages and Human-Centric Computing*. IEEE, 41–48.

[39] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on visual languages and human centric computing*. IEEE, 3–10.

[40] Franco Landriscina. 2013. *Simulation and learning*. Springer.

[41] Andreas Lieberoth. 2015. Shallow gamification: Testing psychological effects of framing an activity as a game. *Games and Culture* 10, 3 (2015), 229–248.

[42] Xiaodong Lin, Cindy Hmelo, Charles K Kinzer, and Teresa J Secules. 1999. Designing technology to support reflection. *Educational technology research and Development* 47, 3 (1999), 43–62.

[43] Mathias Löwe, Jennifer Villareale, Evan Freed, Aleksanteri Sladek, Jichen Zhu, and Sebastian Risi. 2021. Dealing with Adversarial Player Strategies in the Neural Network Game iNNk through Ensemble Learning. In *The 16th International Conference on the Foundations of Digital Games (FDG) 2021*. 1–10.

[44] Tim Marsh. 2016. Slow serious games, interactions and play: Designing for positive and serious experience and reflection. *Entertainment computing* 14 (2016), 45–53.

[45] Richard E Mayer, Amanda Mathias, and Karen Wetzell. 2002. Fostering understanding of multimedia messages through pre-training: Evidence for a two-stage theory of mental model construction. *Journal of Experimental Psychology: Applied* 8, 3 (2002), 147.

[46] Elisa D Mekler, Ioanna Iacovides, and Julia Ayumi Bopp. 2018. " A Game that Makes You Question..." Exploring the Role of Reflection for the Player Experience. In *Proceedings of the 2018 annual symposium on computer-human interaction in play*. 315–327.

[47] Marvin Minsky. 1974. A framework for representing knowledge.

[48] Avantika Mohapatra. 2020. Designing for AI: A collaborative framework to bridge the gap between designers and data scientists, and enabling designers to create human-centered AI products and services.

[49] Donald A Norman. 1983. Some observations on mental models. *Mental models* 7, 112 (1983), 7–14.

[50] Donald A Norman and Daniel G Bobrow. 1979. Descriptions: An intermediate stage in memory retrieval. *Cognitive Psychology* 11, 1 (1979), 107–123.

[51] Donald A Norman and David E Rumelhart. 1978. Accretion, tuning and restructuring: Three modes of learning. *Semantic Factors in Cognition. Hillsdale, NJ: Lawrence Erlbaum* (1978), 37–53.

[52] Stephen J Payne. 2003. Users' mental models: The very ideas. *HCI models, theories, and frameworks: Toward a multidisciplinary science* (2003), 135–156.

[53] Sean Penney, Jonathan Dodge, Claudia Hilderbrand, Andrew Anderson, Logan Simpson, and Margaret Burnett. 2018. Toward foraging for understanding of StarCraft agents: An empirical study. In *23rd International Conference on Intelligent User Interfaces*. 225–237.

[54] Jan Piskur, Peter Greve, Julian Togelius, and Sebastian Risi. 2015. Braincrafter: An investigation into human-based neural network engineering. In *2015 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 2199–2206.

[55] K Andrew R Richards and Michael A Hemphill. 2018. A practical guide to collaborative qualitative data analysis. *Journal of Teaching in Physical Education* 37, 2 (2018), 225–231.

[56] Sebastian Risi, Joel Lehman, David B D'Ambrosio, Ryan Hall, and Kenneth O Stanley. 2015. Petalz: Search-based procedural content generation for the casual gamer. *IEEE Transactions on Computational Intelligence and AI in Games* 8, 3 (2015), 244–255.

[57] Yvonne Rogers, Helen Sharp, and Jenny Preece. 2011. *Interaction design: beyond human-computer interaction*. John Wiley & Sons.

[58] D Rumelhart and E Schemata. 1980. The building blocks of cognition. *Theoretical issues in reading comprehension: Perspectives from cognitive psychology, linguistics, artificial intelligence, and education* (1980).

[59] David E Rumelhart. 1984. Schemata and the cognitive system. (1984).

[60] Pamela Ann Savage-Knepshield. 2001. *Mental models: Issues in construction, congruency, and cognition*. Rutgers The State University of New Jersey-New Brunswick.

[61] Roger C Schank and Robert P Abelson. 2013. *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology Press.

[62] Joe Tullio, Anind K Dey, Jason Chalecki, and James Fogarty. 2007. How it works: a field study of non-technical users interacting with an intelligent system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 31–40.

[63] Josep Valls-Vargas, Jichen Zhu, and Santiago Ontañón. 2017. Graph grammar-based controllable generation of puzzles for a learning game about parallel programming. In *Proceedings of the 12th International Conference on the Foundations of Digital Games*. 1–10.

[64] Jennifer Villareale, Ana V Acosta-Ruiz, Samuel Adam Arcaro, Thomas Fox, Evan Freed, Robert C Gray, Mathias Löwe, Panote Nuchprayoon, Aleksanteri Sladek, Rush Weigelt, et al. 2020. iNNk: A Multi-Player Game to Deceive a Neural Network. In *Extended Abstracts of the 2020 Annual Symposium on Computer-Human Interaction in Play*. 33–37.

[65] Jennifer Villareale, Colan F. Biemer, Magy Seif El-Nasr, and Jichen Zhu. 2020. Reflection in Game-Based Learning: A Survey of Programming Games. In *International Conference on the Foundations of Digital Games*. 1–9.

[66] Barry J Wadsworth. 1984. *Piaget's theory of cognitive and affective development* (third edition ed.). Longman Publishing.

[67] Yvonne Waern. 1985. Learning computerized tasks as related to prior task knowledge. *International Journal of Man-Machine Studies* 22, 4 (1985), 441–455.

[68] Yvonne Waern. 1990. On the dynamics of mental models. In *Selected papers of the 6th Interdisciplinary Workshop on Informatics and Psychology: Mental Models and Human-Computer Interaction 1*. 73–93.

[69] Michael D. Williams, James D. Hollan, and Albert L. Stevens. 1983. Human Reasoning About a Simple Physical System. In *International encyclopedia of the social and behavioral sciences*. Lawrence Erlbaum Associates, 131–153.

[70] John R Wilson and Andrew Rutherford. 1989. Mental models: Theory and application in human factors. *Human Factors* 31, 6 (1989), 617–634.

[71] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–13.

[72] Georgios N Yannakakis and Julian Togelius. 2018. *Artificial intelligence and games*. Vol. 2. Springer.

[73] Nick Yee. 2006. Motivations for play in online games. *CyberPsychology & behavior* 9, 6 (2006), 772–775.

[74] Nick Yee. 2015. *Game Motivation Model*. https://quanticfoundry.com/

[75] Yan Zhang. 2009. The construction of mental models of information-rich web spaces: The development process and the impact of task complexity. (2009).

[76] Yan Zhang. 2013. The development of users' mental models of MedlinePlus in information searching. *Library & information science research* 35, 2 (2013), 159–170.

[77] Jichen Zhu and Santiago Ontañón. 2019. Experience management in multi-player games. In *2019 IEEE Conference on Games (CoG)*. IEEE, 1–6.

[78] Jichen Zhu and Santiago Ontañón. 2020. Player-centered AI for automatic game personalization: Open problems. In *International Conference on the Foundations of Digital Games*. 1–8.

[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 ACM Conference on Human Factors in Computing Systems (CHI '21)*.

Received February 2022; revised June 2022; accepted July 2022