It Takes Two to Tango: Towards Theory of AI's Mind

Arjun Chandrasekaran[*,1] **Deshraj Yadav**[*,2,] **Prithvijit Chattopadhyay**[*,2] **Viraj Prabhu**[*,2]
Devi Parikh[1]

[1]Georgia Institute of Technology [2]Virginia Tech

{carjun, parikh}@gatech.edu, {deshraj, prithv1, virajp}@vt.edu

Abstract

Theory of Mind is the ability to attribute mental states (beliefs, intents, knowledge, perspectives, etc.) to others and recognize that these mental states may differ from one's own. Theory of Mind is critical to effective communication and to teams demonstrating higher collective performance. To effectively leverage the progress in Artificial Intelligence (AI) to make our lives more productive, it is important for humans and AI to work well together in a team. Traditionally, there has been much emphasis on research to make AI more accurate, and (to a lesser extent) on having it better understand human intentions, tendencies, beliefs, and contexts. The latter involves making AI more human-like and having it develop a theory of our minds.

In this work, we argue that for human-AI teams to be effective, humans must also develop a theory of AI's mind – get to know its strengths, weaknesses, beliefs, and quirks. We instantiate these ideas within the domain of Visual Question Answering (VQA). We find that using just a few examples (50), lay people can be trained to better predict responses and oncoming failures of a complex VQA model. Surprisingly, we find that having access to the model's internal states – its confidence in its top-k predictions, explicit or implicit attention maps which highlight regions in the image (and words in the question) the model is looking at (and listening to) while answering a question about an image – do not help people better predict its behavior.

1. Introduction

Background. Internal states of an agent are often inaccessible to other agents. Some primates have overcome this limitation by acquiring, through natural selection, the ability to make sense of, and (to an extent) successfully predict the behavior of other agents [38]. Indeed, our ability to estimate the beliefs, feelings, and actions of other agents in novel situations forms the basis of human social cognition [46]. The capacity to attribute mental states[1] to other agents that are different from one's own, is called *Theory of Mind* [53][2].

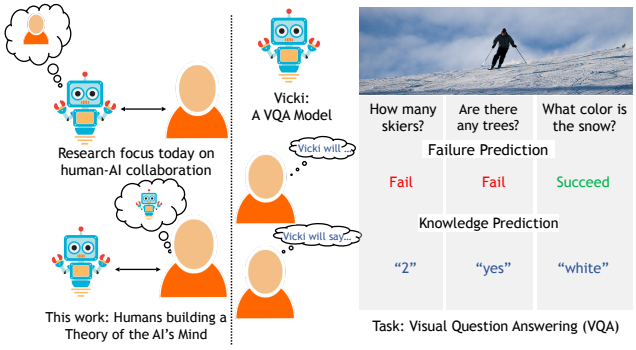


Figure 1: Current emphasis of research in human-AI collaboration is on AI modeling a human teammate's mental state (left top). We argue that for human-AI teams to be effective, humans must also have a model of the AI's strengths, weaknesses, and quirks. That is, humans must develop a Theory of AI's Mind (left bottom). In this paper, we instantiate these ideas in the context of Visual Question Answering (right). Human subjects predict the success or failure (Failure Prediction), and output responses (Knowledge Prediction) of a VQA model (we call Vicki).

Humans frequently attribute mental states to fellow humans (conspecifics) and can make reasonable inferences about their behavior. However, upon encountering a novel agent, for instance, an artificially intelligent (AI) agent, can humans estimate its behavior? Research suggests that to understand new entities such as a robot, humans project existing preconceptions and social constructs upon them [26]. However, as recent research has shown [17, 28], the behavior of an AI agent is often quite different from that of a human – sometimes in ways that are surprising. Thus, inferences based on existing social constructs or preconceptions may fail while estimating the behavior of AI agents. In this work, we consider the novel problem of improving a person's estimate of the behavior of a complex AI agent by

* Denotes equal contribution.

[1]including beliefs, intents, desires, knowledge, etc.

[2]Inferences about behavior by attributing mental states to an agent is a *theory* because a) mental states are not directly observable, and b) the system can be used to predict the behaviors of other agents.

familiarizing the person with the agent. Further, we explore whether representations of the agent's internal states aid in this process.

Motivation. As AI progresses, we find ourselves working with AI agents increasingly often. Intelligent virtual assistants like Siri, Cortana, Google Assistant, and Alexa make our lives more convenient. Doctors collaborate with IBM's Watson [24, 59], dividing work based on their expertise to make better informed diagnoses [35]. Visually-impaired users are starting to rely on computer vision algorithms to interpret the world around them [73, 43, 11]. In-vehicle AI in autonomous cars leverage humans' experience to make decisions in unpredictable situations [70]. There is also an increased interest in using computer vision algorithms for medical imaging [64].

Clearly, in each of these cases it is critical for the human to have a sense for what the AI is good at (vs not), or when the AI might fail and should not be trusted. The human-AI team will be more effective if the human collaborating with the AI agent has a deeper understanding of the AI agent's behavior. However, AI research has traditionally placed much of the burden on the AI to play its part in the team: to be more accurate [32, 56, 65, 57, 39], more human-like [4, 36, 48, 23, 3, 54, 50, 10, 13], understand our intentions [49, 67, 68], beliefs [22], tendencies [18], contexts [55], and mental states [20, 19]. In this work, we argue that for human-AI teams to be effective, Theory of Mind must go both ways. Humans must also understand the AI's beliefs, knowledge, and quirks. See Fig. 1.

Effective teams. Effective communication involves considering a teammate's background knowledge, abilities, preferences and modifying one's interactions accordingly [31]. Indeed, recent studies [21, 72] conclude that the most effective teams are those with members who, among other traits, demonstrated good Theory of Mind abilities. In this work, we consider two tasks that we believe demonstrate varying degrees to which a human understands an AI team member. The first task, Failure Prediction (FP), involves estimating whether an agent in a specific situation will succeed or fail. This is especially relevant for human-AI teams since AI agents often fail in ways that are different from humans [28, 78, 7]. Accurately estimating the success or failure of an agent in a specific situation is one way to measure whether a person understands the strengths and weaknesses of an agent. The second task, Knowledge Prediction (KP), involves estimating the exact response of an agent in a given situation. Making an accurate estimation of the response of the agent requires a deeper understanding of its behavior.

Our setup. Effective communication in human teams is task-oriented, with grounding in a common knowledge base [16]. In our work, we consider an AI agent that is capable of grounded natural language communication with humans. Specifically, we consider an agent trained to perform the multi-modal task of Visual Question Answering (VQA) [5, 42, 66, 11, 27]. Given an image and a free-form open-ended natural language question about the image, the AI agent's task is to answer the question accurately. Call this agent – a VQA model – Vicki. VQA is applicable to scenarios where humans (e.g., visually impaired users, surveillance analysts, etc.) actively elicit information from visual data. It naturally lends itself to human-machine teams. The human teammates in our experiments are from Amazon Mechanical Turk (AMT). In Failure Prediction (FP), we show AMT subjects an image and a question about the image, and ask them to estimate if Vicki will correctly answer the question. In Knowledge Prediction (KP), subjects are asked to estimate Vicki's exact response.

We study the extent to which humans can accurately estimate the behavior of Vicki. Then, we explicitly aid humans in developing a theory of Vicki's mind by (1) familiarizing them with Vicki's actual behavior during a training phase and (2) exposing them to Vicki's internal states via several existing 'explanation' modalities. We evaluate if these explanation modalities aid humans in accurately estimating Vicki's behavior (FP and KP).

Applicability of Theory of Mind to AI. Humans have an innate tendency to anthropomorphize [34]. We often attribute human traits to non-human entities. For instance, Human-Robot Interaction (HRI) research finds that people attribute a 'personality' and 'mind' to non-human entities like robots. Interestingly, their perception can also vary based on surface features like the robot's appearance and behavior [12, 26]. Thus, it is evident that humans attempt to understand and reason about actions of non-human entities like robots by attributing mental states to them.

Theory of Mind involves attributing mental states to other agents. For human agents, these mental states encompass a large range of states, such as desires, beliefs, and knowledge. Some of these notions, like beliefs and knowledge, are clearly applicable to AI. While a Theory of a Human Mind might appear to involve many complex mental states, in practice, it is measured by a fairly simple test – "reading the mind in the eyes" [9][3]. In a similar vein, we propose the two tasks of FP and KP as simple yet effective techniques to measure a person's understanding and estimation of an AI agent's behavior, i.e., their Theory of an AI's mind (ToAIM)[4].

Contributions. The contributions of this work are:

1. We advocate a line of research to study the extent to which humans can build a Theory of AI's Mind (ToAIM) and develop approaches to aid the process.

[3]This test involves looking at a photo of a human's eyes and choosing one of two adjectives that better describes the person's mental state.

[4]On the subject of whether AI can have a mind at all, a number of philosophers suggest that it can. For instance, in 'society of mind' [45], Minsky says that a mind simply emerges as a result of complex interactions between many smaller non-intelligent entities which he calls agents.

2. As a specific instantiation of this, we consider the problem of VQA where the AI's task is to answer a free-form natural language question about an image.
3. We conduct large-scale human studies to measure the effectiveness of training, and of different explanation modalities, in helping humans accurately predict the successes, failures,and output responses of a VQA model on question–image pairs. To the best of our knowledge, this is the first evaluation that measures whether interpretability mechanisms do, in fact, allow humans to build a model of AI.
4. We will make the interfaces we used for our human studies publicly available, to enable others to both evaluate a person's Theory of AI's Mind and further investigate the effectiveness of training and of explanation modalities in improving the same.
5. Our key findings are that (1) humans are indeed capable of predicting successes, failures, and outputs of the VQA model better than chance. (2) explicitly training humans to familiarize themselves with the model by using just a few examples improves their performance (3) existing explanation modalities do not enhance human abilities at predicting the model's behavior. While most prior work on interpretabilty has focused on qualitatively demonstrating the role of such explanation modalities in improving human trust, our findings indicate that they do not yet help humans build more accurate mental models of AI. We believe that as computer vision and AI technology matures, developing improved modalities that can aid humans in this process will be critical towards developing successful human-AI teams.

2. Related Work

AI with a theory of (human) mind. A number of works in AI attempt to develop agents with an understanding of human characteristics and behavior. AI agents employing computer vision have been trained to predict the motivations [67], intentions [49], actions [68], tendencies [18], contexts [55], etc., of humans. In addition, Scassellati [60] examines theories that explain the development of Theory of Mind in children and their applicability to building robots with similar capabilities. More recently, in the domain of abstract scenes, Eyesenbach et al. [22] address the problem of identifying incorrect beliefs in people. The ability to identify false beliefs [71] in other agents is considered an important milestone in the development of Theory of Mind in an agent [8]. Unlike these works where AI agents "understand" humans, our work addresses the converse problem – to have humans understand AI agents, their quirks, weaknesses, and beliefs.

Explainable AI. Recently, there has been a thrust in the direction of "explainable" AI agents in vision-related tasks. **Introspection vs Justification:** Generating explanations for classification decisions has attracted considerable interest. Several works propose introspective explanations based on internal states of a decision process [77, 61, 30, 79], while others generate justifications consistent with model outputs [58, 33, 51]. Riberio et al. [58] explain the predictions of a classifier by learning an interpretable model locally around the prediction. Hendricks et al. [33] develop a justification system that produces explanations consistent with visual recognition decisions. **Natural language vs Visual explanations:** Prior art has assessed the usefulness of natural language explanations of model decisions in improving model trust [33]. MacLeod et al. [41] investigate the role of *phrasing* of a model's confidence in blind and visually impaired persons' trust in image captioning models. Park et al. [51] propose a pointing and justification model for VQA that can both justify predictions in natural language and also point to visual evidence. **Explicit vs Implicit attention:** There is a line of work in designing models that explicitly attend to relevant parts of their input for vision tasks such as object recognition [6, 47], image captioning [75, 15], and VQA [40, 76, 74]. In contrast, recent work by Zhou et al. [80] and Selvaraju et al. [61] expose implicit attention for predictions from CNN-based models as visual explanations.

Across these works, the focus is on making AI agents more transparent and capable of explaining their decisions in order to build trust. In our work, we explore the role explanation modalities play in improving a human's model of the AI, as measured by the human's accuracy at predicting the AI's success, failure, and output responses.

Failure Prediction. There exists prior art that deals with building models that predict failure modes of systems [7, 78]. Whereas these works employ statistical models to predict failure modes of a *base system*, we evaluate the role a training phase as well as explanation modalities play when *humans* perform the same task. In addition to predicting the success or failure of AI agents, we also train humans to more accurately predict the "knowledge", i.e., the actual output of an AI agent.

Humans adapting to technology. A few works [69, 52] observe human strategies while adapting to the limited capabilities of an AI agent in interactive language games. For instance, in a human-AI game of charades, humans modify strategies such as word selection, turn length, and prosody, to adapt to the robot's limited perceptive abilities. While both these works observe that humans dynamically adapt their behavior while interacting with an AI on a particular task, in our work we explicitly measure to what extent humans have formed an accurate model of the AI. We also evaluate the role that explanation/interpretability modalities play in helping humans build a more accurate model.

3. Meet Vicki

We instantiate the idea of humans building a Theory of AI's Mind in the VQA task. Our AI agent (that we call

Vicki) is a VQA model trained to answer a free-form natural language question about an image. Concretely, we use the VQA model by Lu et al. [40]. It is a hierarchical coattention model that models the question at multiple levels of granularity (words, phrases, entire question) and at each level, has explicit attention mechanisms on the image (where to look) as well as the question (which words and phrases to listen to). Among the different variants introduced in [40], we use the alternating co-attention model trained with VGG-19 [63] as the CNN to derive image-representations.

Vicki was trained on the VQA dataset [5] train split containing 248349 QI pairs, and outputs one of a 1000 possible answers (most frequent in the train split). Its accuracy on the VQA dataset (test-standard) is 62.2%[5] (human accuracy is 83.3%), which was the state-of-the-art at publication [40] and is still competitive today. Despite being 4.7% less accurate than the current state-of-the-art VQA model [37], Vicki's image and question attention maps provide access to its 'internal states' while making a prediction. These maps highlight the regions of the image and words of the question that Vicki attends to. This presents an opportunity to assess the role such explanation modalities can play in aiding humans better predict Vicki's behavior. Among the various settings explored in [40], we use the *question-level* image and question attention maps in our experiments.

Vicki is Quirky. There are several factors that contribute to Vicki being quirky, in a predictable fashion. Some of these quirks are well-known in VQA literature [2]. **Vision is not perfect:** Vicki, like most other vision models, has a limited capability to understand the image. Observing Vicki's behavior during its failures demonstrates its quirks. For instance, when the question asks the color of a small object in the scene, say a soda can, Vicki may simply respond with the most dominant color in the scene. This is clearly evident when we observe the distribution of Vicki's responses across a diverse set of images [2]. **Language is not perfect:** Vicki has a limited capability to understand free-form natural language. Vicki seems to converge on a predicted answer after listening to just half the question 49% of the time [2]. So in many cases, it answers questions based only on the first few words of the question alone [2]. **Vicki cannot reason:** Vicki has no mechanism to leverage external knowledge and reason about common sense. Vicki is poor at compositionality – it is unable to disentangle and recompose concepts seen in training to generalize to unseen test concepts [2]. Vicki does not have an explicit counting mechanism [14]. So it often defaults to the popular answer "2" for "How many" questions. **Vicki cannot say much:** Since Vicki is a 1000-way classifier, it only has a fixed set of utterances. **Vicki answers every question:** Vicki was trained only on questions that were relevant to the image. Thus, Vicki does not know how to say "That doesn't make sense." or "There *is* no woman in this image." when asked "What color is the woman's shirt?" on an image that does not contain a woman. Thus, when posed with a question that is irrelevant to the image, Vicki is forced to provide an answer from its limited vocabulary. Interestingly, because Vicki is a deterministic function of the question and image, observing its response across QI-pairs often gives us a sense for what it might be basing its responses on. **Vicki may ignore the image:** Vicki picks up on the language priors that are inherent in the world which are easier to leverage than complicated visual signals. For example, when the question "What color is the banana?" is asked, Vicki often ignores the image and answers "yellow". **Vicki is biased:** Vicki is very likely to answer "yes" to a yes/no question, and answer "white' to a "what color" question due to biases inherent in the VQA dataset that it was trained on [29].

[5] http://www.visualqa.org/roe.html

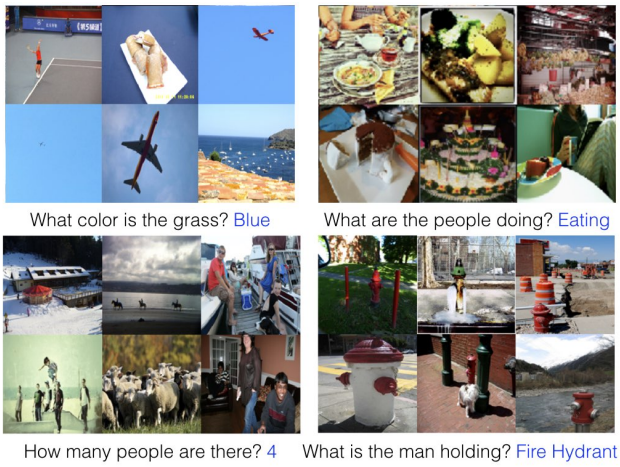
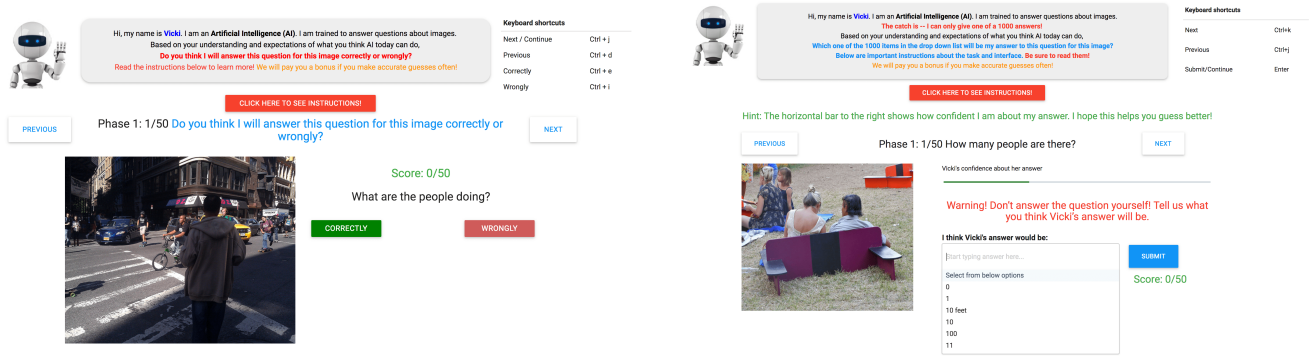


Figure 2: These montages highlight some of Vicki's quirks. For a given question, Vicki has the same response to each image in a montage. Common visual patterns (that Vicki presumably picks up on) within each montage are evident.

To get a sense for this, see Fig. 2. The patterns are clear. In top-left, even when there is no grass, Vicki tends to latch on to one of the dominant colors in the image. For top-right, even when there are no people in the image, Vicki seems to respond with what people could *plausibly* do in the scene if they were present. A priori, one (especially lay people) may not expect this. But when exposed to several examples of Vicki's responses, it is conceivable that subjects may begin to have an understanding of Vicki's behavior and consequently form a theory of its Mind.

4. Meet the tasks

We present two tasks that can measure a human's understanding of the capabilities of an AI agent such as Vicki. These tasks are especially relevant to human-AI teams since they are analogous to measuring if a human teammate's trust in an AI teammate is well-calibrated, and if a human can estimate the behavior of an AI in a specific scenario.



(a) The Failure Prediction interface.

(b) The Knowledge Prediction interface.

Figure 3: (a) A person guesses if an AI agent (Vicki) will answer this question for this image correctly or wrongly. (b) A person guesses what Vicki's exact answer will be for this question for this image.

Failure Prediction (FP). In this task, we study the ability of a human to predict the success or failure of Vicki. That is, given an image and a question about the image, we measure how accurately a person can predict if Vicki will successfully answer the question. A person can presumably predict the failure modes of Vicki reasonably well if they have a good sense of Vicki's strengths and weaknesses. A collaborator who performs well on this task can accurately determine whether they should trust Vicki's response to a question about an image. Please see a snapshot of the FP interface in Fig. 3a. Note that we do not show the human what Vicki's predicted answer is.[6]

Knowledge Prediction (KP). In this task, we measure the capability of a human to develop a deeper understanding of Vicki's behavior. Given an image and a question, a person guesses Vicki's exact response (answer) from a set of its output labels (vocabulary). Recall that Vicki can only say one of a 1000 things in response to a question about an image. Please see a snapshot of the KP interface in Fig. 3b. We provide subjects a convenient dropdown interface with autocomplete to choose an answer from Vicki's vocabulary of 1000 answers.

In FP, a good understanding of Vicki's strengths and weaknesses might lead to good human performance. However, KP requires a deeper understanding of Vicki's behavior, deeply rooted in its quirks and beliefs. In addition to reasoning about Vicki's failure modes, one has to guess its exact response for a given question about an image. Note that KP measures subjects' ability to take reality (the image the subject sees) and translate it to what Vicki might say. High performance at KP is likely to correlate to high performance at the reverse task – take what Vicki says and translate it to what the image really contains. This can be very helpful when the visual content (image) is not directly available to the user. Explicitly measuring this is part of future work. A person who performs well at KP has likely successfully modeled a more fine-grained behavior of Vicki than just modes of success or failure. In contrast to typical efforts where the goal is for AI to approximate human abilities, KP involves measuring a human's ability to approximate a neural network's behavior!

5. Perception of AI

Before introducing people to Vicki and gauging their expectations from a modern VQA system, we attempt to assess their general impressions of present-day AI. We then observe correlations in their expectations from Vicki with their familiarity with AI, their estimates of AI's capabilities, and their demographic and socio-economic background. We ask each subject to fill out a survey with questions aimed to collect three types of information:

1. **Background information.** We collect basic demographic information such as age, gender, educational qualifications, type of residential area, and profession. We also collect socio-economic background information such as employment status and income group.

[6] Otherwise, given an image from the COCO dataset (everyday consumer images) and a question from the VQA dataset (mostly mundane questions about everyday objects and scenes), it would be trivial for the human to verify if Vicki's predicted answer is right or wrong. In general, one might wonder why a human would need Vicki to answer questions if they are already looking at the image. This may be true for the VQA dataset, but outside of that there are scenarios where the human either does not know the answer to a question of interest (e.g., the species of a bird), or the amount of visual data is so large (*e.g.*, long surveillance videos) that it would be prohibitively cumbersome for them to sift through it. Note that even in this scenario where the human does not know the answer to the question, a human who understands Vicki's failure modes from past experience would know when to trust its decision (*e.g.*, if the bird is occluded, or the scene is cluttered, or the lighting is bad, or the bird pose is odd, Vicki will fail). Moreover, the idea of humans predicting the AI's failure (and ToAIM in general) also applies to other scenarios where the human may not be looking at the image, and hence needs to work with Vicki (*e.g.*, blind user, or a human working with a tele-operated robot). In these cases too, it would be useful for the human to have a sense for the contexts and environments and/or kinds of questions for which Vicki can be trusted. In this work, as a first step, we focus on the first scenario where the human is looking at the image and a question while predicting Vicki's failures and responses.

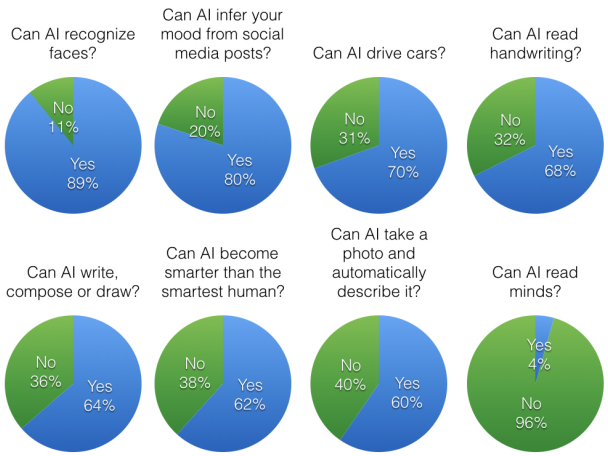


Figure 4: People's estimate of AI's capabilities.

2. **Familiarity with computers and AI.** We ask subjects if their jobs involve computers, if they know how to program, how much time they spend in front of a computer or smartphone, and their familiarity with popular AI assistants such as *Siri*, *Alexa* and *Google Assistant*. We also ask if they are aware of recent advances in AI, especially those trending in popular media, such as *Watson* [25], *AlphaGo* [62], machine learning, and deep learning.

3. **Estimates of AI's capabilities.** We ask subjects their duration and source of exposure to AI and gather their impressions on the capabilities of modern-day AI systems on a range of tasks. We also ask them about their understanding, trust and sentiment towards modern AI systems, as well as their expectations and predictions for AI in the future.

Please find the full list of survey questions and distributions of responses in the appendix. As an interesting tidbit: Fig. 4 shows what % of subjects think certain tasks are "solved". 80-90% of the subjects think AI today can recognize faces and infer your mood from social media posts. 65-70% of subjects think AI today can recognize handwriting, be creative (write, compose, draw), or drive a car. They are more split on whether AI can describe an image in a sentence. However, most (96%) agree that AI today cannot read our minds! Interestingly, 62% of subjects think that AI can become smarter than the smartest human.

6. Perception of VQA

To set the baseline for our specific task, we measure people's current estimates about VQA models. To this end, we briefly introduce Vicki to subjects as an "AI trained to answer questions about images". We then ask subjects to use their current understanding and expectation of what AI agents can do, to estimate the behavior of Vicki. Subjects fill out the survey described above before doing this task.

We study the ability of humans to estimate Vicki's behavior via the Failure Prediction and Knowledge Prediction tasks. For both tasks, we randomly sample questions from the set of ∼1,400 most frequent questions in the validation set of the VQA dataset [5]. A description of our experimental setup for each task follows.

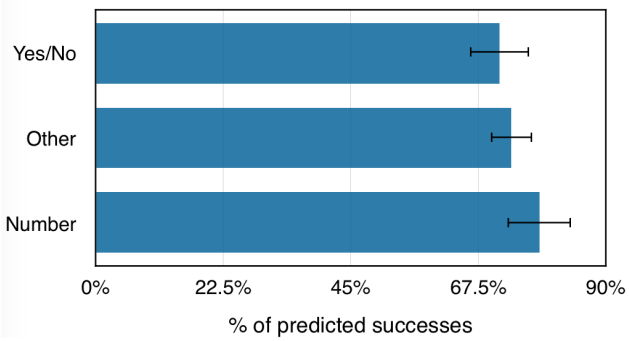


Figure 5: Optimism regarding Vicki: people's estimate of the AI's success in answering a question, across different answer-types.

6.1. Failure Prediction (FP)

In this task we show a question, an image on which this question was asked in the VQA dataset, and ask subjects if they think Vicki's response would be correct or wrong. To get ground truth, similar to the VQA accuracy metric [5], we check if Vicki's response matched at least 3 of 10 human-provided answers in the VQA dataset. Overall, a total of 88 unique subjects participated in our study, providing responses on 1000 QI–pairs. On average, subjects accurately guessed whether Vicki would answer the question correctly (success) or not (failure) 59.88% of the time. The accuracy of always guessing success is 61.52%. While subjects' performance seems lower than this, normalizing for the prior of each class (success vs. failure), always guessing success drops to 50% but humans are at 54.24%. This shows that even without prior exposure to Vicki, human subjects can predict its failure better than chance.

We further measure people's optimism about Vicki's abilities. Fig. 5 shows the percentage of QI-pairs that subjects predicted Vicki would answer correctly for different answer types. We find that subjects expect Vicki to answer questions whose answers are numbers (e.g., counting questions starting with "how many") correctly quite often. Interestingly, today's VQA models are in fact quite ineffective at counting. The VQA leaderboard shows significant improvements in performance on "other" questions over time, but improvements on "number" questions has stalled [1].

Overall, subjects demonstrated an average optimism – as measured by % of "correctly" (success) predictions – of 75.46%. Interestingly, subjects who first heard about AI only in the past 6 months and those who thought that it could read minds are amongst the most optimistic (mean optimism >90%). Those who use a Personal Assistant (PA) on their smartphones frequently (PA usage >3 times a day), as well as older or retired populations (age>60 years) are amongst the least optimistic (mean optimism <57%)!

6.2. Knowledge Prediction (KP)

In the KP task, we ask subjects what they think Vicki would say in response to a question about an image. Note that the VQA dataset only contains questions about an image that are relevant to the image. In the VQA dataset collection protocol, annotators were looking at the image while asking questions. So a question "What color is the man's shirt?" would only be asked for an image that contains a man wearing a shirt.

As an interesting twist intended to elicit Vicki's quirky behavior described in Sec. 3, we also paired images with random (and likely irrelevant [54]) questions (e.g., "What are the people doing?" on an image that may not contain people). Recall that Vicki is forced to respond with a limited vocabulary (one of 1000 answers). These samples are useful to measure a person's understanding of an agent's responses to any given stimulus – including those that come from a distribution under which the agent has not been trained. Note that FP cannot be evaluated on irrelevant images. The notion of a "correct" answer is ill-defined if a question is not relevant to an image.

We performed the KP task[7] on 1000 QI-pairs (700 relevant and 300 irrelevant). We collected 25 responses to each pair. A total of 173 unique subjects participated in our study. The accuracy achieved by predicting Vicki's most popular answer('yes') is 15.79%. Subjects were able to accurately predict Vicki's response 24.81% of the time. Interestingly, younger subjects (age $<$20 years) performed the best (mean accuracy of 35.12%), while those who had heard of AI very recently (in the past month), and those who spent less than 1 hour daily on devices, were amongst the worst performers (mean accuracy $<$20%).

7. Familiarizing people with Vicki

In this section we describe our experimental setup to familiarize subjects with Vicki's behavior. We approach this in two ways – by providing instant feedback about Vicki's actual behavior on each QI pair once the subject responds, and by exposing subjects to various explanation modalities that reveal Vicki's internal states.

Challenges. Collecting data for this setup is challenging for a couple of reasons: (1) Each subject has to go through a training phase to become familiar with Vicki before we can test them. This results in each task on AMT being unusually long and expensive. It also reduces the subject pool down to those willing to participate in long tasks. (2) Once a subject does one task for us, they cannot do another task because the training / exposure to Vicki would leak over. This means we need as many subjects as tasks. This makes data collection quite slow. In light of these challenges, to systematically evaluate the roles of training and exposure to Vicki's internal states, we focus on a small set of questions.

Data. We identify a subset of questions in the VQA [5] validation split that occur more than 100 times. We select 7 diverse questions from this subset that are representative of the different types of questions (counting, yes/no, color, scene layout, activity, etc.) in the dataset[8]. For each of the 7 questions, we then sample a set of 100 images. For FP, the 100 images per question are random samples from the set of images on which the question was asked in the VQA validation split (VQA-val). For the KP task, these 100 images are random images from VQA-val. Ray et al. [54] found that randomly pairing an image with a question in the VQA dataset results in about 79% of pairs being irrelevant. Recall that this combination of relevant and irrelevant question-image pairs allows us to test subjects' ability to develop a robust understanding of Vicki's behavior across a wide variety of inputs.

Task setup. Each human study is comprised of 100 QI-pairs where a single question is asked across 100 images. The motivation behind keeping the question constant is to make it easier for the subject to pick up trends in Vicki's responses across images. The annotation task is broken down into a train phase where the person is shown 50 QI-pairs, and a test phase where we evaluate subject's performance on the remaining 50 QI-pairs.

7.1. Does feedback help?

To familiarize subjects with Vicki, we provide them with instant feedback during the train phase. Immediately after a subject responds to a QI–pair, we show them whether Vicki actually answered the question correctly or not (in FP) or what Vicki's response was (in KP). In the train phase, subjects are also shown a live score of how well they are doing and are allowed to scroll through feedback for previous images (of course, they are not allowed to change their answers to previous images). Once training is complete, no further feedback (including running score) is provided and subjects are asked to draw from the intuition they have built in training to best answer all questions in the test phase. Subjects are also paid a bonus if they do particularly well.

To evaluate the role of instant feedback, we have 2 subjects do our study with and without instant feedback each, for each of the questions (7) and each task (FP and KP). This results in a total of 28 human studies (with 28 unique human subjects). Even without feedback, subjects still go through all 100 images.

[7] We ensured that subjects who perform a KP task are not allowed to perform an FP task since subjects who have performed a KP task are familiar with the set of answers that Vicki is capable of producing which influences their expectation of what Vicki can or cannot do.

[8] What kind of animal is this? What time is it? What are the people doing? Is it raining? What room is this? How many people are there? What color is the umbrella?

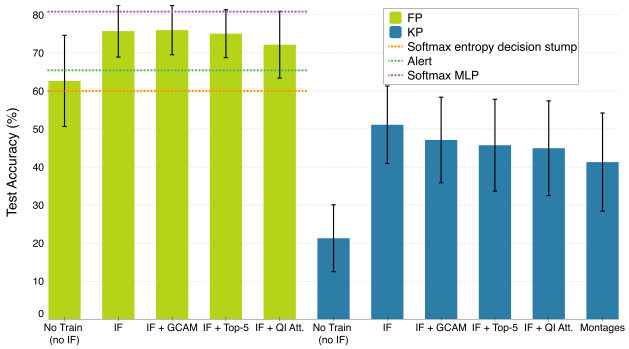


Figure 6: Average performance across subjects for both tasks: Failure Prediction (FP) and Knowledge Prediction (KP), with and without instant feedback (IF), and with various explanation modalities. Error bars are 95% confidence intervals from 1000 bootstrap samples. Note that the dotted lines are various machine approaches applied to FP.

In FP, always answering correctly would result in an accuracy of 58.29%. We find that subjects do slightly better and achieve 62.66% accuracy on FP, even without prior familiarity with Vicki (No Train). Thus, subjects are already slightly better calibrated with an AI's capabilities than unbridled optimism (or pessimism). Further, we find that subjects that receive training as instant feedback (IF) achieve 13.09% (absolute) higher mean accuracies than those who do not (see Fig 6); IF vs No Train (No IF) for FP (green).

In KP, answering each question with Vicki's most popular answer overall ('no') would lead to an accuracy of 13.4%. Additionally, answering each question with Vicki's most popular answer *for that question*[9] leads to an accuracy of 31.43%. Interestingly, subjects who are unfamiliar with Vicki (No Train) achieve 21.27% accuracy – better than the most popular answer overall prior, but worse than the question-specific prior over Vicki's answers. The latter is understandable as subjects unfamiliar with Vicki do not know which of its 1000 possible answers are more likely a priori for each question.

We find that mean absolute performance in KP with IF is 51.11%, 29.84% higher than KP without IF (see Fig 6; IF vs No Train (No IF) for KP (blue)). Subjects thus considerably outperform both the 'most popular answer' and 'most popular answer per question' priors. It is apparent that just from a few (50) training examples, subjects learn to generalize beyond Vicki's favorites among it's vocabulary of 1000 answers. Additionally, the 29.84% improvement over No Train for KP is significantly larger than that for FP (13.09%). This is understandable because a priori (No Train setting), KP is a much harder task as compared to FP, due to the increased space of possible subject responses given a QI-pair, and the combination of relevant and irrelevant QI-pairs in the test phase.

Questions such as 'Is it raining?' have strong language priors – to these Vicki often defaults to the most popular answer ('no'), irrespective of image. We observe that on such questions, subjects perform considerably better in KP once they develop a sense for Vicki's inherent bias via instant feedback. For open-ended questions like 'What time is it?', feedback helps subjects (1) narrow down the 1000 potential options to the subset that Vicki typically answers with – in this case time periods such as 'daytime' rather than actual clock times and (2) identify correlations between visual patterns and Vicki's answer (as seen in Fig. 2). In other cases like 'How many people are in the image?' the space of possible answers is clear a priori, but after IF subjects realize Vicki is not good at detailed counting but does base its count predictions on coarse signals of the scene layout.

In Sec. 3, we described how montages (refer to Fig.2) help highlight Vicki's quirks. In order to test the effectiveness of such montages as a teaching tool, we also experimented with a modification of the KP + IF setting (two unique subjects per question participated in this setting, resulting in an additional 14 human studies). In the train phase of this new setting, instead of individual images, subjects are shown a series of *montages*, each containing 4 to 16 images across which Vicki gave the *same* answer to the question. The objective remains the same – to guess what that answer was (with IF provided after each guess). The test phase is kept identical to the KP + IF test phase, with a single image per question and no IF. We find that subjects achieve 41.6% mean accuracy in the test phase of this setting, which is lower than the mean accuracy in the test phase of the KP + IF setting (51.1%). Interestingly, mean accuracy in the train phase of the montage setting is 68.7%, significantly higher than the mean accuracy in the train phase of the KP + IF setting (49.3%). This seems to indicate that while montages make it much easier to guess Vicki's response correctly by picking out patterns (as seen in Fig. 8 and Fig. 9), the focus on identifying commonalities between groups of images interferes with the ability to pick up on image-level patterns. As a result, subjects do not generalize well to individual images at test time, resulting in worse performance. Keeping the train and test tasks identical (individual images in both cases) is more effective.

VQA Researchers. Just as an anecdotal point of reference, we also conducted experiments across experts with varying degree of familiarity with agents like Vicki. We observed that a VQA researcher had an accuracy of 80% versus a computer vision (but not VQA) researcher who had 60% in a shorter version of the FP task without instant feedback. Clearly, familiarity with Vicki plays a critical role in how well a human can predict its oncoming failures. Our studies

[9] Vicki's most frequent answer (in the train set) to each question is as follows: What kind of animal is this? (Dog) What time is it? (Daytime). What are the people doing? (Standing) Is it raining? (No) What room is this? (Kitchen) How many people are there? (1) What color is the umbrella? (Black)

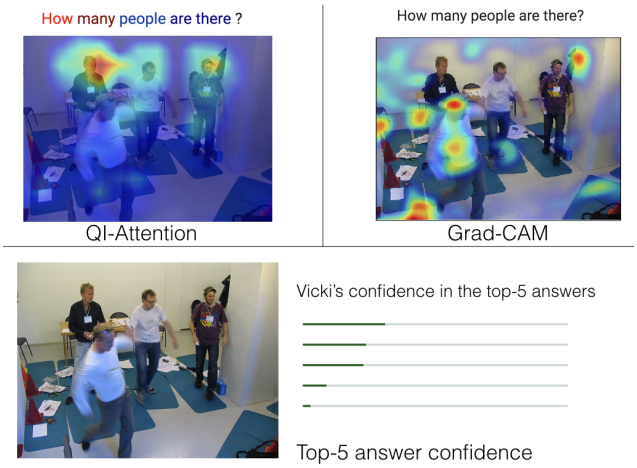


Figure 7: Screenshots of the interfaces of different explanation modalities that we show subjects.

examine the extent to which lay people can be made familiar with Vicki to better predict its behavior.

7.2. Do explanation modalities help?

In this section, we briefly describe the different explanation modalities that we utilize to expose Vicki's internal states to the human subject. In addition to an image and question about the image, we also show the subject one of the three explanation modalities described below. Subjects are asked to use these as hints to perform the task (FP or KP) more accurately. Subjects can leverage the training phase (with instant feedback and a running score) to learn how best they would like to leverage these hints.

We experiment with 3 qualitatively different explanation modalities (see Fig 7):

Confidence of top-5 predictions. We show subjects Vicki's confidence in its top-5 answer predictions from its vocabulary as a bar plot[10]. If Vicki is relatively more confident in its top-1 prediction, it is more likely to be right. If Vicki is confused about the top-5 predictions, it is more likely to be wrong. **Attention maps.** Recall that Vicki is the co-attention VQA model proposed by Lu et al [40] which jointly reasons about image and question attention (Sec 3). Thus, along with the image we show subjects the spatial attention map over the image that indicates the regions that Vicki is looking at and an attention map over each word of the question highlighting the relative importance of words in the question for Vicki, while producing an answer. We show subjects a legend to interpret what the colors in each attention map indicate. **Grad-CAM.** In contrast to explicit attention maps described above, we experiment with an implicit attention map. We use the CNN visualization technique by Selvaraju et al. [61], using the attention maps corresponding to Vicki's most confident answer.

[10]Of course, we don't show the actual top-5 predictions, just the confidence in the predictions.

We have 2 subjects perform each of our tasks (2) for each of the explanation modalities (3) for each question (7) resulting in a total of 84 tasks (and unique subjects). Across all studies (including those described in earlier sections), we have collected over $65k$ responses from 415 unique subjects. Conducting studies in-house in controlled environments at this scale would be prohibitive.

To put human FP accuracies (using explanation modalities) in perspective, we experiment with a few automatic approaches to detect Vicki's failure from its internal states. We find that a decision stump on Vicki's confidence in its top answer or the entropy of its 1000-way softmax output results in FP accuracy of 60% on our test set. We train a Multilayer Perceptron (MLP) neural network on Vicki's output softmax and predict success vs failure. This achieves an FP accuracy of 81%[11]. Training an MLP which takes as input question features (average word2vec embeddings [44] of words in the question) concatenated with image features (fc7 from VGG-19) to predict success vs failure (which we call ALERT following [78]) achieves an FP accuracy of 65%. Note that these methods are trained on about 66% of the VQA–val set (∼81k examples, rest used for validation). Human subjects are trained on only 50 examples.

Accuracies of subjects in the test phase of both tasks (FP and KP) for different settings of the explanation modalities are summarized in Fig. 6. Recall, all studies that include an explanation modality also include instant feedback [12] (IF) and a running score during training. For reference, we also show performance of subjects with no explanation modality both with and without IF. We observe that on both tasks, subjects shown explanation modalities along with IF show no statistically significant improvement in performance over those shown just IF. In fact, in some cases performance is worse. While piloting these tasks ourselves, we found that it was easy to "overfit" to the explanation modalities and hallucinate patterns when none may exist. While the works introducing some of these modalities assessed their interpretability qualitatively or measured their role in improving human trust, our preliminary hypothesis is that these modalities may not yet help human-AI teams be more accurate in a goal-driven collaborative setting because they do not yet help humans predict the AI's behavior more accurately.

[11]Showing a visualization of this score to a human may make for a good "explanation modality" for FP! Exploring this is part of future work.

[12]In real–world settings, we consider familiarizing via instant–feedback, followed by showing explanation modalities, as the natural progression for acquainting subjects with Vicki. Hence, we evaluate the role explanation modalities play on top of instant feedback. Nevertheless, for sake of completeness, studying the effect of showing explanation modalities on subject performance, independent of instant feedback, is part of future work.

8. Conclusion

We posit that as computer vision (and AI in general) makes progress, human-AI teams are imperative. We argue that for these teams to be effective, it is not only important for the AI to be capable of modeling the intentions, beliefs, strengths and weaknesses of the human, but *also* for the human to build a Theory of the AI's Mind (ToAIM). **Take-home message #1**: We should pursue research directions to help humans build models of the strengths, weaknesses, quirks, and tendencies of AI. This is especially relevant in computer vision where input signals are high dimensional and the models we train are becoming ever more complex. We instantiate these ideas in the domain of Visual Question Answering (VQA). We propose two tasks that help measure the extent to which a human "understands" a VQA model (we call Vicki) – Failure Prediction (FP) and Knowledge Prediction (KP) – where given an input instance (question–image pair) the human has to predict whether Vicki will answer the question correctly or not, and what Vicki's exact answer will be. We evaluate the roles that familiarity with Vicki and explanation modalities that expose the internal states of Vicki play. **Take-home message #2**: Lay people indeed get better at predicting Vicki's behavior using just a few (50) "training" examples. **Take-home message #3**: Surprisingly, existing explanation modalities that are popular in computer vision do not help make Vicki's failures more predictable. In fact, humans seem to overfit to the additional information provided and perform slightly worse at KP in the presence of explanation modalities. **Take-home message #4**: Clearly, much work remains to be done in developing improved explanation modalities that do in fact help make AI more predictable to a human.

This work just scratches the surface, and numerous avenues of further exploration remain. Studying other vision models (AI agents in general) at varying points on the interpretability vs performance spectrum for other tasks, evaluating other existing explanation modalities, and conducting human studies at an even larger scale are natural extensions. Relevant to the increased interest in building interpretable models, this work presents novel opportunities to evaluate explanation modalities grounded in specific tasks (FP and KP). Finally, it would be exciting to close the loop and evaluate the extent to which improved human performance at FP and KP translates to improved success of human-AI teams at accomplishing a shared goal. Co-operative human-AI games may be a natural fit for such evaluation.

Acknowledgements. We would like to acknowledge the countless hours of effort provided by the workers on Amazon Mechanical Turk. We thank Satwik Kottur for his help with data analysis, and for many fruitful discussions. This work was funded in part by an NSF CAREER award, ONR YIP award, Sloan Fellowship, ARO YIP award, Allen Distinguished Investigator award from the Paul G. Allen Family Foundation, Google Faculty Research Award, Amazon Academic Research Award to DP. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government, or any sponsor.

References

[1] VQA Challenge Leaderboard . http://www.visualqa.org/roe.html, 2017. 6

[2] A. Agrawal, D. Batra, and D. Parikh. Analyzing the behavior of visual question answering models. *arXiv preprint arXiv:1606.07356*, 2016. 4, 12

[3] H. Agrawal, A. Chandrasekaran, D. Batra, D. Parikh, and M. Bansal. Sort story: Sorting jumbled images and captions into stories. *arXiv preprint arXiv:1606.07493*, 2016. 2

[4] J. Andreas and D. Klein. Reasoning about pragmatics with neural listeners and speakers. *arXiv preprint arXiv:1604.00562*, 2016. 2

[5] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015. 2, 4, 6, 7

[6] J. Ba, V. Mnih, and K. Kavukcuoglu. Multiple object recognition with visual attention. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 3

[7] A. Bansal, A. Farhadi, and D. Parikh. Towards transparent systems: Semantic characterization of failure modes. In *European Conference on Computer Vision*, pages 366–381. Springer, 2014. 2, 3

[8] S. Baron-Cohen. *The evolution of a theory of mind*. na, 1999. 3

[9] S. Baron-Cohen, S. Wheelwright, J. Hill, Y. Raste, and I. Plumb. The reading the mind in the eyes test revised version: A study with normal adults, and adults with asperger syndrome or high-functioning autism. *Journal of child psychology and psychiatry*, 42(2):241–251, 2001. 2

[10] A. C. Berg, T. L. Berg, H. Daume, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos, et al. Understanding and predicting importance in images. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3562–3569. IEEE, 2012. 2

[11] J. P. Bigham, C. Jayant, H. Ji, G. Little, A. Miller, R. C. Miller, R. Miller, A. Tatarowicz, B. White, S. White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342. ACM, 2010. 2

[12] E. Broadbent, V. Kumar, X. Li, J. Sollers 3rd, R. Q. Stafford, B. A. MacDonald, and D. M. Wegner. Robots with display screens: a robot with a more humanlike face display is perceived to have more mind and a better personality. *PloS one*, 8(8):e72589, 2013. 2

[13] A. Chandrasekaran, A. K. Vijayakumar, S. Antol, M. Bansal, D. Batra, C. Lawrence Zitnick, and D. Parikh. We are humor beings: Understanding and predicting visual humor. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4603–4612, 2016. 2

[14] P. Chattopadhyay, R. Vedantam, R. S. Ramprasaath, D. Batra, and D. Parikh. Counting everyday objects in everyday scenes. *CoRR*, abs/1604.03505, 2016. 4

[15] K. Cho, A. Courville, and Y. Bengio. Describing multimedia content using attention-based encoder-decoder networks. *IEEE Transactions on Multimedia*, 17(11):1875–1886, 2015. 3

[16] H. H. Clark. Using language. 1996. 2

[17] A. Das, H. Agrawal, C. L. Zitnick, D. Parikh, and D. Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *arXiv preprint arXiv:1606.03556*, 2016. 1

[18] R. Dubey, J. Peterson, A. Khosla, M.-H. Yang, and B. Ghanem. What makes an object memorable? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1089–1097, 2015. 2, 3

[19] R. El Kaliouby and P. Robinson. Mind reading machines: Automated inference of cognitive mental states from video. In *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, volume 1, pages 682–688. IEEE, 2004. 2

[20] R. El Kaliouby and P. Robinson. Real-time inference of complex mental states from facial expressions and head gestures. In *Real-time vision for human-computer interaction*, pages 181–200. Springer, 2005. 2

[21] D. Engel, A. W. Woolley, L. X. Jing, C. F. Chabris, and T. W. Malone. Reading the mind in the eyes or reading between the lines? Theory of mind predicts collective intelligence equally well online and face-to-face. *PloS one*, 2014. 2

[22] B. Eysenbach, C. Vondrick, and A. Torralba. Who is mistaken? *arXiv preprint arXiv:1612.01175*, 2016. 2, 3

[23] F. Ferraro, N. Mostafazadeh, I. Misra, A. Agrawal, J. Devlin, R. Girshick, X. He, P. Kohli, D. Batra, C. L. Zitnick, et al. Visual storytelling. *arXiv preprint arXiv:1604.03968*, 2016. 2

[24] D. Ferrucci, A. Levas, S. Bagchi, D. Gondek, and E. T. Mueller. Watson: beyond jeopardy! *Artificial Intelligence*, 199:93–105, 2013. 2

[25] D. A. Ferrucci. Introduction to this is watson. *IBM Journal of Research and Development*, 56(3.4):1–1, 2012. 6

[26] S. R. Fussell, S. Kiesler, L. D. Setlock, and V. Yew. How people anthropomorphize robots. In *Human-Robot Interaction (HRI), 2008 3rd ACM/IEEE International Conference on*, pages 145–152. IEEE, 2008. 1, 2

[27] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are you talking to a machine? dataset and methods for multilingual image question. In *Advances in Neural Information Processing Systems*, pages 2296–2304, 2015. 2

[28] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1, 2

[29] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *arXiv preprint arXiv:1612.00837*, 2016. 4

[30] Y. Goyal, A. Mohapatra, D. Parikh, and D. Batra. Towards transparent ai systems: Interpreting visual question answering models. *arXiv preprint arXiv:1608.08974*, 2016. 3

[31] B. Grosz. What question would turing pose today? *AI Magazine*, 2012. 2

[32] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 2

[33] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell. Generating visual explanations. In *European Conference on Computer Vision*, pages 3–19. Springer, 2016. 3

[34] M. Hutson. *The 7 laws of magical thinking: How irrational beliefs keep us happy, healthy, and sane*. Penguin, 2012. 2

[35] J. Joseph and L. Friedman. IBMs Watson Helps Employees Tackle Cancer . https://bestdoctors.com/blog/2017/01/10/press-release/, 2017. [Online; accessed 17-March-2008]. 2

[36] S. Kazemzadeh, V. Ordonez, M. Matten, and T. L. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, pages 787–798, 2014. 2

[37] J.-H. Kim, K.-W. On, J. Kim, J.-W. Ha, and B.-T. Zhang. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*, 2016. 4

[38] C. Krupenye, F. Kano, S. Hirata, J. Call, and M. Tomasello. Great apes anticipate that other individuals will act according to false beliefs. *Science*, 354(6308):110–114, 2016. 1

[39] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37. Springer, 2016. 2

[40] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297, 2016. 3, 4, 9

[41] H. MacLeod, C. L. Bennett, M. R. Morris, and E. Cutrell. Understanding blind peoples experiences with computer-generated captions of social media images. 3

[42] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in Neural Information Processing Systems*, pages 1682–1690, 2014. 2

[43] Microsoft. Microsoft Cognitive Services Computer Vision API. https://www.microsoft.com/cognitive-services/en-us/computer-vision-api, 2017. [Online; accessed 17-March-2008]. 2

[44] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013. 9

[45] M. Minsky. *Society of mind*. Simon and Schuster, 1988. 2

[46] J. P. Mitchell. Inferences about mental states. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 364(1521):1309–1316, 2009. 1

[47] V. Mnih, N. Heess, A. Graves, et al. Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212, 2014. 3

[48] B. Mutlu, J. Forlizzi, and J. Hodgins. A storytelling robot: Modeling and evaluation of human-like gaze behavior. In *Humanoid robots, 2006 6th IEEE-RAS international conference on*, pages 518–523. IEEE, 2006. 2

[49] N. M. Oliver, B. Rosario, and A. P. Pentland. A bayesian computer vision system for modeling human interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000. 2, 3

[50] D. Parikh and K. Grauman. Implied feedback: Learning nuances of user behavior in image search. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 745–752, 2013. 2

[51] D. H. Park, L. A. Hendricks, Z. Akata, B. Schiele, T. Darrell, and M. Rohrbach. Attentive explanations: Justifying decisions and pointing to the evidence. *arXiv preprint arXiv:1612.04757*, 2016. 3

[52] H. R. Pelikan and M. Broth. Why that nao?: How humans adapt to a conventional humanoid robot in taking turns-at-talk. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4921–4932. ACM, 2016. 3

[53] D. Premack and G. Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(04):515–526, 1978. 1

[54] A. Ray, G. Christie, M. Bansal, D. Batra, and D. Parikh. Question relevance in vqa: Identifying non-visual and false-premise questions. *arXiv preprint arXiv:1606.06622*, 2016. 2, 7

[55] A. Recasens, A. Khosla, C. Vondrick, and A. Torralba. Where are they looking? In *Advances in Neural Information Processing Systems*, pages 199–207, 2015. 2, 3

[56] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016. 2

[57] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 2

[58] M. T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016. 3

[59] T. Savvy. Watson will see you now: a supercomputer to help clinicians make informed treatment decisions. 2015. 2
[60] B. Scassellati. Theory of mind for a humanoid robot. *Autonomous Robots*, 12(1):13–24, 2002. 3
[61] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *arXiv preprint arXiv:1610.02391*, 2016. 3, 9
[62] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016. 6
[63] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4
[64] H. Song, A.-D. Nguyen, M. Gong, and S. Lee. A review of computer vision methods for purpose on computer-aided diagnosis. *Journal of International Society for Simulation Surgery*, 3(1):1–8, 2016. 2
[65] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv preprint arXiv:1602.07261*, 2016. 2
[66] K. Tu, M. Meng, M. W. Lee, T. E. Choe, and S.-C. Zhu. Joint video and text parsing for understanding events and answering queries. *IEEE MultiMedia*, 21(2):42–70, 2014. 2
[67] C. Vondrick, D. Oktay, H. Pirsiavash, and A. Torralba. Predicting motivations of actions by leveraging text. In *Proceedings of the IEEE Conference on Comslputer Vision and Pattern Recognition*, pages 2997–3005, 2016. 2, 3
[68] C. Vondrick, H. Pirsiavash, and A. Torralba. Anticipating visual representations from unlabeled video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 98–106, 2016. 2, 3
[69] S. I. Wang, P. Liang, and C. D. Manning. Learning language games through interaction. *arXiv preprint arXiv:1606.02447*, 2016. 3
[70] M. Wayland. Nissan self-driving system teams AI with human advisers. http://www.detroitnews.com/story/business/autos/foreign/2017/01/05/nissan-sam/96224020/, 2017. [Online; accessed 17-March-2008]. 2
[71] H. Wimmer and J. Perner. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128, 1983. 3
[72] A. W. Woolley, C. F. Chabris, A. Pentland, N. Hashmi, and T. W. Malone. Evidence for a collective intelligence factor in the performance of human groups. *science*, 2010. 2
[73] S. Wu, J. Wieland, O. Farivar, and J. Schiller. Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1180–1192. ACM, 2017. 2
[74] H. Xu and K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision*, pages 451–466. Springer, 2016. 3
[75] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, volume 14, pages 77–81, 2015. 3
[76] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–29, 2016. 3
[77] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 3
[78] P. Zhang, J. Wang, A. Farhadi, M. Hebert, and D. Parikh. Predicting failures of vision systems. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3566–3573, 2014. 2, 3, 9
[79] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*, 2014. 3
[80] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016. 3

Appendix

This appendix is organized as follows: We first provide access to the interfaces we used to train subjects to provide a sense for both tasks, and briefly summarize our code-release plan. Next, we provide more qualitative examples of montages (introduced in Fig.2 of main paper) that highlight Vicki's quirks, and additionally share insights on Vicki from subjects who completed the tasks. Finally, we share details and analysis of the survey we conducted on AMT.

A. Interfaces

We will make the interfaces used for our human studies publicly available, to enable others to both evaluate a person's Theory of AI's Mind and further investigate the effectiveness of instant feedback and explanation modalities in improving the same. To enable readers to experience the FP and KP tasks firsthand, we provide a link to the interfaces we used to train subjects: https://deshraj.github.io/TOAIM/. We also provide links to videos demonstrating each task: FP – https://youtu.be/Dcs7GOmTAns and KP – https://youtu.be/f_1ikwCuG4Q. Note that for illustration, we show just a single setting of the respective task in each interface and video.

B. Vicki's Quirks

We present more examples in Fig. 8 and Fig. 9 that highlight Vicki's quirks. Recall that there are several factors which lead to Vicki being quirky, many of which are well known in VQA literature [2]. As we can see across the examples in Fig. 8 and 9, Vicki exhibits these quirks in a somewhat predictable fashion. At first glance, the primary factors that seem to decide Vicki's response to a question given an image are the properties and activities associated with the salient objects in the image, in combination with the language and the phrasing of the question being asked. This is evident when we look across the images (see Fig. 8 and 9) for question-answer (QA) pairs such as – *What are the people doing? Grazing*, *What is the man holding? Cow* and *Is it raining? No*. As a specific example, notice the images for the QA pair *What color is the grass? Blue* (see Fig. 8) – Vicki's response to this question is the most dominant color in the scene across all images even though there is no grass present in any of them. Similarly, for the QA pair *What does the sign say? Banana* (see Fig. 9) – Vicki's answer is the salient object across all the scenes.

Interestingly, some subjects did try and pick up on some of the quirks and beliefs described previously, and formed a mental model of Vicki while completing the Failure Prediction or Knowledge Prediction tasks. We asked subjects to leave comments after completing a task and some of them shared their views on Vicki's behavior. We share some of those comments below.
The abbreviations used are Failure Prediction (FP), Knowledge Prediction (KP) and Instant Feedback (IF).

1. **FP**
   - *"These images were all pretty easy to see what animal it was. I would imagine the robot would be able to get 90% of the animals correct, unless there were multiple animals in the same photo."*
   - *"I think the brighter the color the more likely they are to get it right. Multi-colored, not so sure."*
   - *"I'd love to know the answers to these myself."*
2. **FP + IF**
   - *"This is fun, but kind of hard to tell what the hints mean. Can she determine the color differences in multi-colored umbrellas or are they automatically marked wrong because she only chooses one color instead of all of the colors? It seems to me that she just goes for the brightest color in the pic. This is very interesting. Thank you! :)"*
   - *"I didn't quite grasp what the AI's algorithm was for determining right or wrong. I want to say that it was if the AI could see the face of the animal then it guessed correctly, but I'm really not sure."*
3. **FP + IF + Explanation Modalities**
   - *"Even though Vicki is looking at the right spot doesn't always mean she will guess correctly. To me there was no rhyme or reason to guessing correctly. Thank you."*
   - *"I think she can accurately know a small number of people but cannot know a huge grouping yet."*
   - *"I would be more interested to find out how Vickis metrics work. What I was assuming is just color phase and distance might not be accurate."*
4. **KP**
   - *"Time questions are tricky because all Vicki can do is round to the nearest number."*
   - *"there were a few that seemed like it was missing obvious answers - like bus and bus stop but not bus station. Also words like lobby seemed to be missing."*
5. **KP + IF**
   - *"Interesting, though it seems Vicki has a lot more learning to do. Thank you!"*
   - *"This HIT was interesting, but a bit hard. Thank you for the opportunity to work this."*
6. **KP + IF + Explanation Modalities**
   - *"You need to eliminate the nuances of night time and daytime from the computer and choose one phrasing "night" or "day" Vicki understands. The nuance keeps me and I'm sure others obtaining a higher score here on this task."*
   - *"I felt that Vickie was mistaken as to what some colors were for the first test which probably carried over and I tried my best to recreate her responses."*
7. **KP + IF + Montages**
   - *"I am not sure that I ever completely understood how Vicki thought. It seemed it had more to do with what was in the pictures instead of the time of day it looked in the pictures. If there was food, she chose noon or morning, even though at times it was clearly breakfast food and she labeled it noon."*
   - *"It doesn't seem very accurate as I made sure to count and took my time assessing the pictures."*
   - *"it is hard to figure out what they are looking for since there isn't many umbrellas in the pictures"*

On a high-level reading through all comments, we found that subjects felt that Vicki's response often revolves around the most salient object in the image, that Vicki is bad at counting, and that Vicki often responds with the most dominant color in the image when asked a color question. In Fig. 10 we show a word cloud of all the comments left by the subjects after completing the tasks. From the comments, we observed that subjects were very enthusiastic to familiarize themselves with Vicki, and found the process engaging. Many thought that the scenarios presented to them were *interesting* and *fun*, despite being *hard*. We used some basic elements of gamification, such as performance–based reward and narrative, to make our tasks more engaging; we think the positive response indicates the possibility of making such human–familiarization with AI engaging even in real–world settings.

C. Survey Questions

In this section we describe the survey that subjects were asked to answer before carrying out the FP and KP tasks described in Sections 6.1 and 6.2 of the main paper. These questions attempt to assess the subjects' general impressions of present-day AI, and can be broken down into 3 categories – Population Demographics, Exposure to AI, and Perception of AI.

As part of the survey, subjects were asked a few subjective questions about their opinions on present–day AI's capabilities. Specifically, they were asked to list tasks that they thought AI is capable of performing *today* (see Fig. 11), will be capable of *in the next 3 years* (see Fig. 12), and will be capable of *in the next 10 years* (see Fig. 13).

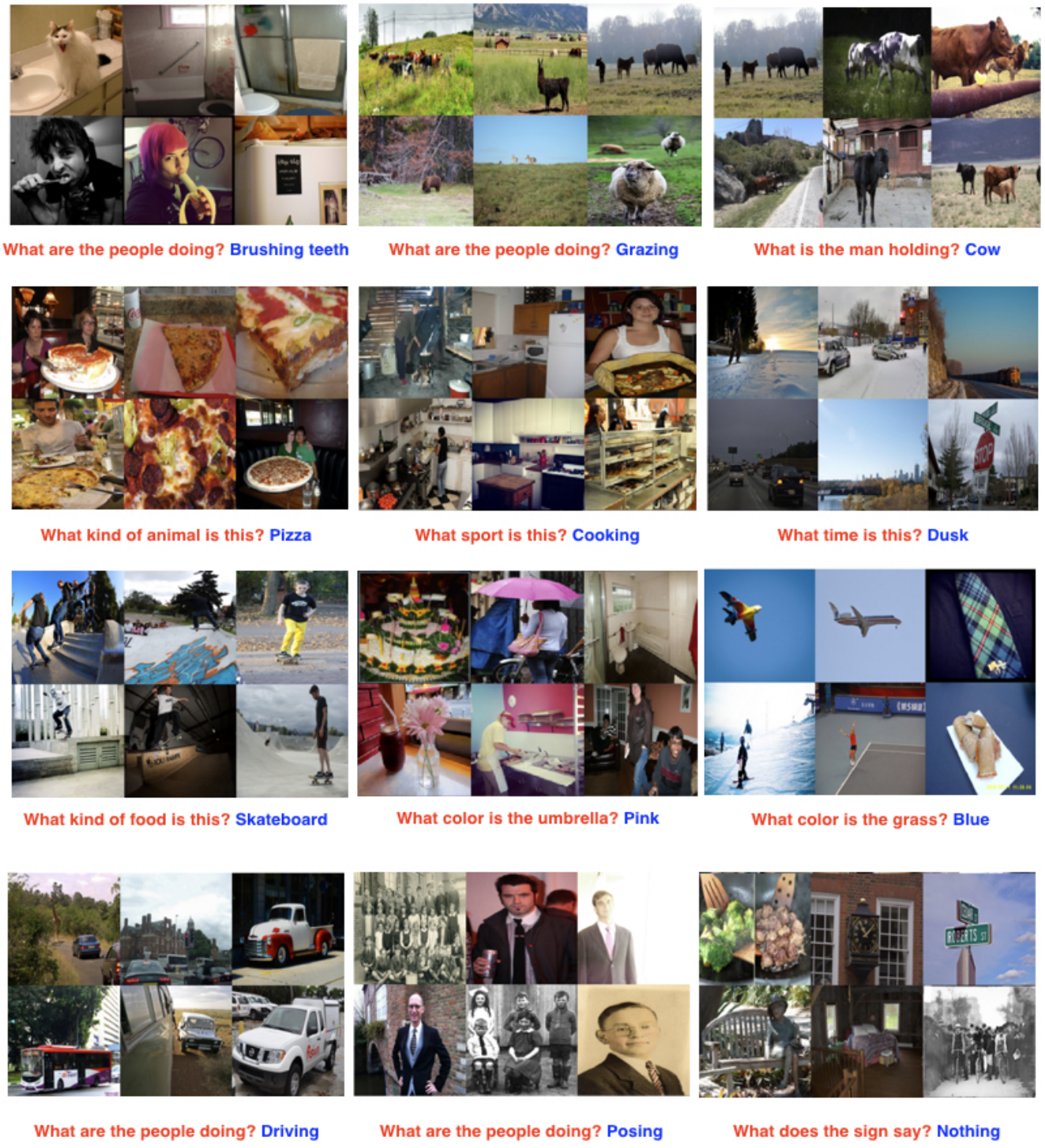


Figure 8: Given a question (red) we show images for which Vicki gave the same answer (blue) to the question to observe Vicki's quirks.

Figure 9: Given a question (red) we show images for which Vicki gave the same answer (blue) to the question to observe Vicki's quirks.

Figure 10: We show a word cloud of all the comments left by subjects after completing the tasks across all settings. From the frequency of positive comments about the tasks, it appears that subjects were enthusiastic to familiarize themselves with Vicki.

We also asked how *they* think AI works (see Fig. 14). In Fig. 11, 12 and 13, we show word clouds corresponding to what subjects thought about the capabilities of AI. We also share some of those responses below.

1. Name three things that you think AI today can do.

Predict sports games; Detect specific types of cancer in images; Control house temp based on outside weather; translate; calculate probabilities; Predictive Analysis; AI can predict future events that happen like potential car accidents; lip reading; code; Facial recognition; Drive cars; Play Go; predict the weather; Hold a conversation; Be a personal assistant; Speech recognition; search the web quicker.

2. Name three things that you think AI today can't yet do but will be able to do in 3 years.

Fly planes; Judge emotion in voices; Predict what I want for dinner; perform surgery; drive cars; manage larger amounts of information at a faster rate; think independently totally; play baseball; drive semi trucks; Be a caregiver; anticipate a person's lying ability; read minds; Diagnose patients; improve robots to walk straight; Run websites; solve complex problems like climate change issues; program other ai; guess ages; form conclusions based on evidence; act on more complex commands; create art.

3. Name three things that you think AI today can't yet do and will take a while (> 10 years) before it can do it.

Imitate humans; be indistinguishable from humans; read minds; Have emotions; Develop feelings; make robots act like humans; truly learn and think; Replace humans; impersonate people; teach; be a human; full AI with personalities; Run governments; be able to match a human entirely; take over the world; Pass a turing test; be a human like friend; intimacy; Recognize things like sarcasm and humor.

Interestingly, we observe a steady progression in subjects' expectations of AI's capabilities, as the time span increases. On a high-level reading through the responses, we notice that subjects believe that AI today can successfully

Figure 11: A word cloud of subject responses to "Name three things that you think AI today can do."

Figure 12: A word cloud of subject responses to "Name three things that you think AI today can't yet do but will be able to do in 3 years."

Figure 13: A word cloud of subject responses to "Name three things that you think AI today can't yet do and will take a while (> 10 years) before it can do it."

perform tasks such as *machine translation*, *driving vehicles*, *speech recognition*, *analyzing information and drawing conclusions*, etc. (see Fig. 11). It is likely that this is influenced by the subjects' exposure to or interaction with some form of AI in their day-to-day lives. When asked about what AI can do three years from now, most subjects suggested more sophisticated tasks such as *inferring emotions from voice tone*, *performing surgery*, and even *dealing with climate change issues* (see Fig. 12). However, the most interesting trends emerge while observing subjects' expectation of what AI can achieve in the next 10 years

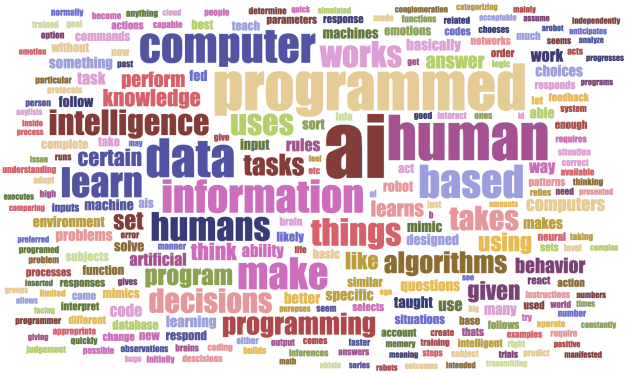


Figure 14: A word cloud of subject responses when asked to describe how *they* think AI today works.

(see Fig. 13). A major proportion of subjects believe that AI will gain the ability to *understand and emulate human beings*, *teach human beings*, *develop feelings and emotions* and *pass the Turing test*.

We also observe how subjects think AI works (see Fig. 14). Mostly, subjects believe that an AI agent today is a system with high computational capabilities that has been programmed to simulate intelligence and perform certain tasks by exposing it to huge amounts of information, or, as one of subjects phrased it – *broadly AI recognizes patterns and creates optimal actions based on those patterns towards some predefined goals*. In summary, it appears that subjects have high expectations from AI, given enough time. While it is uncertain at this stage how many, or how soon, these feats will actually be achieved, we believe that building Theory of AI's mind skills will help humans generally become more active and effective collaborators in human–AI teams.

We now provide a full list of all questions in the survey. In Fig. 15, 16 and 17, we also break down the 321 subjects that completed the survey by their response to each question.

1. How old are you?
   (a) Less than 20 years
   (b) Between 20 and 40 years
   (c) Between 40 and 60 years
   (d) Greater than 60 years
2. What is your gender?
   (a) Male
   (b) Female
   (c) Other
3. Where do you live?
   (a) Rural
   (b) Suburban
   (c) Urban
4. Are you?
   (a) A student
   (b) Employed
   (c) Self-employed
   (d) Unemployed
   (e) Retired
   (f) Other
5. To which income group do you belong?
   (a) Less than 5000$ per year
   (b) 5,000-10,000$ per year
   (c) 10,000-25,000$ per year
   (d) 25,000-60,000$ per year
   (e) 60,000-120,000$ per year
   (f) More than 120,000$ per year
6. What is your highest level of education?
   (a) No formal education
   (b) Middle School
   (c) High School
   (d) College (Bachelors)
   (e) Advanced Degree
7. What was your major?
   (a) Computer Science / Computer Engineering
   (b) Engineering but not Computer Science
   (c) Mathematics / Physics
   (d) Philosophy
   (e) Biology / Physiology / Neurosciences
   (f) Psychology / Cognitive Sciences
   (g) Other Sciences
   (h) Liberal Arts
   (i) Other
   (j) None
8. Do you know how to program/code?
   (a) Yes
   (b) No
9. Does your full-time job involve:
   (a) No computers
   (b) Working with computers but no programming/coding
   (c) Programming/Coding
10. How many hours a day do you spend on your computer/laptop/smartphone?
   (a) Less than 1 hour
   (b) 1-5 hours
   (c) 5-10 hours
   (d) Above 10 hours
11. Do you know what Watson is in the context of Jeopardy?
   (a) Yes
   (b) No
12. Have you ever used Siri, Alexa, or Google Now/Google Assistant?
   (a) Yes
   (b) No
13. How often do you use Siri, Alexa, Google Now, Google Assistant, or something equivalent?
   (a) About once every few months
   (b) About once a month
   (c) About once a week

(d) About 1-3 times a day
(e) More than 3 times a day

14. Have you heard of AlphaGo?
(a) Yes
(b) No

15. Have you heard of Machine Learning?
(a) Yes
(b) No

16. Have you heard of Deep Learning?
(a) Yes
(b) No

17. When did you first hear of Artificial Intelligence (AI)?
(a) I have not heard of AI
(b) More than 10 years ago
(c) 5-10 years ago
(d) 3-5 years ago
(e) 1-3 years ago
(f) In the last six months
(g) Last month

18. How did you learn about AI?
(a) School / College
(b) Conversation with people
(c) Movies
(d) Newspapers
(e) Social media
(f) Internet
(g) TV
(h) Other

19. Do you think AI today can drive cars fully autonomously?
(a) Yes
(b) No

20. Do you think AI today can automatically recognize faces in a photo?
(a) Yes
(b) No

21. Do you think AI today can read your mind?
(a) Yes
(b) No

22. Do you think AI today can automatically read your handwriting?
(a) Yes
(b) No

23. Do you think AI today can write poems, compose music, make paintings?
(a) Yes
(b) No

24. Do you think AI today can read your Tweets, Facebook posts, etc. and figure out if you are having a good day or not?
(a) Yes
(b) No

25. Do you think AI today can take a photo and automatically describe it in a sentence?
(a) Yes
(b) No

26. Other than those mentioned above, name three things that you think AI today can do.

27. Other than those mentioned above, name three things that you think AI today can't yet do but will be able to do in 3 years.

28. Other than those mentioned above, name three things that you think AI today can't yet do and will take a while ($>$ 10 years) before it can do it.

29. Do you have a sense of how AI works?
(a) Yes
(b) No
(c) If yes, describe in a sentence or two how AI works.

30. Would you trust an AI's decisions today?
(a) Yes
(b) No

31. Do you think AI can ever become smarter than the smartest human?
(a) Yes
(b) No

32. If yes, in how many years?
(a) Within the next 10 years
(b) Within the next 25 years
(c) Within the next 50 years
(d) Within the next 100 years
(e) In more than 100 years

33. Are you scared about the consequences of AI?
(a) Yes
(b) No
(c) Other
(d) If other, explain.

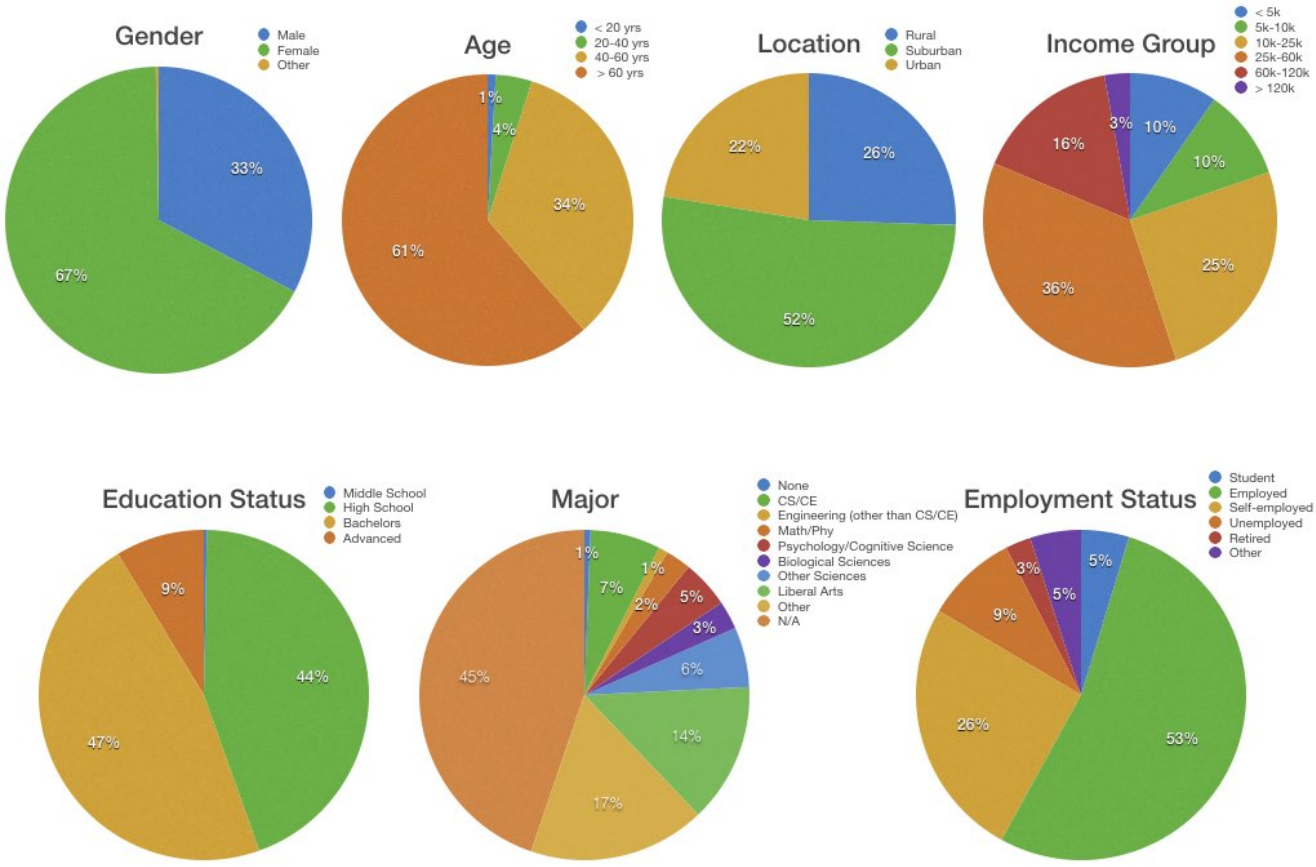


Figure 15: Population Demographics (across 321 subjects)

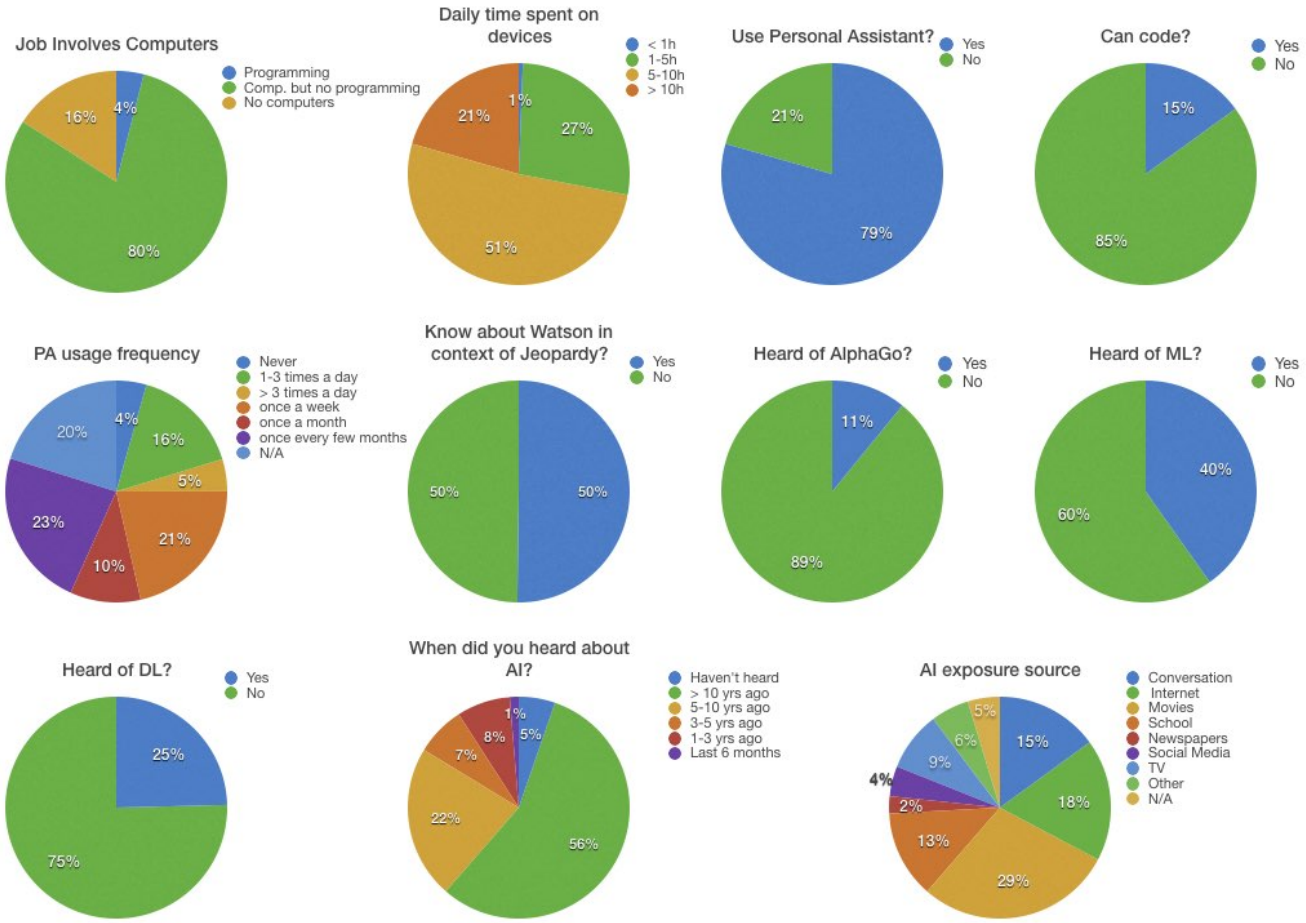


Figure 16: Technology and AI exposure (across 321 subjects)

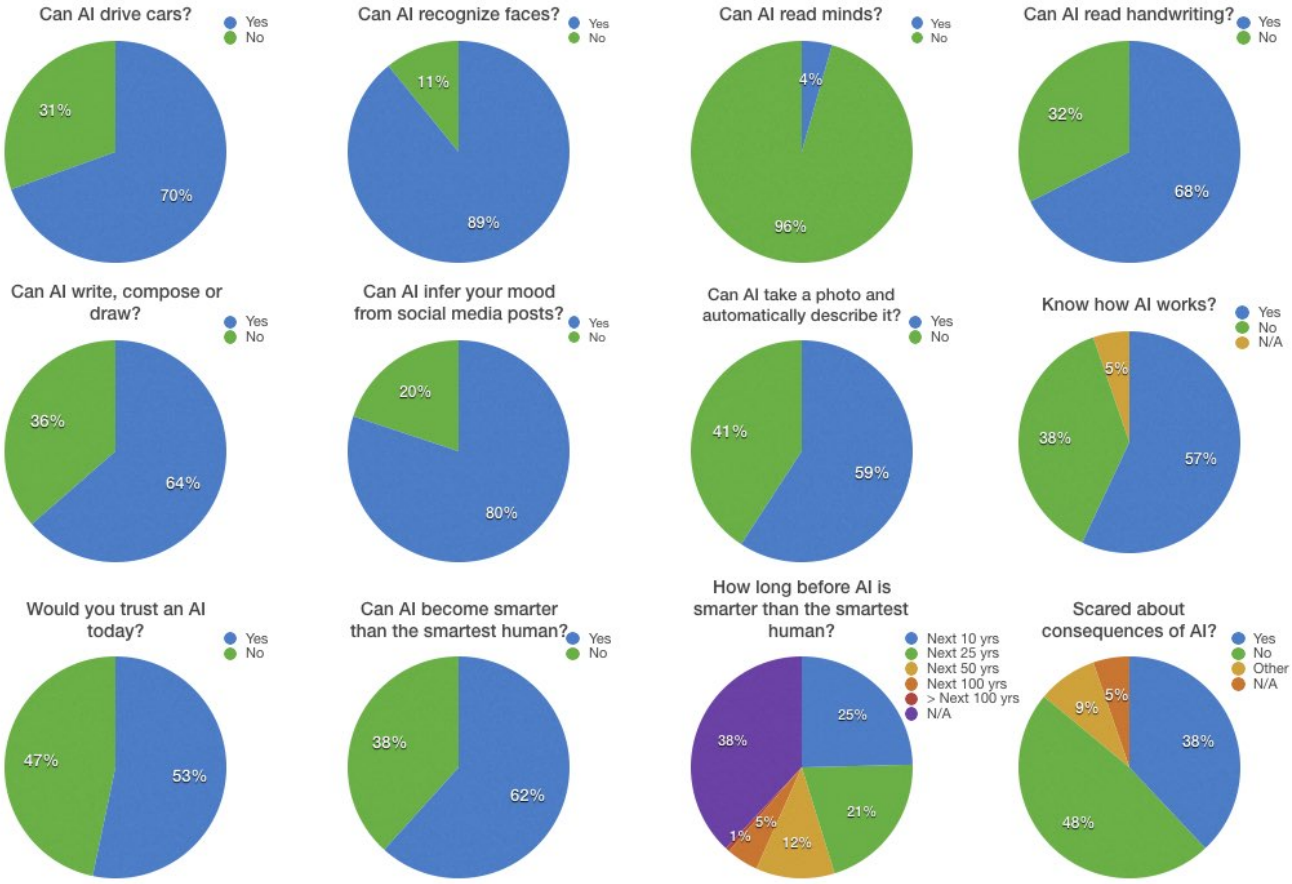


Figure 17: Perception of AI (across 321 subjects)