Do explanations make VQA models more predictable to a human? Supplementary Material

Arjun Chandrasekaran[*,1] **Viraj Prabhu**[*,1] **Deshraj Yadav** [*,1]
Prithvijit Chattopadhyay[*,1] **Devi Parikh**[1,2]
[1]Georgia Institute of Technology [2]Facebook AI Research
{carjun, virajp, deshraj, prithvijit3, parikh}@gatech.edu

1 Introduction

This supplementary material is organized as follows. We first discuss various visual recognition scenarios in which a human might rely on an AI, and motivate the need for building a model of the AI in such scenarios. Following this, we discuss the tasks (see Tasks, Sec. 3 of the main paper) – Failure Prediction (FP) and Knowledge Prediction (KP) in more detail and provide video demonstrations of the AMT interfaces associated with the same. Next, we describe an in-house Failure Prediction study we conducted with Computer Vision researchers as subjects. We then provide a few more qualitative examples of montages that highlight the quirks (see Agent, Sec. 3 of the main paper) which make Vicki predictable, and additionally share insights on Vicki from subjects who completed the tasks, in Sec. 5. Finally, we describe an AMT survey we conducted to gauge public perception of AI, and provide a list of questions and qualitative analyses of results.

2 Visual Recognition Scenarios

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. For instance, if the bird is occluded, or the scene is cluttered, or the lighting is bad, or the bird pose is odd, Vicki will likely fail. Moreover, the idea of humans predicting the AI's failure 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.

[*]Denotes equal contribution.

3 Tasks and Interfaces

Our proposed tasks of FP and KP are designed to measure a human's understanding of the capabilities of an AI agent such as Vicki. As mentioned before, the 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 the human can estimate the behavior of an AI in a specific scenario.

Failure Prediction. Recall that in FP, 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 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. 3(a) of the main paper. Note that we do not show the human what Vicki's predicted answer is.

Knowledge Prediction. In KP, 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. We provide subjects a convenient dropdown interface with autocomplete to choose an answer

from Vicki's vocabulary of 1000 answers. Please see a snapshot of the KP interface in Fig. 3(b) of the main paper.

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, 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 and question that the subject sees) and translate it to what Vicki might say. High performance at KP is likely to correlate with high performance at the reverse task – i.e., to reconstruct the input image based on Vicki's prediction. 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 humans ability to approximate a neural network's behavior!

We used different variants of the base interfaces for both Failure Prediction and Knowledge Prediction tasks on Amazon Mechanical Turk (AMT). These variants are characterized by the presence/absence of different explanation modalities used in train or test time.

The interfaces we used to train subjects are available at `https://deshraj.github.io/TOAIM/`. To enable the readers to experience the FP and KP tasks firsthand, we also include videos demonstrating each task with this supplementary document. Note that for illustration, we provide videos for only one setting, for each of the FP and KP tasks.

4 FP with VQA and Vision Researchers

Just as an anecdotal point of reference, we also conducted experiments across experts with varying degrees of familiarity with agents like Vicki. We observed that a VQA researcher achieved an accuracy of 80% versus a computer vision (but not VQA) researcher who achieved 60% in a shorter version of the FP task without instant feedback. Thus, familiarity with Vicki might play a role in how well a human can predict its oncoming failures or successes.

5 Vicki's Quirks

We present some additional examples in Fig. 1 and Fig. 2 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 (Agrawal et al., 2016). As we can see across both examples, 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. 1 and 2) 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. 1) – 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. 2) – 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 um-*

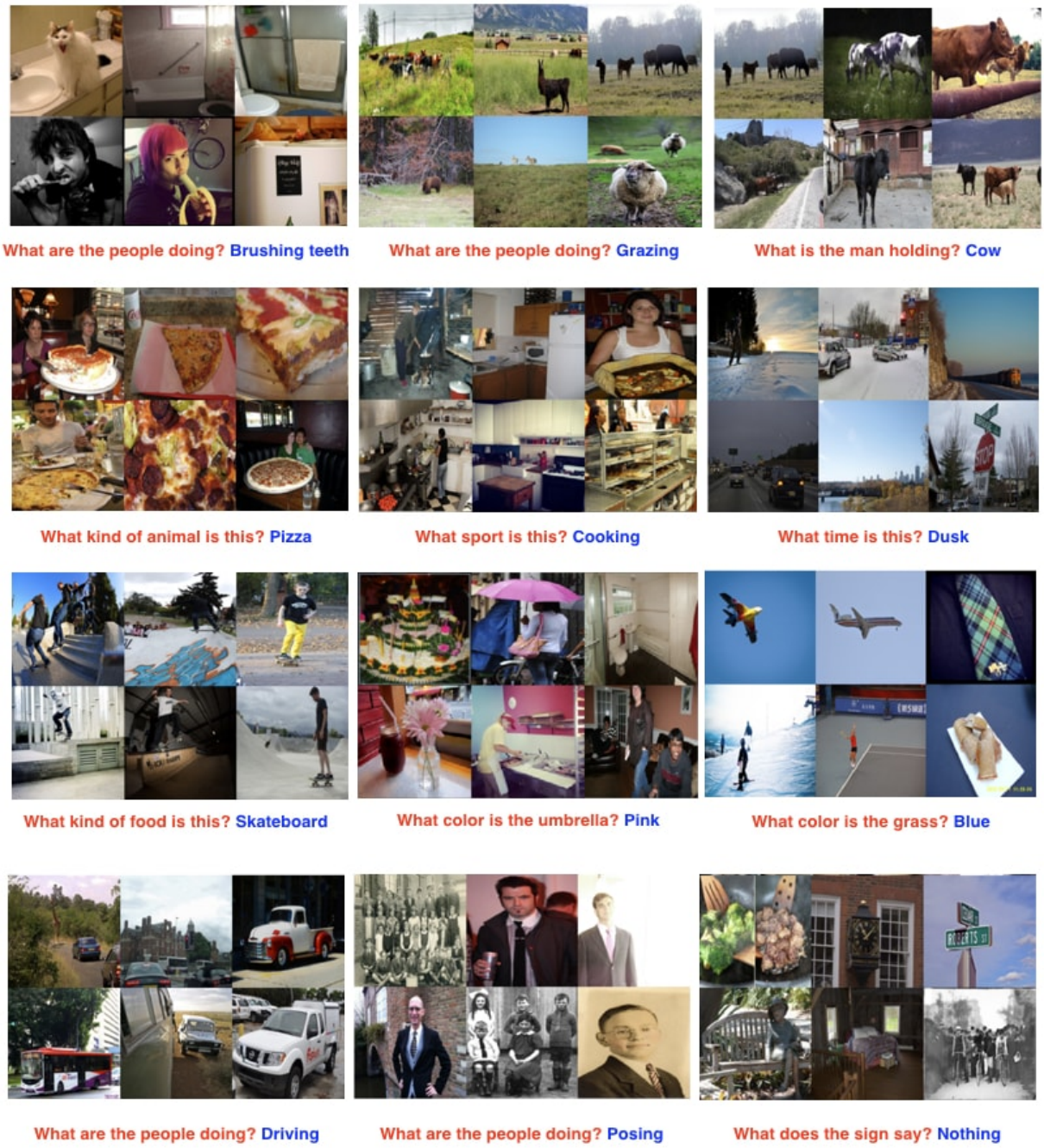


Figure 1: Given a question (red) we show images for which Vicki gave the same answer (blue) to the question to observe Vicki's quirks.

brellas 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*

Figure 2: Given a question (red) we show images for which Vicki gave the same answer (blue) to the question to observe Vicki's quirks.

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.*

Figure 3: 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 the subjects who left comments were enthusiastic to familiarize themselves with Vicki.

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*

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. 3, 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.

6 Perception of AI

In addition to measuring the subjects' capabilities to predict Vicki's behavior, we also conducted a survey to assess their general impressions of present-day AI. Specifically, we asked them to fill out a survey with questions focusing around three types of information - "Background Information", "Familiarity with Computers and AI" and "Estimates of AI's capabilities".

In Fig. 5, 6 and 7, we break down the 321 subjects that completed the survey by their response to each question.

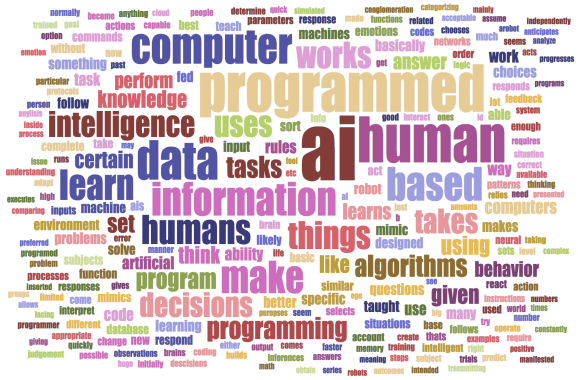
As part of the survey, subjects were also asked a few subjective questions about their opinions on present–day AI's capabilities. These include multiple-choice questions focusing on some specific capabilities of AI ("Can AI recognize faces?", "Can AI drive cars?", etc.) – responses to which are summarized in Fig. 7. The subjects were also asked to specifically list tasks that they thought AI is capable of performing *today* (see Fig. 4a), will be capable of *in the next 3 years* (see Fig. 4b), and will be capable of *in the next 10 years* (see Fig. 4c). We also asked how *they* think AI works (see Fig. 4d). In Fig. 4a, 4b and 4c, 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;*

(a) A word cloud of subject responses to "Name three things that you think AI today can do."

(b) 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."

(c) 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."



(d) A word cloud of subject responses when asked to describe how *they* think AI today works.

Figure 4: Word clouds corresponding to responses from humans for different questions.

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 perform tasks such as *machine translation*, *driving vehicles*, *speech recognition*, *analyzing information and drawing conclusions*, etc. (see Fig. 4a). 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. 4b). However, the most interesting

trends emerge while observing subjects' expectation of what AI can achieve in the next 10 years (see Fig. 4c). 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 ask subjects how they think AI works (see Fig. 4d). One of the subjects phrases it as – *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 a model of the AI's skillset will help humans generally become more active and effective collaborators in human–AI teams.

We now provide a full list of questions the subjects were asked in the survey.

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

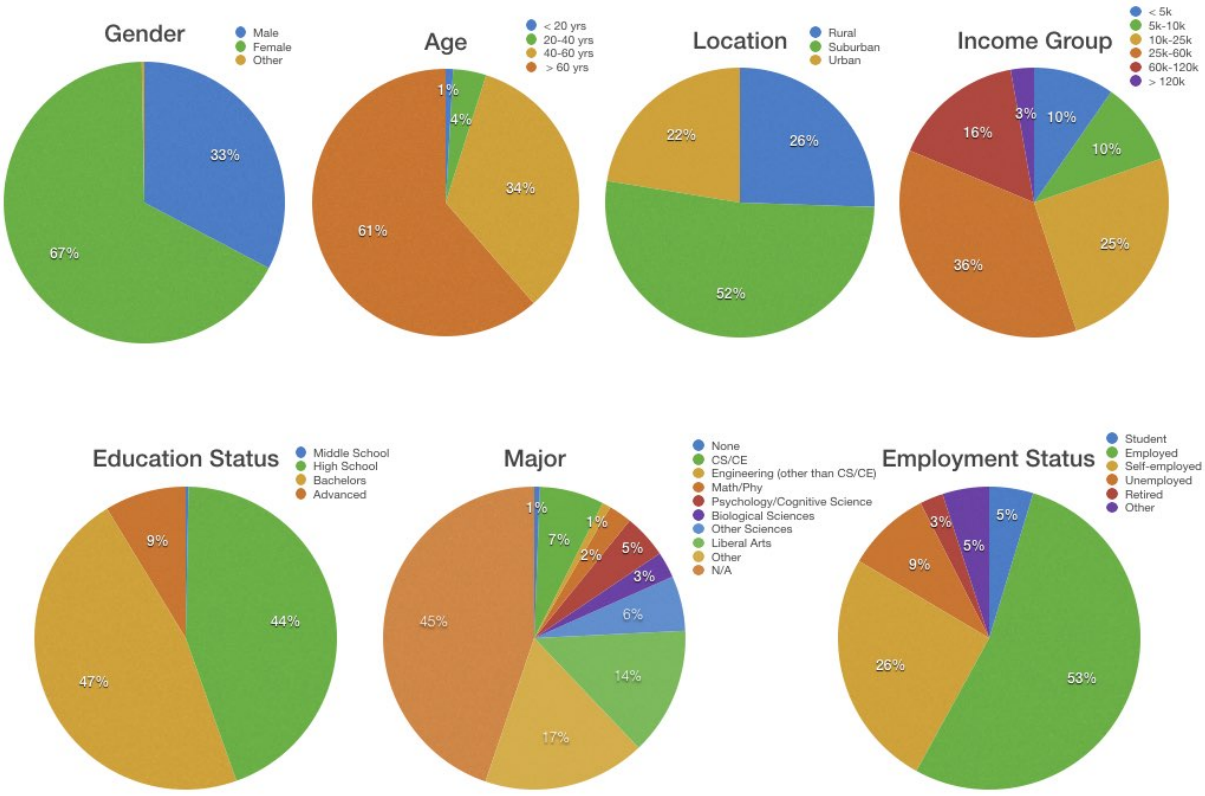


Figure 5: Population Demographics (across 321 subjects)

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

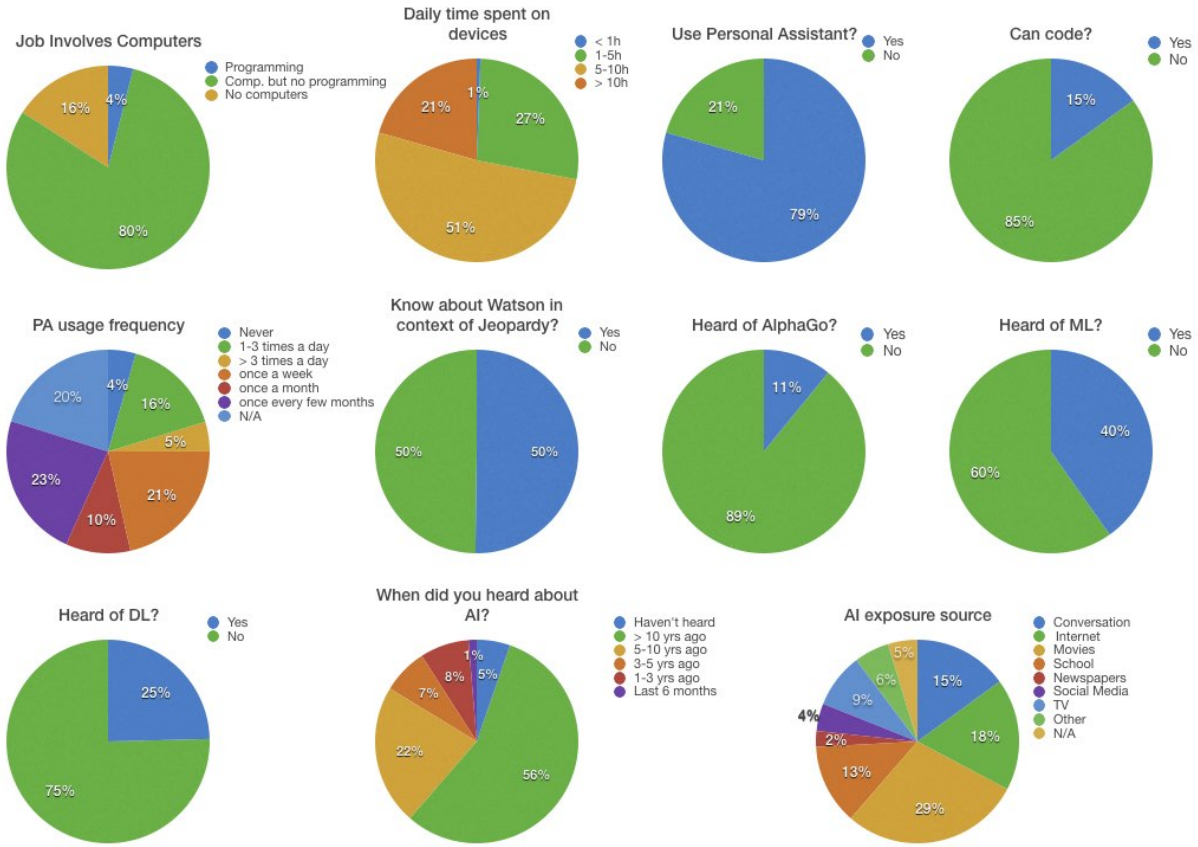


Figure 6: Technology and AI exposure (across 321 subjects)

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.

References

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. *arXiv preprint arXiv:1606.07356*.

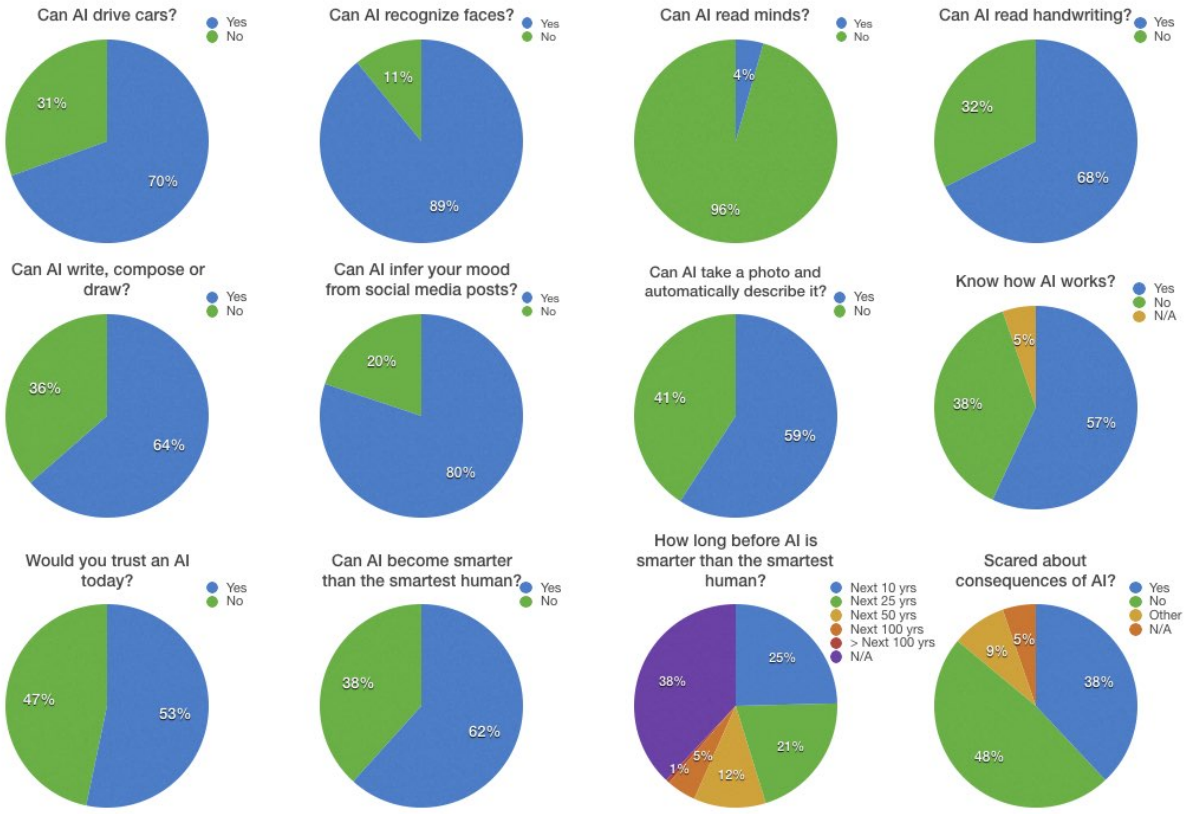


Figure 7: Perception of AI (across 321 subjects)