Mental Models of AI Agents in a Cooperative Game Setting

Katy Ilonka Gero*, Zahra Ashktorab†, Casey Dugan†, Qian Pan†, James Johnson†, Werner Geyer†, Maria Ruiz‡, Sarah Miller‡, David R Millen‡, Murray Campbell§, Sadhana Kumaravel§, Wei Zhang§
Columbia University*, IBM Research AI†§, IBM Watson‡
NYC, NY, USA*, Cambridge, MA, USA†‡, Yorktown, NY, USA§
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 AI in a cooperative word guessing game. We run think-aloud studies in which people play the game with an AI agent; 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 for modeling AI systems, propose that understanding the underlying technology is insufficient for developing appropriate conceptual models (analysis of behavior is also necessary), and suggest future work for studying the revision of mental models over time.

Author Keywords
Artificial intelligence; mental models; conceptual models; games; word games; AI agents; think-aloud.

CCS Concepts
•**Human-centered computing** → *Empirical studies in HCI; HCI theory, concepts and models;* •**Computing methodologies** → *Artificial intelligence;*

INTRODUCTION
Mental models define how we interact with the world. When we sit down to drive a car, or explain how lights work to a child, or look for a file on our computer, we use our mental models to make sense of the world and act on it. A mental model, which is an individual's understanding of how a system works or behaves, is often distinguished from a conceptual model, which is an expert or designer's understanding of the system. Conceptual models are developed slowly and purposefully by people with extensive knowledge of the system. Mental models, in contrast, are developed quickly and often unconsciously by people who know far less about the system but still desire to use it. Discrepancies between the two can lead to a host of problems, ranging from misunderstanding and confusion to the abandonment of a system altogether.

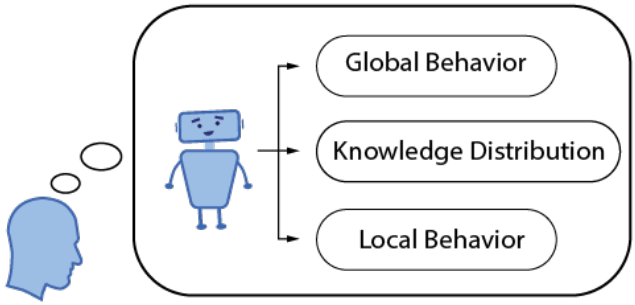

Figure 1: A mental model of an AI agent has three components: behavior at a large scale, the agent's knowledge of various topics, and behavior at the scale of an individual output.

As AI systems appear in high-stakes environments, such as decisions about who to hire [10] or diagnosing diseases [6], 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. This has spurred HCI researchers into Explainable AI [8, 25, 26, 15], but part of the difficulty of this research is a lack of conceptual models of AI systems. A rich understanding of the underlying AI technology does not always lead to a rich understanding of how an AI system will behave. The fact is, for now, many AI systems remain somewhat idiosyncratic in their behavior.

Yet these AI systems are being used, and we are developing mental models that guide our use of and reasoning about these technologies. Many important questions remain open. In the

CHI '20, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6708-0/20/04 ...$15.00.
http://dx.doi.org/10.1145/3313831.3376316

context of a cooperative word guessing game, we pose the following research questions:

RQ1 What should conceptual models of AI systems include?

RQ2 How do users develop mental models of AI systems?

RQ3 What encourages *accurate* mental models of AI systems?

AI systems are used in a wide range of situations and no one use case is a perfect representative. 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 [27] and more recently has gained popularity in AI research [21, 4].

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.

We ran two studies to investigate what appropriate conceptual models of AI systems look like and how users develop mental models of AI systems. The first is an in-person, think-aloud study, in which participants play Passcode with an AI agent while thinking out loud. This study allowed us to identify the important aspects of a mental model and get a qualitative understanding of how people think about AI systems.

The second is a large-scale online study, in which participants played 5 or 10 rounds of the game and then filled out a survey which probed their mental model of the AI agent. This study showed us who makes accurate estimations of the AI agent, and points us towards why these people do so.

This paper makes the following contributions:

- An example conceptual model of an AI agent consisting of three key components (global behavior, knowledge distribution, and local behavior) based on the technological structure and training procedure, as well as an analysis of actual behavior.
- A thematic analysis of a think-aloud study (n=11) in which participants played Passcode with an AI agent, illustrating the comments and concerns that arise when trying to understand an AI technology.
- 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 the mental model, but that participants who won more often did have more accurate models.

RELATED WORK

Mental Model Theory

We draw heavily on an existing foundation of research when creating our own framework for conceptual and mental models and probing how mental models change over time. In the field of design, Norman [20] considers four distinct things in the consideration of mental models: the *target system* which is the actual system a person uses; the *conceptual model* of the target system which is "invented to provide an appropriate representation of the target system" and tend to be developed methodically by experts; the *mental model* of the target system which is evolved by users through interaction with the target system; and the *scientist's conceptualization* of the mental model, which is a model of a model. Through studies of human error and human-machine interaction, Norman observes that mental models are incomplete, limited, unstable, unscientific, parsimonious, and lack firm boundaries. Norman finds that mental models value utility over accuracy.

Greca and Moreira [11], considering mental models in the context of science education, further discuss how instruction on a conceptual model does not lead to students' acquiring perfect copies of it. They also observe that physicists use distinct mental models when engaging with different phenomena, though they use a shared conceptual model when presenting their results. In the educational context, they note that modification of initial mental models is difficult, and suggest enriching existing models rather than overhauling them.

Mental Models of AI Systems

Various work in HCI has tackled how people model AI systems, though few study the deep neural network-based systems which are becoming increasingly popular. Kulesza et. al [13, 14] study mental models of an intelligent music recommender system; they quantify people's mental models with a survey and find that a 15 minute tutorial (with an experimenter) significantly increased the soundness of participants' mental models, as did high fidelity written explanations. Bansal et. al [2] look at the effect of different kinds of AI errors on people's mental models, using performance as an indicator of a mental model.

Moreso than users' explicit mental models, research on AI systems in HCI has focused on explainability and trust. Rutjes et. al [22] argue for capturing a user's mental model and using it while generating explanations. Miller [17], in a comprehensive review of social science related to explainable AI, invokes the concepts of mental models through ideas of reconciling contradictions and our desire to create a shared meaning. Yin et. al [29] look at the effect of stated and observed accuracy of machine learning models on people's trust of the system, finding that the effect of stated accuracy can change depending on the observed accuracy. Relatedly, Bansal et. al [3] look at the effect of updates to AI technology in human-AI teams, finding that updates that increase AI system performance can hurt overall team performance. We believe work on explainability and trust would benefit from independent studies on mental models, which is what we do in this paper.

Models and Inference in Language Games

Our work was heavily inspired by the study of mental models in the context of language games, which has a long history in linguistics [27] as well as AI research. In the context of AI research, understanding the mental models of others is a key element of communication that AI must acquire. The 'Taboo' word guessing game was put out as an AI challenge [21], as the game "forces agents to speculate about their partner's understanding of the domain, rather than just performing inference on their own knowledge"; a similar AI challenge was suggested for the game 'Hanabi' [4]. Xu and Kemp [28] and

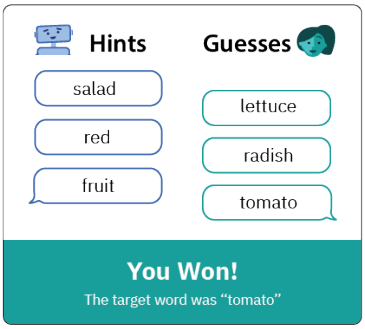

Figure 2: Example round of the game Passcode, hints provided by the AI agent and guesses provided by the participant.

Shen et. al [23] study logs of people playing word guessing games to understand how people tend to provide and receive hints. In tandem there is work in which AI agents are developed to play these games [9, 12, 1], both to test theories about communication as well as to contribute to larger AI goals.

SYSTEM DESIGN

Passcode: A Cooperative Word Guessing Game

To learn about the mental models of AI agents that people develop, we use a simple cooperative word guessing game which we call Passcode. It's a two person game, in which one person has a target word in mind and gives one-word hints to get their partner to guess the target word. The player who gives hints is referred to as the 'giver', and the player who guesses is referred to as the 'guesser'. The game starts with the giver giving a hint like 'toast'; after each hint the guesser must make a guess, in this case maybe 'bread'. If the guess is correct, the game is over and the players win. In our version, the game is web-based, such that single-word hints and guesses are typed and displayed to both players[1]. Figure 2 shows a typical round (i.e. guessing of one target word) of gameplay.

Passcode is cooperative, meaning both players are on the same team and a win for one player is a win for the other. The benefit of this type of game is that there is no reason, for any player, to hide or disguise any information about themself or their strategy. The structure of the game introduces opaqueness; the goal is overcome that opaqueness. The game is explicitly about trying to understand what your partner is thinking, which makes it an ideal environment to study mental models.

AI Agent Description

We use a reinforcement learning-based AI agent trained to play Passcode, developed by a team of AI researchers. In fact we use two AI agents – 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. As with many AI systems, these AI agents perform quite well at the game, but

[1]Passcode is similar to some other word games, like the television game 'Password', and the card games 'Taboo' and 'Codenames'.

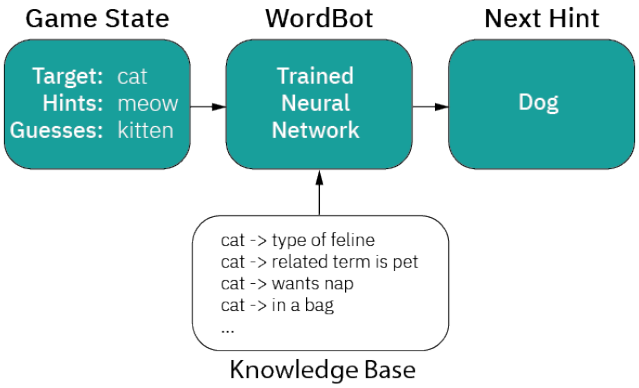

Figure 3: Diagram of the 'giver' AI agent, called WordBot. WordBot is a trained neural network, which has encoded information from the training data. In addition, information from a knowledge base is used as input along with the game state.

are not perfect. Below, we describe the high-level technical details of the system. However, the technical details are not the main contribution of this paper and for that reason further details are omitted because the system is in submission in another venue.[2]

The AI agents are pre-trained using two datasets of word associations [19][3]. They also always have access to the ConceptNet knowledge base [24], which encodes common sense knowledge. After a period of pre-training, the giver and guesser agents are trained in a reinforcement learning framework by playing with each other. The architecture of the giver agent is such that it considers the most recent (if incorrect) guess of its partner when selecting a new hint. The architecture of the guesser is such that it considers all hints equally, regardless of the order in which they were given. In use (i.e. outside of training) neither agent retains any information about gameplay from game to game. For instance, given the same hints, the guesser agent will always guess the same word, regardless of past rounds of gameplay.

We consider what a conceptual model (i.e. an appropriate mental model) of the AI agent would look like. It is our belief that a precise description of the neural network architecture and training procedure does not represent an appropriate conceptual model for players. A conceptual model should reflect its actual behavior, which might differ from its intended behavior.

For the rest of the paper we will focus only on the AI agent for the giver. 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.

[2]The development of the AI agents was done independently of the work presented in this paper.

[3]The second dataset is a collection of taboo cards from the website http://playtaboo.com.

AI Agent Characterization

Let us consider the behavior of the AI agent for the giver, which we called 'WordBot'. 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 ConceptNet knowledge base, 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.

To characterize WordBot, we run the following analyses. These include analysis of gameplay mechanics as well as knowledge categories. In considering WordBot's knowledge, we selected categories for which the system had access to knowledge in ConceptNet, yet did not generate meaningful hints from many words in that category.

Adjusting hints. How often does WordBot change its hint given the previous guess? We analyze 600 games across 20 random target words (with 5 hints per game). For each target word, we generate 30 unique guess sequences using the related terms for the target word from ConceptNet. We measure how many of the 30 guess sequences resulted in hint sequences that differed from the most common sequence. Hint sequences differ from most common sequence 11.8% of the time. The majority of differences are in the last hint, so this number drops dramatically when you consider not all players see all 5 hints, e.g. it is 4.5% when considering the first 3 hints.

Knowledge about food/cooking. We look at 20 food or cooking words: 5 cooking verbs, 5 cooking objects, 5 raw foods, and 5 processed foods.[4] Compared to a baseline of 20 random words, we look at the number of related terms in ConceptNet (the knowledge base available to WordBot), and the number of bad hints generated by WordBot (its actual behavior). (The number of related terms for a given target word is the number of English terms in the AI agent's vocabulary found in ConceptNet, querying the 'related to' relation for the target word. The number of bad hints is determined by the authors of the paper; bad hints are those either totally unrelated, e.g. 'plant' for 'paris', or too vague, e.g. 'thing' for 'pot'.) Results are in Table 1, where we compare these measures to a baseline set of 20 random words. We see that WordBot both has less access to information about food/cooking (mean 15.6 relations/word compared to baseline of 23.8), and produces bad hints more often (11.5% of the time to baseline of 4.0%).

Knowledge about geography/places. We look at 20 geography or place words: 5 well-known cities, 5 well-known countries, and 10 geological landforms.[5] Similar to above, results are in Table 1; we see that WordBot both has about as much access to information about geography/places as the baseline (mean 21.3 relations/word compared to baseline of 23.8), and *yet* produces bad hints much more often (36.4% of the time to baseline of 4.0%). This means that although WordBot has *access* to information about geography/places, its behavior indicates it is *not* particularly knowledgeable about this category.

Target Word Category	Knowledge Base Data (avg # relations)	WordBot Behavior (% bad hints)
random	23.8	4.0%
food/cooking	15.6	11.5%
place/geography	21.3	36.4%

Table 1: Comparison of information in the knowledge base and the quality of WordBot's behavior. Each target word category contains 20 words. Note that having a high number of Knowledge Base relations *does not* predict WordBot behavior.

Synonym versus antonym hints. Based on an analysis of 20 random target words, we mark each hint (in the most common hint sequence) as a synonym, antonym, or other relation to the target word. Other includes many kinds of relations, such as adjectival, common collocation, or similar type (e.g. 'pot' is a similar type to 'pan'). We find that synonym hints occur 29% of the time, and antonym hints occur 11% of the time.

Thus, we deduce a conceptual model of WordBot from a combination of understanding its structure and training procedure, *and* an analysis 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:

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.

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.
- WordBot has decent knowledge about food/cooking.

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 [18]. 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.

[4]The list is: fry, roast, chop, cook, boil, knife, oven, spoon, pot, pan, apple, pork, broccoli, rice, egg, bread, yogurt, soup, cake, dumpling.
[5]The list is: london, paris, tokyo, rome, beijing, usa, france, china, india, germany, beach, coast, valley, mountain, desert, ocean, field, rainforest, iceberg.

STUDY 1: IN-PERSON THINK ALOUD

In this study, we brought participants 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. We conducted a thematic analysis [5] on the resulting data. This study gave us insight into the important aspects of a conceptual model, the kinds of mental models players develop, and how players come to their beliefs about the system. Additionally, this study guided our development of a 'mental model' survey, used in Study 2 to probe participants' mental models of WordBot.

Experimental Design

11 participants were recruited from a local technology company (IBM), though not all participants worked on technology development (for example some worked in operations). The average age was 22.4 ($\pm$ 2.8) years. The gender ratio was 55% male/45% female. 64% of participants had some exposure to coding. All participants had read about 'artificial intelligence' in the news.

5 of the participants played the game individually and 6 played in teams of 2 (in which they together played one role), resulting in 8 observational studies. We included team-play to encourage out-loud thinking (participants could talk to each other) and in-depth reflection on the AI agent (participants often negotiated or collectively brainstormed their next move).

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 (guesser vs giver) 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 lost the game and 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. Difficulty is measured by the accessibility index, a measure from [19] that reports how often a word is thought of when prompted with other words. For example, 'dog' has a high accessibility index, whereas 'trombone' has a low one. It is related to, though not identical with, frequency of usage.

When participants played as the giver of hints, with the AI agent guessing, they had the target words *minute*, *run*, *polish*, *genius*, *life*. When they played as the guesser, with the AI agent providing hints, they had the target words *vase*, *calm*, *forgive*, *plant*, *fly*.

Participants were given instructions for how to play the game, and instructed to think out loud as they played the game. They were told the study would be audio-recorded. If a participant was being particularly quiet, they were prompted to talk through their thought process.

After playing all 10 games, participants were asked a series of questions in a semi-structured interview designed to elicit information about how the participants thought the AI agent worked, and what they would need to do to better understand it. Participants were then debriefed and allowed to ask questions.

Data Analysis

To conduct the thematic analysis, three of the authors transcribed all of the audio recordings. Then two authors read all of the transcriptions at least twice, on the second reading taking notes of pertinent utterances and themes. Together they developed a series of codes and example utterances for analyzing the data, given our research questions. They then coded all transcripts individually, marking relevant utterances with the appropriate code. Finally, the two authors reviewed all the transcripts together to discuss any disagreements, and formed an agreement on the final coding of the data [16].

Results

Overview

Table 2 describes the 10 codes developed through the thematic analysis, ordered by their prevalence in the transcript. (Prevalence is calculated as the number of utterances marked with a particular code divided by the total number of utterances marked with any 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. Utterances marked with a code were either explicit or implicit statements about these things, or questions or expressions of uncertainty about these things.

Discussion of Prevalent Codes

The most prevalent code was **anomalies/distress/trust**, which included all utterances in response to an unexpected move by the AI agent. 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, for instance giving the hint 'blood' for the target word 'plant', as well as moves that in retrospect made sense though in the moment confused players, like giving the antonym hint 'hectic' for the target word 'chill'.

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 it. 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.

As discussed in the System Design section, the AI agent has no explicit strategy for how to select hints or guesses. The next hint or guess could correspond to a wide range of different semantic relations, from synonymy to adjectival to common collocations. The second most prevalent code was **pattern seeking** in which participants actively sought out an explicit strategy for the AI agent. Interestingly this differs from how

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. 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. 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. 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. 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. 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". 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. 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. 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. 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. P8: *...no one would say 'give' to help us guess 'marriage'.* P9: *Maybe a bot would.*

Table 2: 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.

people typically play the game with other people, based on pilot testing and the authors' own experiences with word guessing games. In games between people, participants tend to understand that there is no one best strategy for selecting hints, and that a given teammate is likely to change strategies based on the situation. Yet, when study playing with an AI agent, people longed to understand and explain all the moves, perhaps because their teammate was a foreign agent they could not directly relate to.

Synonyms/antonyms, AI knowledge, and memory/weighting, the next most prevalent codes, all refer to explicit theories of how the AI agent worked or what it knew. Sometimes participants stated what they thought they knew about the AI agent. At other times participants questioned how the AI agent worked, or what the AI agent knew, using specific examples.

Synonyms/antonyms refers to discussion of whether or not the AI agent used antonyms as hints, or to the AI agent's preference for synonyms. Almost all participants were at first confused by the use of antonyms as hints, and later came to understand it was a commonly used strategy.

AI knowledge was often a discussion of what a participant thought the AI agent would *not* know, resulting in the participant looking for a different hint. Generally participants suspected the AI agent didn't know much about pop culture, for instance that it probably didn't know who 'Jimmy Neutron' (a cartoon character) was, or that 'Sorry' is a pop song by Justin Beiber. However, they also wondered about proper nouns (such as if the AI agent knew that 'polish' was a verb but also an ethnicity) and detailed knowledge (would it know that 'nail polish' smelled bad).

Finally, **memory/weighting** refers to a discussion of what the AI remembers, within a single game or across the many games participants played. Several participants assumed the AI agent remembered past plays because the internet (broadly assumed to have AI involved) seemed to have knowledge of their past usage as demonstrated by targeted advertisements. Others wondered if perhaps it remembered little about their past plays because digital assistants like Siri and Alexa seemed to *not* remember anything from past conversations. The remaining codes were less prevalent, and can be reviewed in Table 2.

Model Components and Survey Development
We uncover three components for modeling an AI agent. These components were developed iteratively through discussions by the entire team of researchers after the thematic analysis of Study 1 was conducted. These components are a framework for describing a conceptual model or a mental model; for example in the System Design section we use them to articulate the conceptual model of WordBot.

The components are: **Knowledge Distribution** which includes conceptions such as whether or not the AI agent knows about specific people or attributes, **Local Behavior** which includes conceptions of what kinds of hints the AI agent is likely to give or respond best to, and **Global Behavior** which includes conceptions of how the AI agent tends to play the game, such as what and how much the AI agent remembers from previous interactions.

Using these results, we developed a set of Likert-scale survey statements, several per aspect, to determine how players thought the AI agent worked.

It's important to consider how a participant may know or learn a correct answer. For instance, let's consider the statement "WordBot knows a lot about pop culture". Given all information about the system it's possible to know that WordBot knows nothing about pop culture – it is not available in any of the training data, and there is no way for it learn it in a reinforcement learning context. However, a participant who has only played the game does not know all this. All they can possibly know is that WordBot never makes any references to pop culture, which suggests (but may not confirm) WordBot knows nothing about pop culture. This is characteristic of mental models – users develop an understanding of a system based on their exposure to and use of that system.

STUDY 2: LARGE-SCALE, ONLINE GAMEPLAY

Experimental Design
To better understand what impacts people's mental models, we ran a large-scale, online study using Amazon Mechanical Turk. For this study we limit participants to only playing as the guesser, and the AI agent, WordBot, playing as the giver.

Participants were allowed a maximum of 5 guesses. If they did not correctly guess the target word after these 5 guesses, they lost the game and were moved to the next game round.[6]

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 consists 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.

- List A ('london'): london, egg, dog, clean, cold, beef, mountain, clothes, help, music
- List B ('china'): china, tomato, hard, friend, time, coast, potato, hair, small, happy

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).[7]

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.[8] In pilot studies, the average time for completion for the 5 game condition was 11 minutes and for the 10 game condition was 13 minutes. Based on this, all participants were paid $3 for the task, or about $15/hour.

Results
113 Amazon Mechanical Turk workers in total participated in the study. Participants could only complete the study once. Any participants who did not complete the correct number of games or the post-game survey were removed. All guesses by the remaining participants were inspected manually, and any participant who clearly had not put in a good faith effort, for instance always guessing whatever hint was given or always guessing the word 'word' regardless of the hints, were removed. This resulted in 89 'good faith' participants.

In addition to the two controlled conditions – the number of required games and the word lists – we classified all participants as 'winners' or 'losers' based on the average number of games a given participant correctly guessed the target word. For instance, if a participant guessed 9 out of 10 target words correctly, they had a win rate of 0.9. 'Winners' were those participants who won the same or more than the median win rate, which was 0.6. 'Losers' were those participants who won less than the median rate. Table 3 gives exact breakdowns of the participants and the conditions.

Demographics and Previous Experience
Participants were asked their age given several range buckets. 53% of our participants were 26-35 years old, with the rest spread somewhat evenly across the buckets 18-25, 36-45, and 45+. All participants had at least a high school diploma, while

[6] In the think-aloud study we allowed 10 guesses, and many participants became agitated toward the end of a game; several requested a 'give up' option. To keep people invested, reduce irritation, and maximize the use of people's time, we decided to limit the maximum length of the game.

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

[8] Development of the survey is discussed at the end of Study 1 Results, and the survey questions in full can be found in the supplementary materials.

	all	winners	losers
wordlist A	42	30	12
5 games	20	14	6
10 games	22	16	6
wordlist B	49	34	13
5 games	28	23	5
10 games	19	11	8
all players	89	64	25

Table 3: Breakdown of the number of participants in the online study. 'Winners' won more than or the same as the median rate, which was 0.6; 'losers' won less than the median rate. Note the uneven split despite splitting on the median – many players won at the median rate, resulting in an uneven split when these players are placed in the winner category.

Question (shortened)	Mean 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 4: 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.

58% also had a bachelor's degree and 3% had an advanced degree. 83% of the participants reported English as their native language; despite a list of 10 other languages, the remaining 17% selected 'other', and many of these reported (using a text field) Malayalam as their native language.

Despite a significant portion of non-native English speakers, 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. As above, these were not predictors of win rate.

Number of Games Played

There were no significant differences between any survey answers for the number of games played. We had thought that playing more games would give participants more time and evidence with which to develop a more accurate mental model. This does not appear to be the case. It could be that the difference between 5 and 10 games is too small to make a difference. A promising direction would be to look at the difference between one session of 10 games, and multiple sessions of 10 games, in which many games (perhaps 100 games) are played over several days. Given this, for the rest of the analysis we group 5 and 10 game players together.

Word List

We tried to balance the word lists for difficulty and exposure to particular concepts. While individual words vary in difficulty, overall the word lists were quite balanced – for word list A ('london') the average win rate was 0.63, while for word list B ('china') the average win rate was 0.65 (p-value is 0.88; no significant difference).

There was only one question in the survey which had a significant difference between the word lists. Participants with word list A ('london') reported that WordBot used antonyms hints more than word list B ('china') did (mean response from wordlist A: 4.4, mean response from word list B: 3.2, difference: 1.2, p-value of 0.00). A close examination of the hints suggests this is true: looking at all hints provided wordlist A ('london') had 20% antonym hints, whereas wordlist B ('china') had 0% antonym hints.

This result indicates that participants were certainly responding to the games they played, and were not simply relying on past or existing mental models of AI systems.

Win Rate

For every participant, we calculated their win rate: the fraction of games they played in which they correctly guessed the target word and won. The median win rate was 0.6. We split participants into those who won as much or more than the median win rate, and those who won less than the median win rate. Our theory was that people who win the game more are likely to have more accurate mental models; perhaps they win more *because* they understand the system better.

We did see significant differences between these two groups. Table 4 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.

In the knowledge distribution questions, we see no significant differences. It's possible that to understand knowledge distribution, more exposure to the system is necessary.

Let's examine the local behavior questions. Based on the analysis of the AI agent, we know that synonym hints occur 29% of the time and antonym hints occur 11% of the time. There was no significant difference for the question about synonym rate; both groups overestimate the rate of synonym occurrence. We do see a significant difference for the question about antonym rate: we find that winners, while still overestimating, are significantly closer to the true rate than losers.

We might think that winners are simply *trying harder* than losers, both in playing the game and in answering the survey questions. They might be more reflective, take longer to play the game, and therefore fill out the survey with more correct answers – not because they have better mental models, but because they try harder to use and explain them. However, it turns out this is not the case. The average time for a winner to make a guess is 15.6 seconds; losers take 20.5 seconds (p-value is 0.3; not significant).

To better understand this result we asked 10 participants to explain their answers to the survey questions about global behavior and knowledge distribution.

Winners tended to have good recall for the gameplay, and actively reflected on the gameplay to answer the questions. For instance P2, who had a win rate of 0.8, gave this answer in response to "WordBot knows a lot about food/cooking":

> P2 (0.8) One word was tomato and hints were salad and red so the bot is good with food association.

In contrast, P5, who had a win rate of 0.4, acknowledged that they did not have evidence for their answer but assumed it would know about food associations anyway:

> P5 (0.4) I never saw these words being used here but I expect it to know some of these words.

Winners seem to be more likely to say "I'm not sure" than make a guess – losers are more likely to go with their intuition. Here is a winner, P3 who had a win rate of 0.8, reasoning about global behavior, and if WordBot remembers the past gameplay rounds:

> P3 (0.8) I am not sure if this is the case, all the rounds seemed independent to me.

It could be that a more accurate mental model enables better gameplay, perhaps through a virtuous cycle in which better gameplay encourages more engagement, which in turn further improves gameplay. An alternative explanation is that there is a characteristic of winners that makes them both better at playing the game, *and* better at understanding the AI agent, such as better verbal reasoning skills. This experiment alone cannot distinguish between the two.

DISCUSSION

Our analysis of the AI agent itself and people's interactions with it during the think-aloud resulted in three principal components of a conceptual model of an AI system: global behavior, knowledge distribution, and local behavior. We also developed a set of intuitions about how mental models are formed. We discuss how this work can be generalized to other kinds of AI systems and how our results on the development and accuracy of mental models intersects with research on explainable AI and AI trust. Finally, we address limitations of the study, and unpack how we might better probe, understand, and influence people's mental models in the future.

What Conceptual Models Include

Conceptual models are models of a system developed over time by experts. But AI systems seem somehow different than a calculator or web site because there is often no one-to-one mapping from the design of the system (for instance, the neural architecture) and the behavior of the system.

To address **RQ1**, what should conceptual models of AI systems include, we analyzed both the underlying technology, which includes the neural architecture, the training procedure, and the data used to train it, as well as its actual behavior. The analysis of its behavior was guided by our own experience interacting with the AI system, as well as purposeful observations of others interacting with it. These three ingredients – the underlying technology, observation of interaction, and analysis of behavior – seem necessary to develop an appropriate conceptual model.

Through this we developed three categories or types of features of the AI system: *global behavior*, *knowledge distribution*, and *local behavior*. In the case of our AI agent, which played a word guessing game, these categories have very specific interpretations that have to do gameplay tactics and strategies, and the type of common sense knowledge important for the game. But we believe these categories can guide conceptual model development of all kinds of AI systems.

For example, consider an AI system that makes health recommendations based on test results. In this case, global behavior would refer to whether or not recommendations change over time, what these changes are based on, and if or how the system learns from new examples or feedback. Knowledge distribution would refer to if the system makes equally accurate recommendations over all kinds of test results or types of patients, what kinds of other information it has access to, and whether it uses this knowledge appropriately. Local behavior would refer to why specific recommendations are made, what kinds of recommendations the system can make, and what the most common recommendations are. For example, in Bansal et. al [2] they look at people's mental models of error boundaries–their formulation would fit into local knowledge, i.e. understanding the details of individual decisions.

This framework could be used to design tools that help users understand complex AI systems. Consider again a healthcare recommendation system. Global behavior understanding could be supported by an onboarding activity that specifies how recommendations do or don't change over time. Knowledge distribution understanding could be supported by contextualizing results with information on how well the system performs on different types of input. Local behavior understanding could be supported by generated explanations, or breaking recommendations down into components.

Revising Mental Models in the Face of Anomalies

It is hard to track the development of mental models, but in our think-aloud study we gained some insights into **RQ2**, how do users develop mental models of AI systems. The most common utterances in the think-aloud study had to do with anomalies, distress, and trust – people talked most about their mental model when something unexpected occurred. This is also where we saw the most revision; despite trying to explain an anomaly, when an anomaly persisted people did end up revising their model. Antonym hints showed this clearly:

most people were initially distressed by antonym hints that seemed to contradict the other hints presented, and some even thought that these hints were mistakes. However, after the game concluded they acknowledged that the behavior made sense, revising their mental model.

Miller [17], in his review of insights from social sciences for explainable AI, states two situations in which people desire explanation: 1) when a contradiction occurs, and 2) when shared meaning is desired. This dovetails nicely with our finding that people tend to revise their mental models in the face of anomalies. Considering how to design explanations for AI systems, our results confirm Miller's finding that we should provide explanations to people when anomalies occur, as this is when they are most open to revision and most desire an explanation.

Outside Knowledge, Guessing, and Overestimation

In the Study 2, we studied the accuracy of people's mental models. Thinking of **RQ3**, what encourages accurate mental models of AI systems, we mostly found evidence for the roots of *inaccurate* mental models. Here we relate those findings to techniques which might work against inaccurate models.

People tended to overestimate the AI system's abilities, particularly those who lost the game the most. Sometimes people guessed in the face of not enough information, like P5 from the large-scale studying: "I never saw these words being used here but I expect it to know some of these words." But we also saw people drawing explicitly on past experiences to explain their current experience. For example, P10 from the think-aloud drew on their understanding of online targeted advertising to interpret WordBot's abilities. In this case, they relied on outside knowledge to understand their experience in the game, which resulted in overestimation.

This may be tied to the need for explanation. The second most prevalent utterances from the think-aloud had to do with pattern seeking – people had a strong desire to understand the *patterning* of the hints when they didn't know what to guess next. (We might consider this state, of not knowing what to guess, an anomaly; people may expect, in general, to know what to do next.) A surprisingly common occurrence was the participants trying to put all the given hints into a sentence or a scene, and using that sentence/scene to discover a related word to use as their guess. We expect the world to make sense in specific, often narrative, ways. When an AI systems fails to fit this expectation, it may lead to overestimation.

To encourage accurate mental models, we may have to be pro-active in the face of existing parallels (does an AI system behave differently to other systems) and the absence of signals (does an AI system *not* do something, especially something that might be expected).

Limitations

There are several limitations to this study. Most importantly, there may be differences between the context of a game and the context of, say, using a prediction to make a decision. Although we saw differences between winners – who can be thought of as high performers – and losers, this may or may not transfer to more standard explainable AI scenarios, where a user may be asked to explain how inputs affect outputs. However, despite the different context, our study does seem to follow the Bansal et. al human-AI interaction framework [3], in which 1) the environment provides an input (the target word), 2) the AI suggests an action (the hint), 3) the user makes a decisions based on this (the guess), and 4) the environment produces a reward based on the user decision (win or not).

Our study looks only at short-term exposure, and we found that the difference between 5 and 10 games was too small to detect. While many people are navigating AI systems for the first time, it's also important to consider long-term usage.

Finally, we looked at only one AI agent. Ideally, we could compare AI agents with different features, to fully understand how exposure affects mental models. If one AI agent does adapt to each player, and another does not, will high performing players be sensitive to this? Some work has been done on updating models [3] and comparing game performance between AI models [7]; it would be worthwhile understanding people's sensitivities to model differences.

Future Directions

The question of how winning relates to accurate mental models is important, and dovetails nicely with what kind of interventions could improve people's mental models. Although our test-bed is games, it is in a way a prediction task, and winners are better at predicting what the AI agent knows. How do we confirm the relationship between winning and accurate mental models? And can we improve people's ability for both?

One could use our framework for testing interventions. Participants could play a set number of games to get a baseline win rate, experience an intervention such as an explanation or example, and then play another set of games. This intervention could occur either at a fixed time, after a participant makes a successful prediction, or after a participant makes an incorrect prediction. The difference between the win rate in the first and second set of games can show if the intervention was successful and if the timing of the intervention matters, and surveys could measure if the intervention impacts mental models. Such an experiment may work best in long-term use cases, so we could also confirm how mental models are developed given longer exposure to a system.

CONCLUSION

We studied conceptual and mental models of AI systems in the context of a word guessing game. We developed a conceptual model of an AI agent that plays the word guessing game, 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: an in-person think-aloud study and a large-scale online study. We found that people have existing intuitions about how AI systems work that can upset their understanding of this specific AI agent, that they can revise their mental model in the face of anomalies, and that those who win the game more often have better estimates of the AI agent.

REFERENCES

[1] Kemo Adrian, Aysenur Bilgin, Paul Van Eecke, and others. 2016. A Semantic Distance based Architecture for a Guesser Agent in ESSENCE's Location Taboo Challenge. *DIVERSITY@ ECAI* (2016), 33–39.

[2] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter Lasecki S, Daniel S Weld, and Eric Horvitz. 2019a. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*.

[3] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. 2019b. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2429–2437.

[4] Nolan Bard, Jakob N Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, and others. 2019. The Hanabi Challenge: A New Frontier for AI Research. *arXiv preprint arXiv:1902.00506* (2019).

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

[6] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, and others. 2019. 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*. ACM, 4.

[7] Prithvijit Chattopadhyay, Deshraj Yadav, Viraj Prabhu, Arjun Chandrasekaran, Abhishek Das, Stefan Lee, Dhruv Batra, and Devi Parikh. 2017. Evaluating visual conversational agents via cooperative human-ai games. In *Fifth AAAI Conference on Human Computation and Crowdsourcing*.

[8] 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 (CHI '19)*. ACM, New York, NY, USA, Article 559, 12 pages. DOI:http://dx.doi.org/10.1145/3290605.3300789

[9] Abhishek Das, Satwik Kottur, José MF Moura, Stefan Lee, and Dhruv Batra. 2017. Learning cooperative visual dialog agents with deep reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision*. 2951–2960.

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

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

[12] Serhii Havrylov and Ivan Titov. 2017. Emergence of Language with Multi-agent Games: Learning to Communicate with Sequences of Symbols. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 2149–2159.

[13] 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 (CHI '12)*. ACM, New York, NY, USA, 1–10. DOI: http://dx.doi.org/10.1145/2207676.2207678

[14] 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.

[15] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. 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)*. ACM, New York, NY, USA, Article 487, 12 pages. DOI: http://dx.doi.org/10.1145/3290605.3300717

[16] Nora McDonald, Sarita Schoenbeck, and Andrea Forte. 2019. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. (2019).

[17] Tim Miller. 2018. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* (2018).

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

[19] Douglas L Nelson, Cathy L McEvoy, and Thomas A Schreiber. 2004. The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers* 36, 3 (2004), 402–407.

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

[21] Michael Rovatsos, Dagmar Gromann, and Gábor Bella. 2018. The Taboo Challenge Competition. *AI Magazine* 39, 1 (2018), 84–87.

[22] Heleen Rutjes, Martijn Willemsen, and Wijnand IJsselsteijn. 2019. Considerations on Explainable AI and Users' Mental Models. In *CHI 2019 Workshop on Bridging the Gap Between AI and HCI*.

[23] Judy Hanwen Shen, Matthias Hofer, Bjarke Felbo, and Roger Levy. 2018. Comparing Models of Associative Meaning: An Empirical Investigation of Reference in Simple Language Games. *arXiv preprint arXiv:1810.03717* (2018).

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

[25] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 601, 15 pages. DOI:http://dx.doi.org/10.1145/3290605.3300831

[26] Gesa Wiegand, Matthias Schmidmaier, Thomas Weber, Yuanting Liu, and Heinrich Hussmann. 2019. 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)*. ACM, New York, NY, USA, Article LBW0163, 6 pages. DOI:http://dx.doi.org/10.1145/3290607.3312817

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

[28] Yang Xu and Charles Kemp. 2010. Inference and communication in the game of Password. In *Advances in neural information processing systems*. 2514–2522.

[29] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 279.