Asking goal-oriented questions and learning from answers

Anselm Rothe[1], **Brenden M. Lake**[1,2], **and Todd M. Gureckis**[1]
[1]Department of Psychology, [2]Center for Data Science, New York University

Abstract

The study of question asking in humans and machines has gained attention in recent years. A key aspect of question asking is the ability to select good (informative) questions from a provided set. Machines—in particular neural networks—generally struggle with two important aspects of question asking, namely to learn from the answer to their selected question and to flexibly adjust their questioning to new goals. In the present paper, we show that people are sensitive to both of these aspects and describe a unified Bayesian account of question asking that is capable of similar ingenuity. In the first experiment, we predict people's judgments when adjusting their question-asking towards a particular goal. In the second experiment, we predict people's judgments when deciding what follow-up question to ask. An alternative model based on superficial features, such as the existence of certain key words in the questions, was not able to capture these judgments to a reasonable degree.

Keywords: Bayesian modeling; active learning; information search; question asking

Introduction

The ability to ask questions is a core quality of human cognition. By asking questions, we can actively seek out information that helps us learn about the world and achieve our goals. Skilled question asking involves the ability to adjust questions towards a particular goal as well as a sensitivity to the context, including what was previously asked.

In contrast, machines have difficulty capturing these aspects of human inquiry. Recent work with neural networks has made progress on generating sensible questions about images, such as "What caused this accident?" for an image displaying a crashed motorbike lying on the street (Mostafazadeh et al., 2016; Jain & Schwing, 2017), or about passages of text (Du, Shao, & Cardie, 2017). Such questions can initiate a conversation between human and computer, however these networks are not able to make sense of any answer they might get to their question. As an intermediate solution, neural networks have been trained to predict the answer to their own questions (Johnson et al., 2017). Another ambitious approach has been to train neural networks end-to-end on entire sequences of questions and answers (Lee, Heo, & Zhang, 2018; Strub et al., 2017). However, the networks still learn a fixed question asking strategy and cannot adapt to new goals that were not included in the training regime.

Unlike neural network approaches, people can flexibly adapt their questions based on their goals and the answers they have received. Previous work has looked at how people ask questions based on specific goals (e.g., Graesser, Langston, & Bagget, 1993), or ask follow-up questions (e.g., Nelson, Divjak, Gudmundsdottir, Martignon, & Meder, 2014; Ruggeri, Lombrozo, Griffiths, & Xu, 2016), but little modeling work has been done to test these aspects directly in naturalistic tasks. Here, we study an intuitive question asking task amenable to formal modeling. By systematically manipulating core components of question asking such as goals and previously asked questions, we can compare people's behavior to an ideal observer in a more naturalistic question asking environment. For this purpose, we extend the computational framework by Rothe, Lake, and Gureckis (2018) to handle these facets of flexible question asking.

In the next section, we will introduce the question asking environment, followed by the computational framework and its extensions. We then report two experiments, in which we test people's ability to identify question quality under changing goals (Experiment 1) and after being provided with answers to previous questions (Experiment 2). Finally, alternative models are discussed.

Battleship game environment

We adopt the Battleship task used by Rothe et al. because it enables intuitive question asking for people while still being amenable to formal modeling. In the Battleship task, participants try to discover geometric shapes (i.e., battleships) on a grid (i.e., game board). These ships have varying shapes, colors, and locations (Figure 1). In our setting, there were always exactly three ships on a 6x6 board and each ship got a unique color from the set {blue, red, purple}. Each ship is a rectangle with a width of 1 and a length sampled from the set {2, 3, 4} and its orientation is sampled from the set {horizontal, vertical}. Each ship is randomly placed on the grid, ensuring they do not overlap.

In our experiments, participants face a partly revealed game board, together with a set of natural-language questions that could reveal more information about the board. Participants rank order these questions by quality taking either a particular goal or an already-answered question into account (Figure 2).

Modeling

We develop a Bayesian ideal-observer model of the task, as used in prior work, and discuss extensions to handle goals and previously answered questions.

Bayesian-ideal observer model

What does the hidden game board look like? The player begins with maximal uncertainty about the game board, modeled as a uniform prior belief distribution $p(h)$ over all possible game boards. Then, the player updates this prior via Bayes rule based on the information d presented by the partly revealed game board,

$$p(h|d) = \frac{p(d|h)p(h)}{\sum_{h' \in H} p(d|h')p(h')}, \quad (1)$$

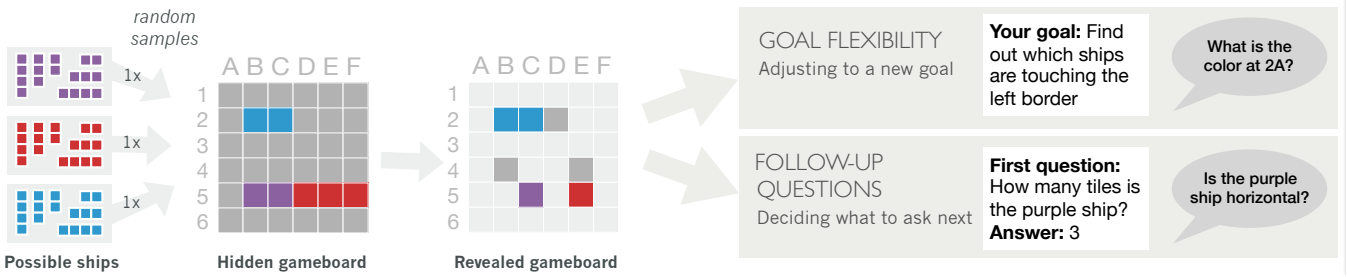

Figure 1: In the Battleship task, three ships are randomly positioned on a game board. The color at each location indicates a ship (blue, red, or purple) or water (dark gray). Participants view a game board that only shows some tiles revealed and many tiles yet unknown (light gray). They can then ask questions to obtain more information. Can people adjust their questions towards specific goals (Experiment 1)? For example for the game board shown above, when the goal is to find out which ships are touching the left border, a question targeting the color of tile 2A would be more useful than the question whether the red ship is horizontal. Do people adjust their questions based on already-answered questions (Experiment 2)? For example, after learning that the red ship is three tiles long, participants might be inclined to keep asking questions about the red ship before addressing the other ships.

where H is the hypothesis space of 1.6 million game boards and $p(d|h)$ the likelihood function, which is 1 if d is consistent with board h, and otherwise 0. The player can now ask a question x to learn more. The answer to the question is assumed to come from an oracle that knows the hidden game board and answers truthfully. We use d again as the label for the answer since it plays the same role as the partly revealed game board before. The likelihood function is again 0 if answer d is inconsistent with board h. Otherwise it is $\frac{1}{n}$, to account for cases where there is more than one valid answer, from which the oracle then chooses uniformly. For instance, for the question "What is the location of one purple tile?" the oracle would indicate the location of one of the n purple tiles on the true game board. Usually though, in our setting there is only one valid answer, $n = 1$ (e.g., yes or no). We now generalize to include the history of previous questions X and their answers D, resulting in

$$p(h|d,D;x,X) = \frac{p(d|h;x)p(h|D;X)}{\sum_{h' \in H} p(d|h';x)p(h'|D;X)}, \qquad (2)$$

where the semi-colon notation indicates that x and X are parameters rather than random variables (for the first question, X and D are empty).

In preparation for the next section, we can compute the posterior predictive probability that d will be the answer to question x via

$$p(d|D;x,X) = \sum_{h \in H} p(d|h;x)p(h|D;X). \qquad (3)$$

Expected Information Gain (EIG)

The player's uncertainty about the hidden game board is measured by the Shannon entropy of the belief distribution (Shannon, 1948; see Crupi, Nelson, Meder, Cevolani, & Tentori, 2018, for a discussion of alternative measures). The Information Gain (IG) of a question x is then defined as the amount by which this uncertainty is reduced when receiving answer d. Since the player does not know the answer at the time of asking, we compute an expected value:

$$\begin{aligned} EIG(x) &= \sum_{d \in A_x} p(d|D;x,X)\Big[I[p(h|D;X)] - I[p(h|d,D;x,X)]\Big] \\ &= \mathbb{E}_{d \in A_x}\Big[I[p(h|D;X)] - I[p(h|d,D;x,X)]\Big], \end{aligned}$$

where $I[\cdot]$ is the Shannon entropy, and A_x are the possible answers to question x. EIG has been used to describe a range of information sampling behavior (see Coenen, Nelson, & Gureckis, 2018, for an overview).

EIG for goal-directed questions. So far, EIG aims to reduce all uncertainty in $p(h)$. In order to only reduce the uncertainty that is relevant for a particular goal, we introduce the goal state space g. To illustrate with an example in the Battleship task, the goal "Find out which ships are touching" has as goal states g the various possibilities of ships that could be touching (i.e., *none*, *blue|red*, *blue|purple*, etc). Furthermore, g is defined as a projection of the hypothesis space h. Table 1 provides a minimal example of such goal projection. We can now measure the quality of a question x with respect to goal g via

$$EIG_{goal}(x,g) = \mathbb{E}_{d \in A_x}\Big[I[p(g|D;X)] - I[p(g|d,D;x,X)]\Big]. \qquad (4)$$

In detail, we compute the belief distribution over the goal states by marginalizing over h (here shown for the posterior, the equivalent is to be done for the prior)

$$p(g|d,D;x,X) = \sum_h p(g|h)p(h|d,D;x,X),$$

where $p(g|h)$ is 1 if h is goal-projected onto g, and 0 otherwise. More simply stated, for each goal state, we sum the belief values from the hypotheses that are projected onto the goal state. The EIG with respect to this goal is then the expected uncertainty reduction in the belief distribution over

Table 1: Simple example of a goal projection. Four hypotheses in h are projected onto two goal states in g. The projection results in a prior belief $p(g)$ of 0.2 for goal state 1, and 0.8 for goal state 2.

p(h)	h	g
0	1	1
0.2	2	1
0.4	3	2
0.4	4	2

these states. For convenience, we will subsume EIG_{goal} under the label EIG outside of this section.

EIG for follow-up questions. With the setup explained so far, the ability to take an already answered question into account comes out-of-the-box for the EIG model. Observed data D can be the visual information provided by the partly revealed board, as well as the verbal information from the answers to previous questions. The resulting knowledge is encoded in the posterior belief distribution, $p(h|d,D;x,X)$.

Experiment 1 – Asking goal-directed questions

In general, people ask different questions when they have different goals. When their goal changes, people should be able to flexibly adapt the questions they want to ask. In this experiment, we investigate whether people's evaluations of question usefulness are sensitive to specific goals.

Participants

Forty participants recruited on Amazon Mechanical Turk, with restriction to the United States pool, were paid a base of \$2 with a performance based bonus of up to \$4.86.

Method

In order to lead participants into a situation in which they wanted to ask a question, we took a number of steps to make them familiar with the Battleship task. First, participants went through a tutorial that presented the game board and the possible colors, sizes, orientations, and positions of the ships. This key information was shown on the side over the whole experiment and additionally checked in a comprehension quiz after the tutorial. Next, participants went through a warm-up phase, in which they began with a completely unidentified game board and clicked on the grid tiles to turn over their color, revealing more of the game board step by step.

Then, participants started the main phase, which consisted of 18 randomized trials. The schema of a trial is shown in Figure 2. Participants first viewed a partly-revealed game board and received a goal. They then ranked six natural-language questions "such that good questions are at the top and not so good questions are at the bottom" by dragging and dropping each question into a sortable list. To make sure that people paid attention to the questions, we displayed them one by one in a random order and people had to press the correct button

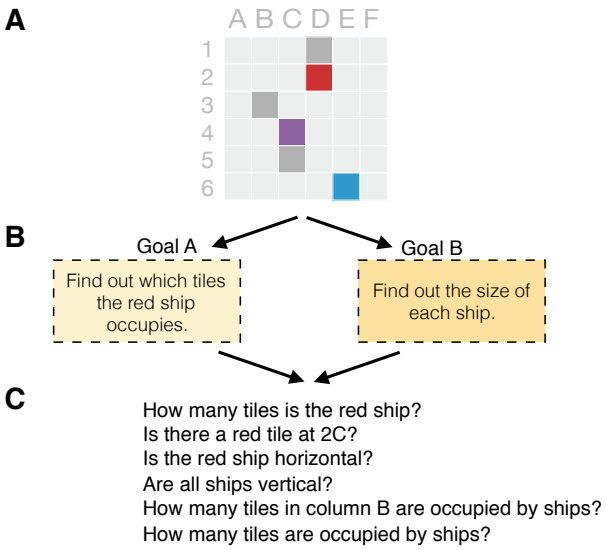

Figure 2: Experimental design of Experiment 1. In a given trial, (A) participants view a partly revealed game board. (B) Participants then receive one of two goals, randomly assigned. (C) Participants rank six questions by quality with respect to the goal. In Experiment 2, Goal A and Goal B are each replaced with an already-answered question.

that described the answer type of the question (either a color, a coordinate on the grid, a number, or yes/no). For each correct response, a bonus of \$0.045 was awarded. Allocating bonuses in this way, rather than basing it on their ranking of questions, discouraged participants from attempting to infer a researcher-preferred ranking of questions.

All participants viewed the same 18 partly-revealed game boards and corresponding question sets. But, as Figure 2 illustrates, the goal they received was randomly chosen from a predefined set of two goals for each context. The 18 game boards and the corresponding questions were the same as in Rothe et al., to ensure maximal comparability across studies (see Rothe et al., 2018, for details on the design of the boards and question sets).

The goals were designed as follows. We created a list of goals that seemed interesting but intuitive, such as "Find out which ships are touching the top border", "Find out which tiles the red ship occupies" which would allow people to ignore the blue and purple ship, or "Find out the size of each ship" which would allow them to ignore the orientation and location of the ships.

For each context, we determined via computer simulation a pair of opposing goals, such that the resulting EIG model scores of the questions were maximally different when evaluated against each goal (as measured by correlation). Examples of these opposing goals are shown as titles of the panels in Figure 3C. The average correlation between model scores within the goal pairs was $r = -0.28$.

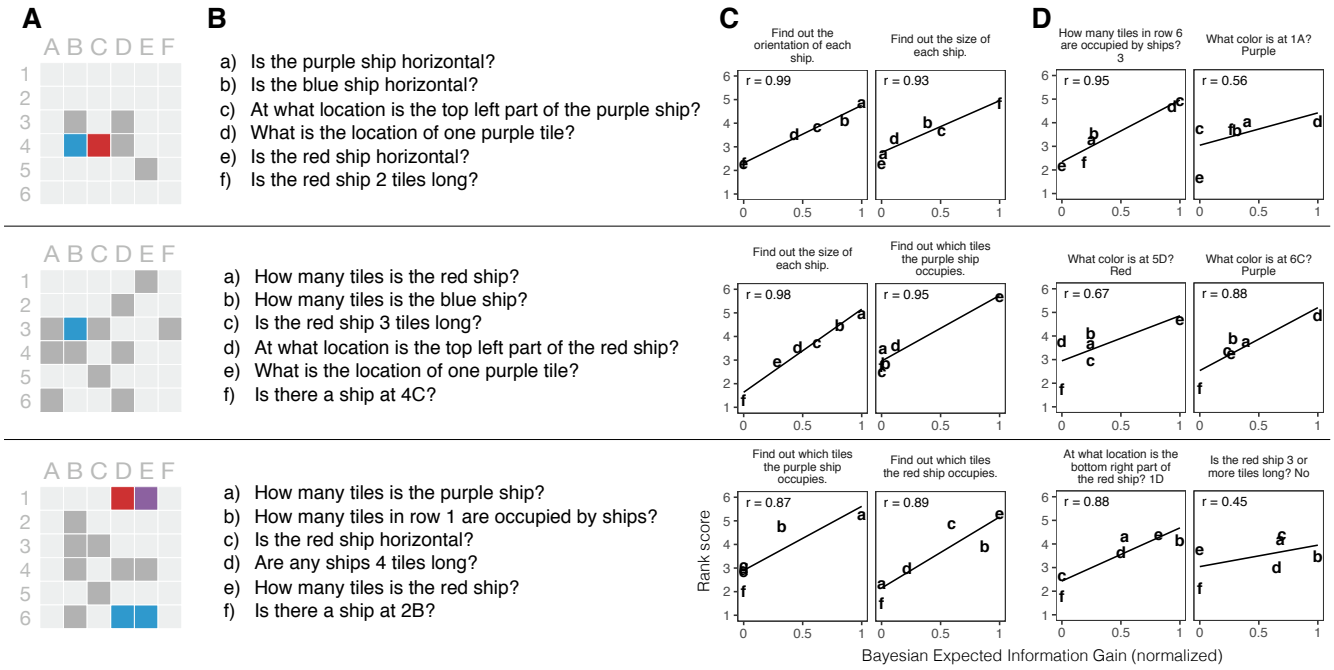

Figure 3: Contexts and questions together with human rankings and model predictions. Three selected trials exemplifying (A) the partly revealed game board, (B) the sets of six questions that were ranked by participants, (C) correlations of human rankings (y-axis; higher is better) and model scores (x-axis) for these questions in Experiment 1, and (D) Experiment 2. The letters a-f in the scatterplots correspond to letters marking the questions to the left. Error bars for $\pm 1SE$ are not plotted as they are only as large as the letters. Model scores are normalized to a maximum of 1.

Results

People's preference for questions were highly sensitive to the specific goals they had. Figure 3C shows how people's rankings of the same questions varied widely depending on the different goals. For example, in the first row in Figure 3, the question "Is the purple ship horizontal?" (marked with the letter **a**) was ranked best for the goal "Find out the orientation of each ship" but very low for the goal "Find out the size of each ship." The different rankings of this and other questions were well captured by the EIG model, which took the respective goal that participants had into account. Figure 3C shows several examples with strong correlations between EIG and human rank scores. Across all contexts, the average Pearson correlation between model scores and human rankings was $r = .84$. In contrast, when we let the EIG model hypothetically take the respective *opposite* goal into account, correlations dropped to an average $r = -.16$.

We also computed an "ignorant" model that ignored the specific goal and instead tried to obtain as much information as possible for the complete game board. A participant whose ratings are well captured by this model is probably ignoring the specific goal and instead plays the original Battleship game. The average correlation for this model was $r = .42$.

Instead of comparing correlation coefficients, we conducted a more sensitive model comparison that takes guessing behavior into account. Model scores were transformed into choice probabilities via the softmax function

$$p(x) = \frac{e^{-\beta M(x)}}{\sum_x e^{-\beta M(x)}},$$

where $M(x)$ is the model score (e.g., EIG(x)) and β is the free temperature parameter, capturing more guessing behavior as $\beta \to 0$. For each model, β was fit per participant to the rankings, and the resulting log-likelihood of the top ranked question computed.

In direct comparison, EIG had higher log-likelihood than EIGopposite, which took the opposite goal into account, for 38 out of 40 participants (95%). EIG also had a higher log-likelihood than EIGignore, which ignored the goal, for 35 out of 40 participants (88%).

We can conclude from this that people are very sensitive towards the specific goals when making question evaluations in our task, and that their evaluations are well predicted by our goal-oriented Bayesian ideal-observer EIG model with zero free parameters.

Experiment 2 – Asking follow-up questions

We test the EIG model further with the very natural task of deciding what to ask next, after a question was already answered.

Participants

A separate set of forty participants recruited on Amazon Mechanical Turk, with restriction to the United States pool, were paid a base of $2 with a performance based bonus of up to $4.86.

Method

The materials and procedure were identical to Experiment 1, except that instead of a goal, a question and its answer were displayed. That is, for each context, there was a predefined set of two already-answered questions from which one was randomly chosen for each participant (cf. Figure 2).

As for the pairs of goals in Experiment 1, we identified via computer simulation pairs of already-answered questions that were as anti-correlated as possible. The following procedure was repeated for each game board context. From a list of 136 unique questions we simulated all possible answers to each question. For example, the question "How many tiles in row 6 are occupied by ships?" (Figure 3D, first panel) has the possible answers {0, 1, ..., 6}. Then, for each question-and-answer combination, we computed what the resulting EIG scores would be for the six questions (Figure 3B) that now served as follow-up question candidates. As before, we created pairs of already-answered questions that had the most different model scores for the follow-up question candidates, as measured by the lowest correlation. The average correlation between model scores within the pairs was $r = 0.02$.

Results

People's rankings of the follow-up questions are generally sensitive to the information provided by the already-answered question. Figure 3D shows the correlations between the EIG model and human rankings. Overall, the average Pearson correlation was $r = 0.71$. When computing EIG by hypothetically taking the opposed already-answered question into account, the average correlation dropped to $r = 0.29$. When computing EIG that ignores the information from the answered question, the average correlation was $r = 0.60$. This suggests that people evaluated the usefulness of the follow-up questions by integrating the verbal information provided by the first question and its answer.

Again, instead of comparing correlation coefficients, we modeled people individually via a softmax function. EIG had a higher log-likelihood than EIGopposite for 28 out of 40 participants (70%). Using a softmax function again and modeling people individually, EIG had a higher log-likelihood than EIGopposite for 28 out of 40 participants (70%). Surprisingly, EIG had a higher log-likelihood than EIGignore for only 21 out of 40 participants (52%). The latter comparison suggests that a fair number of participants were not sensitive to the information from the answered question.

To inspect this result more carefully, we set up a hybrid model that balanced between EIG and EIGignore with a free parameter, $\theta EIG(x) + (1-\theta)EIG_{ignore}$. The balancing parameter θ was fit simultaneously with the softmax guessing parameter β for each person. The resulting distribution suggests

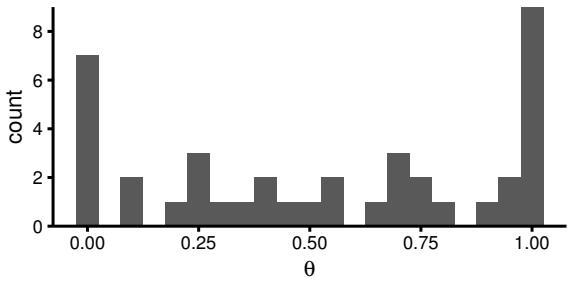

Figure 4: The θ values among participants. A participant's θ could be taken as an indicator for how much she considered the information from the answered question in Experiment 2. The binwidth is .05.

that nine participants (23%) took the information that the answered question provided accurately into account ($\theta > .95$), seven (18%) completely ignored the information ($\theta < .05$), and 24 (60%) exhibited a mixed strategy (Figure 4). Under this analysis it is still possible that the in-between participants used a different strategy, neither captured by EIG nor EIGignore. Yet, the log-likelihoods for these participants were as good as for the others suggesting that they did indeed use a mix of EIG and EIGignore. Thus, a participant's θ could be interpreted as the amount to which she considered the verbal information from the answered question.

Word-based model. We further considered an alternative model that takes a word-based approach. One strategy people might exhibit is to keep focusing on getting information about the ship they already have some details on. For instance, if the answered question provides information that the red ship is horizontal, they might prefer to learn about the size of the red ship before moving on to the next ship. Thus the word-based model looks for signal words that match the already-answered question and the follow-up question. Formally, the Color feature compares the color words {*blue, red, purple, water*} in the answered question with those in the follow-up question candidates. If there exist color words in both questions and they are the same, then the Color feature assigns a 1 to the follow-up question, else a 0. To illustrate, consider the right panel in the third row in Figure 3D. The already-answered question "Is the red ship 3 or more tiles long?" mentions the red ship. Therefore, the model would prefer the follow-up questions **c** and **e** because they also mention the red ship. Indeed, **c** and **e** were both ranked somewhat higher than predicted by the EIG model. Overall, questions that were ranked as best by people had more often matching color words (23%) with the already-answered question than the lower ranked questions (12-21%).

Another strategy that people might employ is to prefer a question of the same type as of the one that was already answered. The Type feature categorizes questions into mutually exclusive groups of *ship orientation, ship size, adjacency, region, location,* and *demonstration* questions. This classifica-

tion follows the one described in Rothe et al. (2018). The Type feature simply assigns a 1 if both questions are classified into the same type, else 0. Illustrating for the same case as above, the feature classifies the answered-question into the *ship size* type, to which also follow-up questions **a**, **d**, and **e** belong, which therefore get a higher score. Overall, questions that people ranked worst were more often of the same question type as the already-answered question (30%) than the higher ranked questions (21-23%).

The word-based model combines both features in a linear combination. We fitted a linear regression using both features as predictors and participants' average rank scores as criterion. However, only little variance in people's rankings could be explained this way, $R^2 = 0.05$.

Discussion

We tested people's preference for questions in two crucial situations: when asking goal-directed questions and when asking follow-up questions. In Experiment 1, we manipulated the goal that people had, while keeping everything else constant. People's rankings of question quality dramatically shifted based on the goal they were assigned. The rankings were well predicted by our Bayesian ideal-observer model of Expected Information Gain (EIG) with zero free parameters. In Experiment 2, we manipulated what already-answered question people received, while keeping everything else constant. Again, people's rankings shifted strongly based on the answered question. However, the picture was less clear than in the first experiment. While generally people's rankings were well predicted by the EIG model, detailed analysis suggested that people varied in the amount to which they integrated the information provided by the already-answered question. An alternative model that approximated question usefulness based on superficial features could not explain human rankings.

So far, neural network approaches to question asking generally struggle with the flexibility that is necessary to take previous answers and goals into account. To reach competitive performance in simple tasks they already need training on large data sets with tens of thousands of questions. In order to add sensitivity towards specific answers and goals would require additional training likely in orders of magnitude more.

One of the strengths of the Bayesian approach is the seamless integration of visual and verbal information. The visual information from the partly revealed game board and the verbal information from the answered question were both integrated into a unified posterior. In our current analysis we only considered varying degrees to which people considered the verbal info from the answered question. It is also possible that people did not perfectly take the visual information from the partly revealed board into account. In future work, we will further explore people's integration of high-level information.

We extended the computational framework to two aspects of question asking—more needs to be done. In our setting, we assumed a reliable, all-knowing oracle that is providing the answers. However, the relationship between question, ground truth, and generated answer is not as deterministic in many real-world settings. For example, in social settings, people need to take into account the knowledge state and goals of their communication partner. This aspect has been elegantly modeled in the Rational Speech Act framework, where a questioner has an internal model of the answered that she simulates recursively before deciding what to ask (Hawkins & Goodman, 2017). We see our approach as complementary to this RSA model. Future work should aim to integrate both.

Acknowledgments

This research was supported by NSF grant BCS-1255538, the John Templeton Foundation "Varieties of Understanding' project, a John S. McDonnell Foundation Scholar Award to TMG, and the Moore-Sloan Data Science Environment at NYU.

References

Coenen, A., Nelson, J. D., & Gureckis, T. M. (2018). Asking the right questions about human inquiry: Nine open challenges. *Psychonomic Bulletin & Review*, 1–41.

Crupi, V., Nelson, J. D., Meder, B., Cevolani, G., & Tentori, K. (2018). Generalized Information Theory Meets Human Cognition: Introducing a Unified Framework to Model Uncertainty and Information Search. *Cognitive Science*, *42*(5), 1410–1456.

Du, X., Shao, J., & Cardie, C. (2017). Learning to Ask: Neural Question Generation for Reading Comprehension. *arXiv:1705.00106v1*.

Graesser, A. C., Langston, M. C., & Bagget, W. B. (1993). Exploring information about concepts by asking questions. *The Psychology of Learning and Motivation*, *29*, 411–436.

Hawkins, R., & Goodman, N. (2017). Questions and answers in dialogue. *PsyArXiv: j2cp6*.

Jain, U., & Schwing, A. (2017). Creativity: Generating Diverse Questions using Variational Autoencoders. *arXiv:1704.03493v1*.

Johnson, J., Hariharan, B., van der Maaten, L., Hoffman, J., Fei-Fei, L., Lawrence Zitnick, C., & Girshick, R. (2017). Inferring and executing programs for visual reasoning. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2989–2998).

Lee, S., Heo, Y., & Zhang, B. (2018). Answerer in questioner's mind: Information theoretic approach to goal-oriented visual dialog. In *Advances in Neural Information Processing Systems 31* (pp. 2584–2594).

Mostafazadeh, N., Misra, I., Devlin, J., Zitnick, L., Mitchell, M., He, X., & Vanderwende, L. (2016). Generating natural questions about an image. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics.*

Nelson, J. D., Divjak, B., Gudmundsdottir, G., Martignon, L. F., & Meder, B. (2014). Children's sequential information search is sensitive to environmental probabilities. *Cognition*, *130*(1), 74–80.

Rothe, A., Lake, B. M., & Gureckis, T. (2018). Do people ask good questions? *Computational Brain & Behavior*, *1*(1), 69–89.

Ruggeri, A., Lombrozo, T., Griffiths, T. L., & Xu, F. (2016). Sources of Developmental Change in the Efficiency of Information Search. *Developmental Psychology*, *52*(12), 2159–2173.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423.

Strub, F., de Vries, H., Mary, J., Piot, B., Courville, A., & Pietquin, O. (2017). End-to-end optimization of goal-driven and visually grounded dialogue systems. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (pp. 2765–2771).