Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)
Sister Conferences Best Papers Track

# Mental Models of AI Agents in a Cooperative Game Setting (Extended Abstract)*

**Katy Ilonka Gero[1], Zahra Ashktorab[2], Casey Dugan[2], Qian Pan[2], James Johnson[2], Werner Geyer[2], Maria Ruiz[3], Sarah Miller[3], David R Millen[3], Murray Campbell[2], Sadhana Kumaravel[2], Wei Zhang[2]**
[1]Columbia University
[2]IBM Research AI
[3]IBM Watson
katy@cs.columbia.edu, {zahra.ashktorab1, qian.pan, sadhana.kumaravel1, maria.ruiz}@ibm.com,
{cadugan, jmjohnson, werner.geyer, millers, david_r_millen, mcam, zhangwei}@us.ibm.com

## Abstract

As more and more forms of AI become prevalent, it becomes increasingly important to understand how people develop mental models of these systems. In this work we study people's mental models of an AI agent in a cooperative word guessing game. We run a study in which people play the game with an AI agent while "thinking out loud"; through thematic analysis we identify features of the mental models developed by participants. In a large-scale study we have participants play the game with the AI agent online and use a post-game survey to probe their mental model. We find that those who win more often have better estimates of the AI agent's abilities. We present three components—global knowledge, local knowledge, and knowledge distribution—for modeling AI systems and propose that understanding the underlying technology is insufficient for developing appropriate conceptual models—analysis of behavior is also necessary.

## 1 Introduction and Related Work

When we sit down to drive a car, or look for a file on our computer, we use a mental model to make sense of the world and act on it. Mental models are developed quickly and unconsciously by users. In contrast, conceptual models are developed slowly and purposefully by experts. Discrepancies between the two can lead to problems, ranging from misunderstanding and confusion to the abandonment of a system.

Through studies of human error and human-machine interaction, Norman [2014] observes that mental models are incomplete, limited, unstable, unscientific, parsimonious, and lack firm boundaries—they value utility over accuracy. Greca and Moreira [2000], considering mental models in the context of science education, find that instruction on a conceptual model does not lead to students' acquiring perfect copies of it, and that modification of initial mental models is difficult, suggesting we enrich existing models rather than overhaul them.

*Originally published as "Mental models of AI agents in a cooperative game setting." in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*.

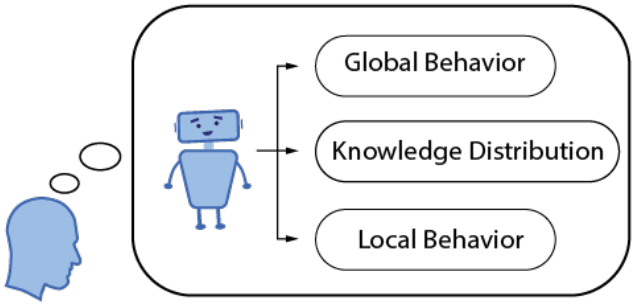


Figure 1: A mental model of an AI agent has three components: behavior at a large scale, the agent's knowledge of various topics, and behavior at the scale of an individual output.

As AI systems appear in high-stakes environments, such as decisions about who to hire [Dickson and Nusair, 2010] or diagnosing diseases [Cai *et al.*, 2019], understanding people's mental models of these systems becomes increasingly important. Additionally, the label 'AI system' may be applied to a variety of technologies, from linear regression-based predictions to neural network-generated images, complicating our ability to learn about them. While some HCI researchers have looked into how people develop mental models of AI systems [Kulesza *et al.*, 2012; Kulesza *et al.*, 2013; Bansal *et al.*, 2019], mostly we have seen research into Explainable AI [Cheng *et al.*, 2019; Wang *et al.*, 2019; Wiegand *et al.*, 2019; Kunkel *et al.*, 2019]. But a rich understanding of the underlying technology does not always lead to a rich understanding of how a system will behave. For now, many AI systems remain idiosyncratic in their behavior.

Many important questions remain open. In the context of a cooperative word guessing game, we pose the following research questions:

1. What should conceptual models of AI systems include?
2. How do users develop mental models of AI systems?
3. What encourages *accurate* mental models of AI systems?

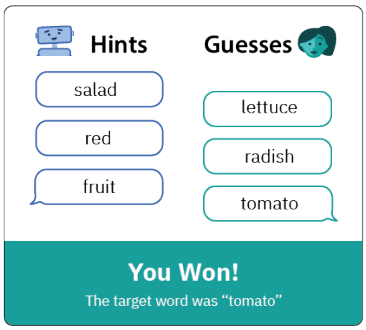


Figure 2: Example round of the game Passcode, hints provided by the AI agent and guesses provided by the participant.

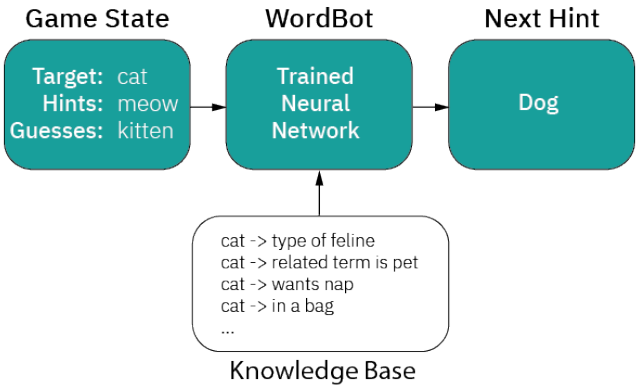


Figure 3: Diagram of the 'giver' AI agent, called WordBot. WordBot is a trained neural network, which has encoded information from the training data. In addition, information from a knowledge base is used as input along with the game state.

## 2 System Design

In this work we focus on cooperative word games, which require understanding what your partner is thinking. Studying mental models in this context has a long history in linguistics [Wittgenstein, 2009] and more recently has gained popularity in AI research [Rovatsos *et al.*, 2018; Bard *et al.*, 2019]. In particular, we use a game called 'Passcode'. In this game one player tries to guess a word that the other player is thinking of; the other player provides one word hints. The game itself is grounded in trying to understand what the other player is thinking, making it an excellent test bed for studying mental models. Figure 2 shows a typical round of gameplay.

We use two reinforcement learning-based AI agents trained to play Passcode–one to play the giver (who has a target word and gives hints) and one to play the guesser (who is trying to guess the target word based on the hints). Each AI agent has a neural network architecture, is pre-trained with word association data [Nelson *et al.*, 2004], has access to a commonsense knowledge graph [Speer *et al.*, 2017], and is trained further in a reinforcement learning framework. As with many AI systems, these AI agents perform quite well at the game, but are not perfect.

We consider what a conceptual model (i.e. an accurate mental model) of the AI agent would look like. A precise description of the neural network architecture and training procedure does not always represent a system's actual behavior, which may differ from its intended behavior.

For the rest of the paper we will focus only on the AI agent for the giver, which we call 'WordBot'. This is an important simplification, as the two AI agents (giver and guesser) have slightly different actual and intended behaviors, given the different roles they play. Figure 3 shows a diagram of the AI agent for the giver.

We take a systematic approach to characterizing the behavior of the agent. For example, we cannot assume that because WordBot has access to a knowledge base, it effectively uses all that information to generate meaningful hints. In the commonsense knowledge graph, there is rich information about Paris—that Paris is the capital of France, that the Eiffel Tower and the Louvre are located there, that it has cafes and boulevards. Yet the hints that WordBot provides for the target word 'paris' are 'city', 'usa', 'plant'—WordBot appears to have very poor knowledge of Paris.

We deduce a conceptual model of WordBot from a combination of understanding its structure and training procedure, *and* a series of analyses of actual results from playing with WordBot. We note that the terminology we use below was developed iteratively and informed by the results of Study 1. We present the following conceptual model of WordBot:

### 2.1 Global Behavior

- WordBot does not remember or adjust its hints based on past rounds.
- WordBot rarely adjusts its hints based on incorrect guesses within a single round.
- WordBot has no explicit hint sequencing strategy.

### 2.2 Knowledge Distribution

- WordBot does not know anything about pop culture (as this is not in the training data).
- WordBot has limited knowledge about geography/places (35% of hints are bad).
- WordBot has decent knowledge about food/cooking (11% of hints are bad).

### 2.3 Local Behavior

- WordBot gives both synonym (29% of the time) and antonym (11% of the time) hints.
- WordBot gives one or more hints that are not highly related to the target word in 4% of games.
- WordBot takes into account multiple senses of a word (if a word has multiple senses).

We note that our use of the term "local behavior" is related to the team "local explanations" as used in the explainable AI literature [Mittelstadt *et al.*, 2019]. The "local behavior" portion of a system model identifies how individual decisions or actions made by a system; "local explanations" seek to explain these individual decisions or actions.

# 3 Study 1: Small-Scale, Think-Aloud Study

## 3.1 Methodology

In this study, we brought 11 participants (recruited from IBM) into the lab either as individuals or as teams of two to play Passcode with WordBot while thinking out loud about their strategy and the strategy of WordBot. Not all participants worked on technology development (for example some worked in operations). The average age was 22.4 ($\pm$ 2.8) years, and 64% of participants had some exposure to coding. All participants had read about 'artificial intelligence' in the news. All participants, either as an individual or on a team, played 5 games as the giver and 5 games as the guesser, the order counter-balanced. The AI agent, WordBot, assumed the other role and participants interacted with WordBot through a simple command-line version of the game. Participants were given a maximum of 10 guesses per game; if they had not won the game within 10 guesses, they moved on to the next round. All participants played the game using the same target words in the same order. These words were randomly selected from the vocabulary of the AI agent and had a range of difficulties. We conducted a thematic analysis [Braun and Clarke, 2012] on the resulting transcripts of what the participants said while playing as well as their responses to a post-play semi-structured interview. This study gave us insight into what were the important aspects of a conceptual model, the kinds of mental models players develop, and how players come to their beliefs about the system.

## 3.2 Results

Table 1 describes the 10 codes developed through the thematic analysis, ordered by their prevalence in the transcripts. All utterances related to the research questions were marked with a code. Not all codes correspond to expressions of a participant's mental model; instead, many correspond to moments when a participant's mental model is used or challenged. Broadly, participants remarked upon what the AI agent knows and how the AI agent plays the game. They either said statements about these things, or gave questions or expressions of uncertainty about these things. These results guided our development of a 'mental model' survey, used in Study 2 to probe participants' mental models of WordBot.

The most prevalent code (18% of all utterances) was **anomalies/distress/trust**. These responses included simple acknowledgement of an unexpected move, distress in which the participant believed they were stupid for not understanding the unexpected move, and concerns about not trusting that the AI agent was making good or meaningful moves. There were several understandably confusing moves from the AI agent, as well as moves that in retrospect made sense (e.g. antonym hints). Some participants were slow to fault the AI agent even when reviewing a game in which some hints were clearly not helpful; instead they would interpret and justify strange moves. Others immediately blamed the AI agent, and were slow to acknowledge that they may have misunderstood how the AI agent was relating words. These moments of confusion forced participants to judge the AI agent in order to progress, and often resulted in a participant changing their mental model when the target word was revealed.

# 4 Study 2: Large-Scale, Online Gameplay

## 4.1 Methodology

To better understand what impacts people's mental models, we ran a large-scale, online study using Amazon Mechanical Turk. For this study we had participants only play as the guesser (the AI agent played as the giver). Participants were allowed a maximum of 5 guesses. We looked at three factors which could influence people's mental models:

- The number of game rounds played
- The target words played (i.e. difficulty, theme, etc.)
- The win rate of the player

Participants played either 5 or 10 game rounds, where each round consisted of trying to guess a single target word, and played one of two wordlists (i.e. the target words to guess). Participants playing only 5 game rounds played on the first five words in the list. The two wordlists were balanced for difficulty, as well as topic—for instance, each word list had the same number of food-related words. Participants saw their words in a random order.

We could not control for how often a participant won or lost, but in analysis split participants up into the top 50% of players ('winners' – those who won the same or more than the median amount) and the bottom 50% ('losers' – those who less than the median amount). The game was developed into an online web application using Flask (a lightweight Python framework for web apps) and React (a Javascript library for building front-end interfaces).[1] Participants first took a short demographic survey, then played 5 or 10 game rounds, and then took a survey that asked questions about how they thought the AI agent worked.[2]

## 4.2 Results

The study resulted in 89 Amazon Mechanical Turk workers participating in the study in 'good faith'.[3] Despite a significant portion of non-native English speakers (17%), we saw no difference in win rate between native English speakers and not. Similarly we saw no difference in win rate for age or education level. We had three questions that asked about participants' familiarity with word games, machine learning, and coding. These were not predictors of win rate. Additionally, there were no significant differences between any survey answers for the number of games played.

We did see significant differences between the 'winners' and 'losers'. Table 2 shows mean survey responses and significance levels. Let's consider the two global behavior questions. Losers tend to believe (more than winners) that WordBot takes into consideration your past incorrect guesses, as well as previous game plays. Both of these are untrue. Winners tend to be unsure, or suspect WordBot does not take into consideration these things. Here it is clear that winners have a better understanding of the global behavior of WordBot than losers; losers tend to *overestimate* WordBot's abilities.

[1] A demo can be found at ibm.biz/wordbot.

[2] Development of the survey was based on Study 1.

[3] All guesses were inspected manually, and any participant who clearly had not put in a good faith effort, for instance always guessing the word 'word' regardless of the hints, were removed.

| Code | Prev | Description and Example Quote |
|---|---|---|
| anomalies/distress/trust | 18% | Noting unexpected behavior from the AI agent, or expressing distress or mistrust in response to unexpected behavior.<br>P6: *Wait so we have 'chill' and 'hectic'. I'm confused.* |
| pattern seeking | 16% | Discussing or questioning specific patterns (within a single game) the AI agent uses to give hints/guesses.<br>P9: *It would make me feel bad if there was a pattern that we were totally missing.* |
| synonyms/antonyms | 15% | Any discussion of synonyms or antonyms as it related to the type or efficacy of hints.<br>P2: *...the fact that it could give antonyms because I thought it would only do synonyms.* |
| AI knowledge | 14% | Discussion of what the AI agent does or does not know, or questioning the same.<br>P2: *I mean it smells but I don't think the AI would know that nail polish smells.* |
| memory/weighting | 12% | Discussion of how much the AI agent remembers, or how much 'weight' is given to subsequent hints/guesses.<br>P4: *I guess I need to look at all four of these equally.* |
| steering | 10% | Noting the need to "steer" the AI agent (or be steered by the AI agent) toward the target word, or questioning how to best get the AI agent "back on track".<br>P10: *How to get them back on track when they start going off...* |
| need for explanation | 7% | Expression of desire for explanation for a single hint/guess or generally for how the AI agent made decisions.<br>P7: *Can I know what the AI is? That would be very useful for me.* |
| reflection | 5% | Explicit reflection on past game plays to inform the next move.<br>P9: *Uhhh I feel like this is another 'minute' situation. This feels familiar.* |
| personification | 3% | Questioning or hesitation about how to describe the AI agent, or explicit discussion of the AI agent as one would a human.<br>P8: *Maybe a different unit of time would lead them – it – down a better path.* |
| perspective taking | 2% | Explicit discussion of the perspective of the AI agent.<br>P8: *...no one would say 'give' to help us guess 'marriage'.* P9: *Maybe a bot would.* |

Table 1: Name, prevalence, description, and example quote of the ten codes found through the thematic analysis of the think-aloud transcripts. Prevalence is calculated as the number of utterances marked with a particular code divided by the total number of utterances marked with a code; there were exactly 100 utterances so it also represents the utterance counts.

| | Mean | | |
|---|---|---|---|
| Question (shortened) | winner | loser | p-value |
| GLOBAL BEHAVIOR | | | |
| **adjusts hints based on guesses** | **3.9** | **4.6** | **.02** |
| **remembers past gameplays** | **3.6** | **4.4** | **.01** |
| KNOWLEDGE DISTRIBUTION | | | |
| knows about pop culture | 3.7 | 4.3 | .16 |
| knows about geography/places | 4.2 | 4.8 | .09 |
| knows about food/cooking | 4.4 | 4.8 | .26 |
| LOCAL BEHAVIOR | | | |
| many synonym hints | 5.0 | 5.1 | .62 |
| **many antonym hints** | **3.5** | **4.6** | **.01** |

Table 2: Results from post-gameplay survey, split by winner/loser. Significant differences bolded. We see that losers over estimate global behavior, and some local behavior. We don't see any differences in knowledge distribution, perhaps because there was not enough exposure to the system.

## 5 Conclusion

We studied conceptual and mental models of AI systems in the context of a word guessing game, Passcode. We developed a conceptual model of an AI agent that plays Passcode, finding three key components of conceptual models for AI systems more generally: global behavior, knowledge distribution, and local behavior. We probed user mental models in two studies. The first was an analysis of a think-aloud study (n=11) in which participants played Passcode with an AI agent, illustrating the themes that arise when trying to understand an AI technology. The second was an online study (n=89) in which participants played Passcode with an AI agent and filled out a survey about their mental model, showing that playing more games did not increase the accuracy of a mental model, but that participants who won more often did have more accurate models. Overall we found that people have existing intuitions about how AI systems work which can upset their understanding of a specific AI agent, and that people tend to revise their mental model in the face of anomalies.

## References

[Bansal *et al.*, 2019] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter Lasecki S, Daniel S Weld, and Eric Horvitz. Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 2019.

[Bard *et al.*, 2019] Nolan Bard, Jakob N Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, et al. The hanabi challenge: A new frontier for ai research. *arXiv preprint arXiv:1902.00506*, 2019.

[Braun and Clarke, 2012] Virginia Braun and Victoria Clarke. Thematic analysis. In *APA handbook of research methods in psychology, Vol. 2.*, pages 57–71. American Psychological Association, 2012.

[Cai *et al.*, 2019] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 4. ACM, 2019.

[Cheng *et al.*, 2019] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 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*, CHI '19, pages 559:1–559:12, New York, NY, USA, 2019. ACM.

[Dickson and Nusair, 2010] Duncan R Dickson and Khaldoon Nusair. An hr perspective: the global hunt for talent in the digital age. *Worldwide Hospitality and Tourism Themes*, 2(1):86–93, 2010.

[Greca and Moreira, 2000] Ileana Maria Greca and Marco Antonio Moreira. Mental models, conceptual models, and modelling. *International journal of science education*, 22(1):1–11, 2000.

[Kulesza *et al.*, 2012] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 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*, CHI '12, pages 1–10, New York, NY, USA, 2012. ACM.

[Kulesza *et al.*, 2013] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 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*, pages 3–10. IEEE, 2013.

[Kunkel *et al.*, 2019] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 487:1–487:12, New York, NY, USA, 2019. ACM.

[Mittelstadt *et al.*, 2019] Brent Mittelstadt, Chris Russell, and Sandra Wachter. Explaining explanations in ai. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 279–288. ACM, 2019.

[Nelson *et al.*, 2004] Douglas L Nelson, Cathy L McEvoy, and Thomas A Schreiber. The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407, 2004.

[Norman, 2014] Donald A Norman. Some observations on mental models. In *Mental models*, pages 15–22. Psychology Press, 2014.

[Rovatsos *et al.*, 2018] Michael Rovatsos, Dagmar Gromann, and Gábor Bella. The taboo challenge competition. *AI Magazine*, 39(1):84–87, 2018.

[Speer *et al.*, 2017] Robert Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[Wang *et al.*, 2019] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 601:1–601:15, New York, NY, USA, 2019. ACM.

[Wiegand *et al.*, 2019] Gesa Wiegand, Matthias Schmidmaier, Thomas Weber, Yuanting Liu, and Heinrich Hussmann. I drive - you trust: Explaining driving behavior of autonomous cars. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI EA '19, pages LBW0163:1–LBW0163:6, New York, NY, USA, 2019. ACM.

[Wittgenstein, 2009] Ludwig Wittgenstein. *Philosophical investigations*. John Wiley & Sons, 2009.