Mental Models of AI Agents in a Semantic Analysis Setting

Derrek Chow
University of Waterloo
Waterloo, Canada
d28chow@uwaterloo.ca

Wenbo Han
University of Waterloo
Waterloo, Canada
w27han@uwaterloo.ca

ABSTRACT

Many of today's AI systems do not provide justification of their responses, yet are becoming more involved in our lives. We investigate how users form mental models of AI systems and the implications of AI systems not providing justification for their responses.

We create an AI agent and a description of its conceptual model. We run a remote think-aloud study where participants interact with our AI agent followed by an interview. We conduct a thematic analysis on our data then discuss prevalent codes and observe patterns of behaviour. We find that people with prior AI knowledge rely on it heavily and establish mental models early, which tend to not change in the face of surprises, in favor of their prior knowledge. We discuss the problematic aspects of these findings and how they relate to AI systems. We propose that AI systems should provide justification for their responses.

Author Keywords

Artificial Intelligence; conceptual model; mental model; semantic analysis; AI agent; think-aloud

CCS Concepts

•**Human-centered computing** → *Empirical studies in HCI;*
•**Computing Methodologies** → *Artificial Intelligence;*

INTRODUCTION

AI systems are playing a more significant, often unnoticed, role in our day to day lives. AI is spreading: from behind the wheel of a car, a voice on a phone, to an answer from an app [9].

Yet it is common for these systems to operate as black boxes from the perspective of its end users. This means the system's inner workings are concealed, viewable only from their inputs and outputs. This concealment is reflective of the complexity and nature of AI systems whose understanding can even evade its makers [2, 4, 6]. Regardless of the cause, we pose that this current standard for AI systems can lead to problematic outcomes.

Today's self-driving car, voice assistant, or recommendation engine offers no explanation as to the results they produce. With insight into these systems being hidden, users are left with limited information to form their mental models of how these systems work.

A mental model is an explanation that an individual forms of how a system works. This differs from its conceptual model which is created by experts of the system. We set out to investigate the implications of the mental models of AI systems that do not provide justification of their responses with the following research questions:

RQ1 How do users develop mental models of AI systems?

RQ2 What are the implications of AI systems not providing justification of their responses to users?

We use the context of linguistics, specifically semantic analysis, for our study. We created an AI agent called WordBot for users to construct a mental model of, and to represent AI systems.

We ran a study where 6 participants interact with WordBot in a remote think-aloud study followed by an interview. The purpose was to gain insights into participant's thought processes during their interaction with the agent. We conducted a thematic analysis on the study data, compared users' mental models with our agent's conceptual model, and discussed our findings.

CONTEXT

Our study utilizes the experimental protocol, linguistic context, AI framework, name "WordBot", and findings of Gero et al. [3] where participants played an AI agent in a game setting. We seek to test and extend their findings with their original research question (RQ2) and our own (RQ1) with a modified setting and participant group.

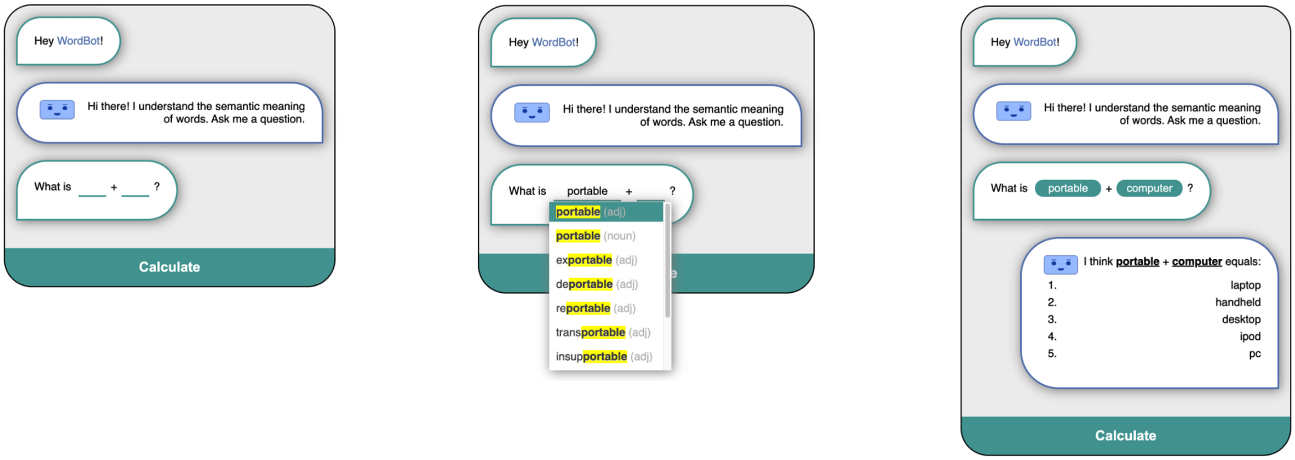
In Gero et al.'s study "All participants had read about 'artificial intelligence' in the news" [3]. In contrast, all our participants had experience with AI systems and a background in Computer Science. We made note to observe how participants having prior AI knowledge could influence how they interacted with the AI agent.

We chose the semantic meaning of words as our interaction because it has a broad interpretation and encourages the user to guess what the agent is thinking. We assume the results from a semantic analysis setting can be comparable to a game setting, yet different enough to provide their own insights and discussion.

SYSTEM DESIGN

WordBot: A Semantic Calculator

To explore the mental models of AI agents that people build, we created an AI agent in the form of a semantic calculator called WordBot. The agent is provided with 2 input words by the user (eg: *cat* and *dog*) in the form of a question (eg: "What



Hey WordBot!

Hi there! I understand the semantic meaning of words. Ask me a question.

What is ___ + ___ ?

Hey WordBot!

Hi there! I understand the semantic meaning of words. Ask me a question.

What is portable + ?

Hey WordBot!

Hi there! I understand the semantic meaning of words. Ask me a question.

What is portable + computer ?

I think portable + computer equals:

1. laptop
2. handheld
3. desktop
4. ipod
5. pc

Figure 1. WordBot's user interface

is *cat* + *dog*?"). The AI agent evaluates the semantic meaning of input words and responds with its answer, which is its top 5 guesses as to the combination of the 2 input words (eg: "I think *cat* + *dog* is: 1) *pet*, 2) *canine*, 3) *puppy*, 4) *raccoon*, 5) *mutt*").

WordBot provides an auto-complete function that prompts a list of words with the common prefix for users to select as they type to ensure valid input. The user interface of WordBot is illustrated in Figure 1.

AI Agent Description

We integrated a deep learning-based AI system into our web application developed by NLPL [5, 1]. The system analyzes words by feeding them into a neural network model with the goal of capturing their semantic meaning. It utilizes the Word2Vec method for word embeddings and is trained on the English Gigaword Fifth Edition archive [8], which is a comprehensive archive of newswire text data collected from several years. Word2Vec converts these words into 300-dimension vectors where words with closed semantic meanings have similar vectorized representations. These words form the basis of WordBot's vocabulary and are its set of valid inputs.

We clean and validate WordBot's vocabulary by taking the intersection of another word embedding model called GloVe [7], and our dataset. We also remove numbers and proper nouns from the set to make the output results more general. WordBot then performs addition on vectors that represent the input words to form a new sum vector. The five vectors with the closest Euclidean distances to this sum vector are returned – they represent the 5 output words. Figure 1 demonstrates the working process of WordBot.

AI Agent Characterization

> "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." (Gero et al., 2020)

Going beyond the underlying technology of WordBot, we analyze the actual behaviour from the AI agent to develop an appropriate conceptual model. We took a systematic approach to characterize the responses of the WordBot.

WordBot's vocabulary of accepted inputs is approximately 62,000 words and their vectorized results, which involve around 6,800 verbs, 33,000 nouns, 2,600 adverbs, and 19,000 adjectives.

It would not be reflective of the conceptual model to test a handful of categories due to the extensive and wide knowledge base. Instead, we aim to create a conceptual model of WordBot by exploring the relationship of its inputs and output. We select 120 pairs of varied input words and analyze 600 output results (2:5 input to output word ratio). These results contain the following aspects that inform our conceptual model of WordBot:

Synonyms

We look at 30 pairs of input words where each pair has similar semantic meanings to each other. 90% of the outputs are synonyms to the inputs, and 6% are antonyms to the inputs. 3% are the comparative or the superlative of the input words. Therefore, WordBot tends to produce synonyms as outputs when the input words are synonyms.

Antonyms

We look at 30 pairs of input words where each pair consists of: a word A, and a word with an opposite semantic meaning to A. 53% of the output words are either synonyms or antonyms of one of the input words. Therefore, when given semantically "opposing" inputs, WordBot tends to produce synonyms of one of the input words (or antonyms of the other) as outputs.

Words in the same category

We look at 30 pairs of input words where each pair of words share the same category. 95% of the output words are in the same category with their input words. The remaining output words are still related to those same categories. Therefore,

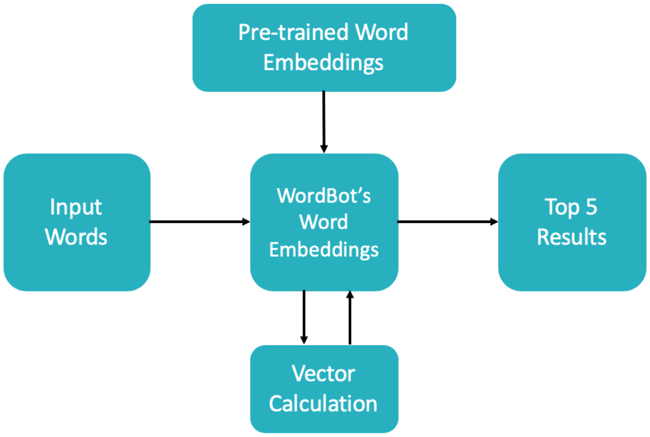


Figure 2. WordBot's working process

when given words in the same category, WordBot tends to output words in that same category.

Compound words

We look at 30 pairs of input words to see if they are combined literally in their output as a compound word (eg: "basket" + "ball" = "basketball"). Only 2 pairs of inputs produced compound words in the output. Therefore, WordBot performs the word combination in a semantic rather than literal way.

From the characteristics above, we created a conceptual model of WordBot using the 3 component framework from Gero et al. [3]:

Global Behavior

- WordBot does not remember or adjust its outputs based on past inputs.

Knowledge Distribution

- WordBot has an extensive knowledge base on words except numbers and proper nouns since such words are excluded from the dataset.
- WordBot considers each word with only one meaning, a word with different speeches are considered as different words.

Local Behavior

- WordBot's inputs are symmetric; the order of inputs has no influence on output results.
- WordBot does not literally combine input words; rather, it returns the semantic combination of words.
- WordBot often produces synonyms of the input words if the inputs are synonyms of each other.
- WordBot rarely produces antonyms of the input words if the inputs are synonyms of each other.
- WordBot often returns words in the same category as the input words.

STUDY: REMOTE THINK ALOUD

In this study, we conducted a video call with participants and had them individually interact with WordBot. We conducted a thematic analysis on the participant's responses during the study and their input to WordBot.

Experimental Design

We recruited 6 participants for the study, they were all students enrolled in the Spring 2020 course of *CS 889: Experimental Methods in HCI* at the University of Waterloo. The gender distribution was 4 male and 2 female. All participants were pursuing their Masters or PhD in Computer Science or a related field at Waterloo and had worked with AI systems before.

Participants interacted with WordBot individually in loosely guided observational studies; they typed and selected input words as shown in Figure 1. Participants were encouraged to think out loud and test WordBot's capabilities as well as try to identify its strengths, weaknesses, and domain of knowledge.

The experiment consisted of a short introduction explaining the study followed by a demo of WordBot. Participants would then interact with WordBot for 15 minutes in the observational study. After, they would be asked a series of questions in a 10 minute semi-structured interview to uncover the mental model they constructed of WordBot.

The experiment was conducted via video call. Participants were asked to share their screen during the observational study; they were told their audio and video would be recorded as well as their responses to WordBot. All participants were shown the same example inputs when during the demo of WordBot.

Data Analysis

To conduct the thematic analysis, we transcribed the participants' recordings using the speech-to-text service Otter.ai. We read through the generated transcripts and checked them against the original recordings for consistency. We then analyzed the transcripts and noted recurring utterances and themes. Together, we developed codes with example utterances; we used our notes and research questions to inform our coding. One author then coded the transcripts to produce the final coding. This was used for our discussion section and compared with the respective findings from Gero et al.

RESULTS

Participants inputted an average of 22 (±5.17) questions out of 132 questions total (or 264 individual words) WordBot. These words consisted of 192 nouns, 58 adjectives, 13 verbs, and 1 adjective. 85.3% of these words occurred once or twice while the remaining occurred 3 to 9 times in total. Table 1 describes the 8 codes from the results of the thematic analysis. We use the 4 most prevalent codes to justify our observations and inform our discussions.

We then compare users' mental model's formed during the study and WordBot's conceptual model by using the 3 concept framework created in *AI Agent Characterization*. We then discuss 3 patterns of behaviour we observed: *Negotiating Outputs and Confirmation Bias*, *Prior AI Knowledge*, and *Establishing Mental Models Early*.

Code	Prevalence	Description and Example Quote
Justification	25%	Explanation as to AI agent's responses. *P4: this is kind of what I was expecting that, you know, makes sense in an additive way, like hot and cold kind of cancel out.*
Prior AI knowledge	17%	Discussion or reference to participants prior experience/knowledge of AI systems. *P1: I'm thinking more of like I did my project on the Google Home, stick with stuff. So things that are kind of related along that type of conversational agent.*
Confirmation	16%	Discussion on how a response validates an assumption about the AI agent. *P1: It says a "year" or "weekend" in time. So that's approximately what I would expect from that.*
Surprises	11%	Noting unexpected result(s) from the AI agent, times when their expectation were challenged. *P3: the word "pet" is here, just not quite what I would expect because it's not the type of animal that you commonly eat*
AI knowledge	10%	Discussion on the AI agent abilities and what it knows or does not know. *P5: the thing I found is that the order of inputs doesn't matter.*
Similarities	9%	Noting words appear to be grouped in the same category or share commonalities. *P6: These words are related to each other somehow and both of them belong to the same category, the output will be some other words in that category.*
Synonyms/antonyms	5%	Any mention of synonyms or antonyms in relation to the input or output words. *P6: So if like, one of the inputs is a subset of the other input, I will get a synonym for like the superset word.*
Weighting/importance	2%	Noting that certain words have a difference weighting or importance than others. *P1: So again it still seems to be putting a lot of weight on here the placement over the person.*

Table 1. Thematic analysis of the participant's transcripts. Prevalence is calculated as the number of utterances of a certain code divided by the number of total utterances of all codes.

Mental and Conceptual Model Comparison

Global Behavior

- There was no indication participants thought WordBot could remember past inputs.

Knowledge Distribution

- All participants appeared to assume each input had a single input which was the one they intended.
- Participants did not comment on the semantic ambiguity of WordBot.

Local Behavior

- A third participant tested for symmetry and concluded the order of the inputs had no influence on the output results.
- All participants mentioned certain outputs sharing categorical and semantic similarity to their inputs at least once.
- A third of participants noticed that the inputs' semantic similarities are preferred over their literal concatenation.
- A third of participants were surprised by the presence of antonyms in the outputs.

Overall, participants were able to detect some aspects of the conceptual model, but their justifications varied. These aspects depended on what area participants focused on during their interactions (eg: P1 focused on synonyms/antonyms, while P2 focused on numbers/mathematics). We discuss these justifications and patterns of behaviour in the section below.

Negotiating Outputs and Confirmation Bias

All participants engaged in some form of "negotiating" in what they believed to be an acceptable response from WordBot for some outputs.

In some cases, even if the results were not inline with their initial assumption, participants indicated a result was acceptable if they could construct some justification in hindsight. This often meant pointing out synonyms, antonyms, or categorical similarities of the output word to either the input words or to their initial guess. For example:

> Input: *leather + alive*
> P3: That'd be a cow, an animal.
> Output: *Snake skin* (2nd result)
> P3: All right. It's not too bad.

P3 initially guessed *cow* or *animal* but accepted *snake skin* as a result even though it was not related to his initial assumption. There were also cases participants did not have an initial assumption in which they relied solely on the input words for justification.

This is not to say participants thought WordBot was always correct. There were many cases WordBot produced unexpected results that participants were confused by and rejected. However, for more ambiguous results, participants appeared to operate within a large margin of acceptability as long as they could "negotiate" a justification between their assumptions and the output. This is shown through the prevalence of the **justification** code, in which most of these situations occurred.

Participants also exhibited confirmation bias in assumptions they made early on. An unexpected outcome of returning multiple outputs was that participants would sometimes selectively ignore "incorrect" outputs as long as a "correct" or agreeable output was present. For example, when guessing *pet + house*, P5 was expecting *kennel* which she found in the output as rank 4. She made no comment on the other outputs which appeared to not agree with her justification.

Prior AI Knowledge

All participants had experience with AI systems beforehand. We found all participants relied heavily on their preexisting AI knowledge when making assumptions about WordBot, this is shown in the prevalence of the **prior AI knowledge** code. Many spoke about their mental model in terms of the possible technology behind the WordBot.

> "So again, going back to that original hypothesis I had that the addition is a union. When I think of addition, I think I'm adding more information to it. I think that goes back to my experience using Google Home. . . " (P1)

> "I've heard of bots that use vector spaces that from what I know would work well with these combinations. So basically every word is a high, high dimension vector. And if you add up different vectors, you'll get a new vector. And the closest word to that vector is usually closely related to what you would get when you add up these two words." (P3)

> "I don't think that this agent has notions of like ordering built into it. So I think it really is just doing some sort of vector operations in an embedding space for these two things." (P4)

Participants were reluctant to revise their mental models in the face of surprises. Their prior AI knowledge was a frequent utterance when justifying WordBot's behavior.

Establishing Mental Models Early

All participants were able to form some mental model of the agent when asked after the study. These models formed quickly early on and were often based in their prior AI knowledge. Within 2 interactions, P4 arrived at mental model based on prior AI knowledge which he cited throughout the study:

> "I probably had some preconceptions, which is not the not ideal when exploring something new but I walked in wondering what it could be under the hood. And seeing that it was associated with semantic meanings with words immediately made me think of things like Word2Vec. I think that was reinforced pretty quickly, even before I started interacting with the interface. Seeing this plus sign here was something that said to me that maybe this presumption is correct. So when I jumped into it, I immediately started trying to think of ways to test if that was the case." (P4)

5 of the 6 participants formed some mental model or theory within the early first half of the study which did not change throughout the rest of their interactions. P1 formed his idea of the "union of terms" while P2, P5, and P6 established that words were being grouped and evaluated within various categories (these occurred in the **similarities** utterances).

Most of the **surprises** utterances occurred in the earlier part of the study, with **confirmation** and **justification** occurring near the end. Whatever mental model participants created, they were able to come to terms with by the end of the study in *Negotiating outputs and confirmation bias*.

DISCUSSION

Comparison with Geo et al. [3]

We found the 3 concept AI framework to be useful and comprehensive in describing both the mental model of our AI agent and its conceptual model. We found aspects of the agent, as discussed above, could be easily categorized into the 3 concepts.

Our coding shared the codes of **synonyms/antonyms, anomalies/distress/trust (as surprises)**, and **AI knowledge**. We believe the code **synonyms/antonyms** is reflective of the shared linguistic context of the studies. But **AI knowledge** and **surprises** could be indicative of themes that emerge when interacting with AI agents in general.

We also compare our findings to the following:

> "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." (Gero et al., 2020)

We found our participants' prior AI knowledge heavily influenced how they formed their mental model. Participants tended to maintain their model which was formed early on in the face of anomalies or surprises. This difference may be due to the AI experience and Computer Science background of our participants in which they had more confidence in their assumptions due to their prior knowledge.

Research Questions

To address **RQ1**: how do users develop mental models of AI systems? We found people with prior AI knowledge developed mental models early into the interaction with the AI agent, and maintained these models for the rest of the study. We also found they relied heavily on their prior AI knowledge to form their models and provide justification for the agent's responses.

The most prevalent codes that appeared were **justification**, **prior AI knowledge**, **confirmation**, and **surprises**. People began with **surprise** utterances which later changed to **confirmation** utterances near the end part of the study. Once P1 had the idea the agent uses a "union of terms", his remaining interactions were spent confirming this idea. Similarly, P4 arrived at his idea that the agent uses Word2Vec almost immediately and conducted his interactions accordingly. P3 was the notable exception, he did not arrive at a mental model quickly, but had the highest number of **surprises** that occurred through his interactions – this differed from the rest of the participants.

To address **RQ2**: what are the implications of AI systems not providing justification of their responses to users? At the beginning of the paper we posed that the current standard for AI systems can lead to problematic outcomes. The current standard for these systems is similar to what participants experienced with our agent: brief interactions with limited information as to the conceptual model of the system. We observed 3 recurring behaviours in people's interaction with the AI agent: confirmation bias, prior AI knowledge, and establishing mental models early. We believe the combination of these behaviors is problematic.

Having prior AI knowledge allows people to establish their mental models early which can be selectively confirmed as they interact with the system. We found the same aspects of the agent's conceptual model could be justified in different ways depending on the mental model of the user. This affords a lot of flexibility to the AI's responses, which can be exacerbated through confirmation bias.

In other words, when an AI system does not provide a justification or insight into its results, users are left to construct their own. By providing multiple answers, utilizing confirmation bias, using other tactics, and not providing answer justification, AI systems can exploit users' tendencies. Thus, the acceptability of an AI system's responses may depend on the user being able to construct a satisfying justification rather than the inherent quality of the system.

Remarks

We may be tolerating poor AI systems by having to construct a justification for their responses instead of the other way around when they are kept as black boxes.

An AI system should provide justification for its responses. When left to interpret and justify the system's outputs, we give great leniency and tolerance to the system. This may result in a compromise of accuracy and quality for a perceived sense of validity or agreement with our mental models.

AI systems have begun to replace or share responsibilities of professionals such as diagnosing diseases, making investments, or recruiting candidates. They are becoming involved in high-stakes decisions across sectors. Therefore, AI systems should be held to at least the minimum of the standards as their human counterparts: decisions should come with justification.

Limitations

Our study has several limitations. We were not able to measure what was a "right" or "wrong" answer from the AI agent's due to the subjective nature of semantic analysis other than what the participants. Thus, it was difficult to measure the accuracy when comparing models. Participants' interaction with our AI agent was only 15 min. It is important to explore how people's mental models mature or develop over longer periods of time and more interactions. Interaction was also limited to 2 input terms with a summation operation, as one participant pointed out:

> "I wish there were more ways to interact with it [WordBot], cause right now I feel like I have a hypothesis and I'm trying to sort of falsify it, but it's tough to do." (P4)

In addition, we only tested a single AI agent (WordBot), it is important to explore a variety of AI systems to understand what aspects may be context/setting-specific vs. what can be concluded about AI systems in general.

The confirmation bias we observed could be in part to users being more lenient or tolerant of acceptable answers because they knew the interactions were part of the study.

CONCLUSION

We studied conceptual and mental models of AI systems using an AI agent in a semantic analysis setting and investigated the implications of users' models when given limited insight into the agent. We created a conceptual model of the AI agent using the framework from Gero et al. [3]

We investigated participants' mental models of the agent in a remote think-aloud study and subsequent interview. We compared people's mental models and the conceptual model of the agent, and found people detected aspects of the conceptual model, but their justifications for these aspects varied.

We conducted a thematic analysis on the data and compared our results to the work of Gero et al. to which most of our findings were in agreement. We found that people with AI experience rely heavily on their prior knowledge, establish mental models early, and are susceptible to confirmation bias. They tend to not update their mental model in the face of surprises, in favor of their prior knowledge.

REFERENCES

[1] Murhaf Fares, Andrey Kutuzov, Stephan Oepen, and Erik Velldal. 2017. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In *Proceedings of the 21st Nordic Conference on Computational Linguistics (Link őping Electronic Conference Proceedings)*. Linkőping University Electronic Press, Linkøpings universitet, Linkőping, Sweden, 1–6. https://ep.liu.se/ecp/article.asp?issue=131&article=037#

[2] Krishna Gade, Sahin Cem Geyik, Krishnaram Kenthapadi, Varun Mithal, and Ankur Taly. 2019. Explainable AI in Industry. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19)*. Association for Computing Machinery, New York, NY, USA, 3203–3204. DOI: http://dx.doi.org/10.1145/3292500.3332281

[3] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R. Millen, Murray Campbell, Sadhana Kumaravel, and Wei Zhang. 2020. Mental Models of AI Agents in a Cooperative Game Setting. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. DOI: http://dx.doi.org/10.1145/3313831.3376316

[4] José Hernández-Orallo and Karina Vold. 2019. AI Extenders: The Ethical and Societal Implications of Humans Cognitively Extended by AI. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19)*. Association for Computing Machinery, New York, NY, USA, 507–513. DOI: http://dx.doi.org/10.1145/3306618.3314238

[5] NLPL. 2020. Vectors/home. (2020). http://wiki.nlpl.eu/index.php/Vectors/home.

[6] Will Knightarchive page. 2017. The Dark Secret at the Heart of AI. MIT Technology Review. (11 April 2017). Retrieved August 9, 2020 from https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/.

[7] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global Vectors for Word Representation.. In *EMNLP*, Vol. 14. 1532–1543.

[8] Parker Robert, Graff David, Kong Junbo, Chen Ke, and Maeda Kazuaki. 2011. English Gigaword Fifth Edition. https://catalog.ldc.upenn.edu/LDC2011T07

[9] Darrell M. West and John R. Allen. 2018. How artificial intelligence is transforming the world. Center for Technology Innovation. (28 April 2018). Retrieved August 9, 2020 from https://www.brookings.edu/research/how-artificial-intelligence-is-transforming-the-world/.