Asking and evaluating natural language questions

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

Abstract

The ability to ask questions during learning is a key aspect of human cognition. While recent research has suggested common principles underlying human and machine "active learning," the existing literature has focused on relatively simple types of queries. In this paper, we study how humans construct rich and sophisticated natural language queries to search for information in a large yet computationally tractable hypothesis space. In Experiment 1, participants were allowed to ask any question they liked in natural language. In Experiment 2, participants were asked to evaluate questions that they did not generate themselves. While people rarely asked the most informative questions in Experiment 1, they strongly preferred more informative questions in Experiment 2, as predicted by an ideal Bayesian analysis. Our results show that rigorous information-based accounts of human question asking are more widely applicable than previously studied, explaining preferences across a diverse set of natural language questions.

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

Cognitive science and machine learning have both explored the ability of learners to ask questions in order to gain information. In an active learning setting, a learning machine is able to query an oracle in order to obtain information that is expected to improve performance. This is often contrasted with passive learning where examples are presented without the advantage of active control. Machine learning research has shown that active learning can speed acquisition for a variety of learning tasks (Cohn, Atals, & Ladner, 1994; MacKay, 1992; Settles, 2009). Interestingly, humans seem to benefit from active learning in similar ways (Castro et al., 2008; Markant & Gureckis, 2015).

Although these studies have revealed common computational principles between human and machine active learning, they have largely sidestepped a hallmark human ability: the capacity for asking rich, sophisticated, and even clever questions using language. Active learning algorithms and most psychological studies emphasize relatively simple types of stereotyped, non-linguistic queries (essentially "What is the category label of document X?", Angluin, 1988; Markant & Gureckis, 2015). In contrast, people can ask far richer questions which more directly target the critical parameters in a learning task. For example, when learning about categories of animals, a child not only can point at examples and request a category label (e.g., "What is that object over there?") but can also ask about characteristic features (e.g., "Do all dogs have tails?"), typical examples (e.g., "What does a lemur look like?"), related categories (e.g., "How do alligators and crocodiles differ?"), and other types of questions which constrain the space of possible concepts (Graesser, Langston, & Bagget, 1993; Mills, Legare, Grant, & Landrum, 2011).

Despite this observation, little is known about how humans generate and evaluate natural language questions, particularly from a computational perspective. For example, how do people search an infinite space of possible questions? How do people evaluate different types of questions within a common currency? In the present paper, we begin to try to answer these questions by comparing the preference people have for asking certain natural language questions to different ways of valuing questions according to an ideal Bayesian analysis.

Studying question asking in the Battleship game

We examine question asking in a simple active learning task called the Battleship game due to its superficial similarity to a single-player version of the popular children's game (Gureckis & Markant, 2009; Markant & Gureckis, 2012, 2014). The goal of the game is to determine the location and size of 3 non-overlapping ships on a 6×6 grid (Figure 1). The ships are horizontal or vertical and can be between 2 and 4 tiles long. During the standard game, a participant sequentially clicks on tiles to reveal either the color of the underlying ship part or an empty water part (*sampling phase*, Figure 1). An efficient active learner seeks out tiles that are expected to reduce uncertainty about the ship locations and avoids tiles that would provide redundant information (e.g., when the hidden color can be inferred from the already revealed tiles). At a certain point, the sampling is stopped and participants are asked to fill in the remaining tiles with the appropriate color, based on their best guess (*painting phase*, Figure 1). The score they receive is a combination of the number of observations made in the sampling phase and the number of correctly painted tiles.

The task is well suited for the present study because the underlying hypothesis space of possible ship configurations is relatively large (1.6 million possible game boards) but is easy to explain to participants prior to the start of the task. In addition, the game is interesting and fun for participants while being amenable to an ideal observer analysis (see below). The major innovation of the present paper is that we set up situations where participants can ask any question they want in natural language (e.g., "Are the ships touching?" or "What is the total area of the ships?"). This allowed us to study rich, natural language question asking in the context of a well understood active learning task.

Models of question evaluation

At times we notice somebody ask a question that strikes us as especially clever. But why do some questions seem better than others? Here we describe a set of models which provide an objective "yardstick" for evaluating the quality of participant's questions with respect to the goals of the task. In a given trial, a player must learn a hidden configuration corresponding to a single hypothesis h in the space of possi-

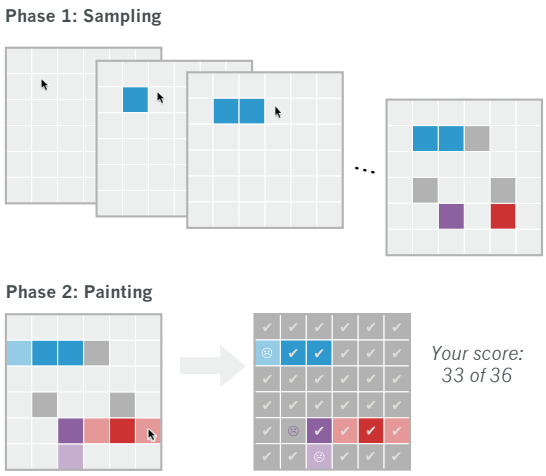

Figure 1: With the goal to locate all three ships (blue, red, purple) on the grid, a participant uncovers tile by tile. At a certain point, this *sampling phase* is stopped and the participant guesses the color of the remaining tiles. For each correctly painted tile one point is awarded.

ble configurations H. We model her prior belief distribution over the hypothesis space, $p(h)$, as uniform over ship sizes. The prior is specified by first sampling the size of each ship from a uniform distribution and second sampling uniformly a configuration from the space of possible configurations given those sizes. The player can make a query x (turning over a tile or asking a natural language question) and receives the response d (the answer). The player can then update her posterior probability distribution over the hypothesis space by applying Bayes' rule,

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

The semi-colon notation indicates that x is a parameter rather than a random variable. The posterior $p(h|d;x)$ becomes the next step's prior $p(h|D;X)$, with X representing all past queries and D representing all past responses,

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

The likelihood function $p(d|h;x)$ is $\frac{1}{n}$ if d is a valid response to the question x (and zero otherwise). The normalizing constant, n, depends on the type of question asked. For example when asking for the coordinates of any one of the tiles that contain a blue ship n is defined by the number of blue ship tiles in the true configuration. However, for most queries that we collected $n = 1$. The posterior predictive value of a new query x resulting in the answer d is defined as $p(d|D;x,X) = \sum_{h \in H} p(d|h;x)p(h|D;X)$.

Expected Information Gain (EIG). According to EIG, the value of a query x is the expected reduction in uncertainty about the true hypothesis, averaged across all possible answers A_x of the query:

$$EIG(x) = \sum_{d \in A_x} p(d|D;x,X)\Big[I[p(h|D;X)] - I[p(h|d,D;x,X)]\Big]$$

where $I[\cdot]$ is the Shannon entropy. EIG is closely related to machine learning approaches to active learning (Settles, 2009) and has a long history of study as a model of human information gathering (Oaksford & Chater, 1994).

Expected Savings (ES). According to ES, a query x is valued according to the expected reduction of errors in the painting task (Figure 1) averaged across all possible answers A_x of the query

$$ES(x) = \sum_{d \in A_x} p(d|D;x,X)\Big[EC[p(h|D;X)] - EC[p(h|d,D;x,X)]\Big].$$

Here $EC[p(h|v)]$ are the Expected Costs when coloring tiles in the painting task according to belief distribution $p(h|v)$,

$$EC[p(h|v)] = \sum_i \sum_l p(l|v;i) \times [C_{hit}\ p(l|v;i) + C_{miss}(1 - p(l|v;i))],$$

where the belief that tile i has color l is given by $p(l|v;i) = \sum_{h \in H} p(l|h;i)p(h|v)$. The choice to actually paint the tile in that color is here given by $p(l|v;i)$ again because we assume a probability matching choice function. $C_{hit} = 0$ and $C_{miss} = 1$ indicate the costs associated with painting a tile correctly or incorrectly, respectively.

Experiment 1 – Question Generation

There is an infinite number of questions that can be asked in any situation. However, most of them would have little or no information value while others would be highly informative. Our first experiment explored how people generate free-form, natural language questions in a modified version of the Battleship game. Our ultimate goal is to relate open-ended natural language questions to models of information utility.

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 \$3.60. Participants were awarded a bonus of \$0.20 for each generated question that was in line with the task rules, encouraging a minimum level of question quality without providing monetary incentives for especially rich and creative questions.[1]

Method. Before eliciting the natural language questions, a number of safeguards were implemented to help the participants understand the task. These included detailed tutorial-like instructions that explained the task and comprehension quizzes to verify understanding. In addition, key task information remained visible throughout the whole experiment.

[1] We decided against paying people based on question quality. For example, participants would have to reason about what we, the experimenters, expect to be good questions.

In a warm-up phase participants played five rounds of the standard Battleship game (i.e., turning over tiles to find the ships) to ensure understanding of the basic game play. Then, in the main phase, participants were given the opportunity to ask free-form questions in 18 trials. We defined 18 different "contexts" which refer to partially revealed game boards (see Figure 2A).

At the beginning of a trial, we introduced participants to a partly revealed game board by letting them click on a predetermined sequence of tiles (which are the past queries X and answers D in Equation 2). We chose this format of tile-uncovering moves, resembling the warm-up phase, to give the impression that a human was playing a game that was paused in an unfinished state. Subsequently, as a comprehension check, participants were asked to indicate the possible colors of each covered tile (e.g., whether the tile could be hiding a piece of the red ship). The task would only continue after all tiles were indicated correctly (or a maximum of six guesses were made).

Next, participants were given the following prompt: "If you had a special opportunity to ask any question about the grid, ships, or tiles - what would you ask?" (represented as x in Equation 2). A text box recorded participants' responses. The only two restrictions were that combinations of questions were not allowed (i.e., putting two questions together with "and" or "or") and questions had to be answerable with a single piece of information (e.g., a word, a number, true/false, or a single coordinate). Thus, participants could not ask for the entire latent configuration at once, although their creativity was otherwise uninhibited. Due to practical limitations participants asked only one question per trial, no feedback was provided and there was no painting phase. We emphasized to participants that they should ask questions as though they were playing the game they already had experience with in the earlier part of the experiment.

Question asking contexts. To produce a variety of different types of partial knowledge states or "contexts" from which people could ask questions, we varied the number of uncovered tiles (6 or 12), the number of partly revealed ships (0 to 3), and the number of fully revealed ships (0 to 2). These factors were varied independently while excluding impossible combinations leading to a total of 18 contexts/trials.

Results

We recorded 720 questions (18 trials $\times$ 40 participants). Questions that did not conform with the rules or that were ambiguous were discarded (13%) along with (3%) which were dropped due to implementation difficulties. The remaining 605 questions (84%) were categorized by type (see Table 1).

Question content. As a first stage of our analysis, we manually coded commonalities in the meaning of questions independent of the specific wording used. For example, the questions "How many squares long is the blue ship?" and "How many tiles is the blue ship?" have the same meaning for our purposes and were formalized as shipsize(blue), where shipsize is a function with parameter value blue. Since the function shipsize also works with red and purple as parameter values, it represents a cluster of analogous questions. Within these functional clusters we then considered the frequency by which such questions were generated across the 18 contexts to get a sense of participant's question asking approach (first column in Table 1).

Table 1: The natural language questions obtained in Exp. 1 were formalized as functions that could be understood by our model. The table shows a comprehensive list. Column N reports the number of questions people generated of that type. Questions are organized into broad classes (headers) that reference different aspects of the game.

N	**Location/standard queries**
24	What color is at [row][column]?
24	Is there a ship at [row][column]?
31	Is there a [color_incl_water] tile at [row][column]?
	Region queries
4	Is there any ship in row [row]?
9	Is there any part of the [color] ship in row [row]?
5	How many tiles in row [row] are occupied by ships?
1	Are there any ships in the bottom half of the grid?
10	Is there any ship in column [column]?
10	Is there any part of the [color] ship in column [column]?
3	Are all parts of the [color] ship in column [column]?
2	How many tiles in column [column] are occupied by ships?
1	Is any part of the [color] ship in the left half of the grid?
	Ship size queries
185	How many tiles is the [color] ship?
71	Is the [color] ship [size] tiles long?
8	Is the [color] ship [size] or more tiles long?
5	How many ships are [size] tiles long?
8	Are any ships [size] tiles long?
2	Are all ships [size] tiles long?
2	Are all ships the same size?
2	Do the [color1] ship and the [color2] ship have the same size?
3	Is the [color1] ship longer than the [color2] ship?
3	How many tiles are occupied by ships?
	Ship orientation queries
94	Is the [color] ship horizontal?
7	How many ships are horizontal?
3	Are there more horizontal ships than vertical ships?
1	Are all ships horizontal?
4	Are all ships vertical?
7	Are the [color1] ship and the [color2] ship parallel?
	Adjacency queries
12	Do the [color1] ship and the [color2] ship touch?
6	Are any of the ships touching?
9	Does the [color] ship touch any other ship?
2	Does the [color] ship touch both other ships?
	Demonstration queries
14	What is the location of one [color] tile?
28	At what location is the top left part of the [color] ship?
5	At what location is the bottom right part of the [color] ship?

At a broader level, there are natural groups of question types (Table 1). While this partitioning is far from the only possible scheme, it helps to reveal qualitative differences between questions. An important distinction contrasts *location/standard queries* with *rich queries*. Location queries ask for the color of a single tile and are the only question type afforded by the "standard" Battleship task (Gureckis & Markant, 2009; Markant & Gureckis, 2012, 2014). Rich queries incorporate all other queries in Table 1 and reference

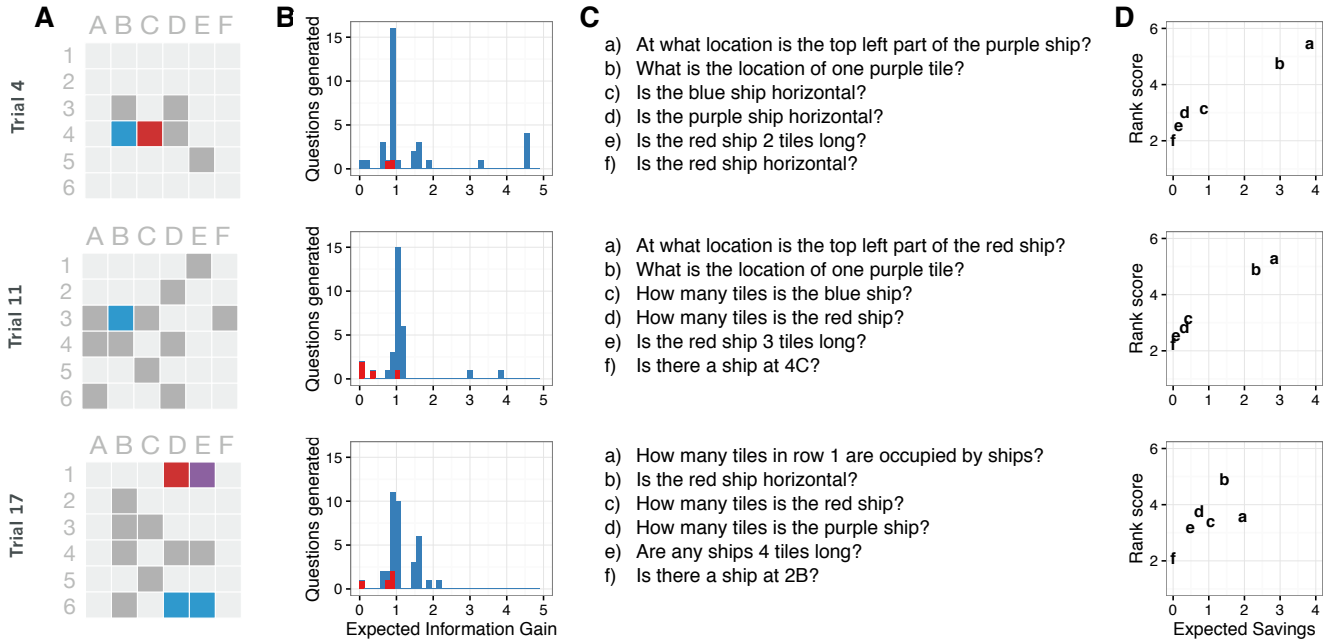

Figure 2: Three selected trials exemplifying (A) the partly revealed configuration, (B) the qualities of the questions (as measured by the Bayesian Expected Information Gain model) that participants generated in Experiment 1 (red = simple queries, blue = rich queries), (C) the six questions that were sampled from those obtained in Experiment 1 and presented in Experiment 2, (D) the values that participants (y-axis) and the Bayesian Expected Savings (ES) model (x-axis) assigned to these questions (higher score means better question).

more abstract properties of the game. Of the rich queries, *demonstration queries* ask for an example given a reference label. In the case of battleship, demonstration queries ask for an example tile of a ship, whereas most other rich queries ask about a part of feature of gameboard configuration. Demonstration queries can be especially helpful in active learning settings where the set of positive examples is relatively small (Cakmak & Thomaz, 2012; Hendrickson, Navarro, & Perfors, in press), as is the case in Battleship. Examples of high value demonstration queries are shown as the first two questions of Trial 4 and Trial 11 in Figure 2C.

Question frequencies. Among all 605 questions, only 13% were of the location/standard query type. In other words, being freed from the constraints of typical active learning studies, our participants creatively invented a vast array of questions. Only 47 questions (8%) were demonstration queries despite the fact that these can be especially useful (see below). In sum there were 139 unique questions that were repeated with different frequencies. The most popular questions was "How many tiles is the red ship?" ($n = 66$), followed by the same question asking about the purple ($n = 64$) or blue ship ($n = 55$). When grouping the questions, almost half of all generated questions ($n = 289$) addressed the size of one or several ships (Table 1). Another large group of questions targeted the orientation of the ships ($n = 116$).

Question quality. A more interesting analysis concerns the overall *quality* of these questions (as objectively assessed by the models described above).[2] One intriguing hypothesis is that there should be a positive relationship between the frequency by which a question is generated in a given context and the objective quality of the question. We used the EIG model to evaluate all 605 questions in their respective contexts. However, counter to our hypothesis, the objectively best questions were generated rarely (see Figure 2B for three example trials). The worst questions were also not generated often, while the most generated questions were in the intermediate range. Indeed, for all 18 contexts, each frequency distribution has a high peak and this peak is always below the maximum end of the distribution.

Interestingly, location/standard queries were generally inferior to richer queries. The mean EIG for location/standard queries was 0.77 compared to 1.26 for the rich queries, demonstrating the effectiveness of the more sophisticated queries (red vs. blue in Figure 2B).

Context specificity. A good question in a certain context is not necessarily a good question in a different context. To estimate the context sensitivity of the generated questions, we permuted the configurations each question was associated with across all 605 questions and evaluated the EIG for each new configuration-question pair. The average EIG across questions in the original data set was larger than in all 100 permutation sets, $p < 0.01$. Thus, people produced a range of questions that were both rich and context sensitive.

[2] Since both models make similar predictions for this section, we only report results for the EIG model here and save the model comparison for the analysis in Experiment 2.

Experiment 2 – Question Evaluation

In Experiment 2 we look more closely at how people *evaluate* natural language questions by having participants select what they viewed as the best question from a set taken from Experiment 1. In addition, we provide participants with the answer to that question.

Participants. 41 participants on Amazon Mechanical Turk were paid \$6 with a potential performance-based bonus of up to \$3.60. The higher payment compared to Exp. 1 was due to a longer experiment duration.

Method. The materials and procedure were nearly identical to Exp. 1, except that participants, rather than generating free-from rich questions from scratch, chose from a list of natural language questions asked by participants in Exp. 1. They received the answer to that question and could utilize this information in a subsequent painting phase.

Participants viewed the same 18 board configurations (contexts) along with a selection of six natural language questions which were sub-sampled from the full list of human-generated questions from the corresponding context in Exp. 1. They were asked to rank the questions for quality by positioning them from best to worst in a sortable list. After ranking they were provided with the answer to the top-ranked question and then had to do the painting task.

The reduction was necessary, as the intention of this experiment was to study question evaluation without the burden of having to consider a large number of possible questions. For sub-sampling, we used a simple algorithmic procedure designed to include the most frequently generated questions, the highest quality questions (according to EIG), and some questions that were neither frequent nor high quality.[3]

To ensure people read each question they ranked, they were asked to classify each question by the form of its possible answers (either a color, a coordinate on the grid, a number, or yes/no, which span all possible answers to the questions in Table 1). Only after a correct classification they were able to continue, in case of a wrong classification they had to wait for 5 seconds.

In contrast to Experiment 1, the bonus was tied to the performance in the painting phase. For each correctly painted tile, we awarded a potential bonus of \$0.10. The bonus was only paid for a single trial, selected by lottery at the end of the experiment. This allowed us to award a higher bonus per tile and also kept people motivated up to the very last trial.

Results

In each trial, participants ranked six rich questions by quality. Subsequently, they received the answer to the top-ranked question. In our analysis, a higher rank score represents a better question (i.e., 6 for the highest position and 1 for the lowest) and we will treat the top-ranked question as the "chosen" question.

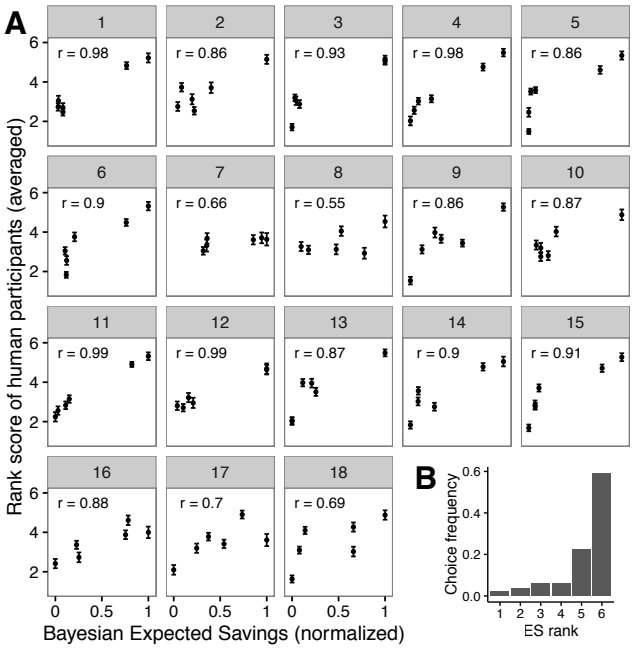

Figure 3: (A) Human rank scores (Experiment 2) and Bayesian Expected Savings (ES) model scores for the six questions per trial. Error bars show 1 se. (B) Human choice frequencies of the six questions ranked by ES, collapsed across all trials. Rank ties were resolved such that the choice counts were split between the ranks.

Figure 3A shows the high correlations between rank scores and question qualities, as measured by the Expected Savings (ES) model (average Pearson correlation $r = 0.84$). This is remarkable given that the model scores come from a Bayesian ideal-observer analysis without any free parameters.

People predominantly chose the best question (Figure 3B). The best question was the most selected question in all trials except 16 and 17, where the second-best question was favored slightly more often. Even more, the ES model scores are well reflected in the choice distributions within each trial (not shown), such that lower ES questions were selected less often (average Pearson correlation $r = 0.87$).

The correlations with the EIG scores are somewhat lower ($r = 0.70$ and 0.75 for ranking and choice, respectively). The different preferences of the models become clear with the example of the question "How many tiles are occupied by ships?" In many contexts this question has a high EIG value because it allows the learner to rule out many hypothesized configurations but a low ES value because such abstract information does often not help much with the painting task. For a more careful comparison of which model provides the best fit to the human rankings, the model utilities were transformed into choice probabilities via a softmax function with one free temperature parameter. For each model, the parameter was fit per participant to the choice data from Experiment 2. We found that ES had a higher log-likelihood for 30 out of 41 participants (73%). In addition, we looked at the

[3]The free-from questions for each context in Experiment 1 were placed in a 2D space with EIG and generation frequency as dimensions. We then sampled 1000 six question subsets and took the sample with the largest average pairwise distance between questions in the subset.

log-likelihood differences between the two models and set an arbitrary threshold to 1.6. We found an above-threshold difference in favor of ES for 26 participants (63%) but for zero participants in favor of EIG. Previous work has suggested that EIG provides a better account of human active learning than ES (Markant & Gureckis, 2012), but this work only considered location/standard queries as opposed to the rich questions we considered here.

Discussion and Conclusions

While humans and machines seem to benefit from active learning in similar ways, people ask far richer and more sophisticated types of questions. Previous experimental and computational work has used Bayesian analysis to model how people play 20 Questions with either pre-determined sets of questions (Cohen & Lake, 2016) or a small number of hypotheses (Ruggeri & Feufel, 2015; Ruggeri, Lombrozo, Griffiths, & Xu, 2015). However, people require no such restrictions. They can construct open-ended queries to resolve uncertainty in massive and novel hypothesis spaces. In this paper, we studied a probabilistic reasoning task which aimed to capture as much complexity as possible while remaining amenable to ideal Bayesian analysis. Most natural language questions can be precisely interpreted as constraints on the hypothesis space, allowing various measures of question quality to be computed exactly.

We draw a number of conclusions from two experiments. When freed from the typical constraints of active learning studies, people generated natural language questions from a rich space of possibilities spanning multiple qualitative types. In every studied context, there were a number of rich queries that were more informative than the best standard queries. Furthermore, questions were highly context sensitive and tuned to the particular partially observed game states that participants saw (as opposed to heuristic selections ignoring the current context). We found that the highest information questions were rarely generated spontaneously (Exp. 1), yet this was not because people do not recognize the quality of the questions (Exp. 2). Importantly, we were able to capture people's evaluations of natural, rich questions by a Bayesian model with zero parameters. In our setting, people's preferences are better described by the cost-sensitive measure Expected Savings rather than the cost-insensitive measure Expected Information Gain.

We also hope these findings will help inspire new, more human-like active learning algorithms. Some queries resemble the features that the generative model of a game board configuration has 'built-in' (e.g., the size or the orientation of a ship), relating to active learning algorithms that ask feature relevance queries for the purpose of classification (Settles, 2011). Unlike these algorithms, however, our participants referred to features that are inductive in nature (e.g., about ships touching each other, about one ship being larger than another, or about ships having parallel orientation). These features are interesting because they reference configural or emergent features which are not explicit in the Bayesian model. By querying these emergent properties, people must have either synthesized new features or transferred structure from related tasks. We hope this type of work will inform more human-like question asking machines.

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

Angluin, D. (1988). Queries and concept learning. *Machine Learning*, *2*, 319-342.

Cakmak, M., & Thomaz, A. L. (2012). Designing robot learners that ask good questions. In *Proceedings of the seventh annual acm/ieee international conference on human-robot interaction* (pp. 17–24). doi: 10.1145/2157689.2157693

Castro, R., Kalish, C. W., Nowak, R., Qian, R., Rogers, T., & Zhu, X. (2008). Human Active Learning. *Advances in Neural Information Processing Systems 21*.

Cohen, A., & Lake, B. M. (2016). Searching large hypothesis spaces by asking questions. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society.*

Cohn, D., Atals, L., & Ladner, R. (1994). Improving generalization with Active Learning. *Machine learning*, *15*(2), 201–221.

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.

Gureckis, T. M., & Markant, D. B. (2009). Active Learning Strategies in a Spatial Concept Learning Game. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society.*

Hendrickson, A. T., Navarro, D. J., & Perfors, A. (in press). Sensitivity to hypothesis size during information search. *Decision*.

MacKay, D. J. C. (1992). Information-Based Objective Functions for Active Data Selection. *Neural Computation*, *4*(4), 590–604. doi: 10.1162/neco.1992.4.4.590

Markant, D. B., & Gureckis, T. M. (2012). Does the utility of information influence sampling behavior? In *Proceedings of the 34th Annual Conference of the Cognitive Science Society.*

Markant, D. B., & Gureckis, T. M. (2014). A preference for the unpredictable over the informative during self-directed learning. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society.*

Markant, D. B., & Gureckis, T. M. (2015). Is It Better to Select or to Receive? Learning via Active and Passive Hypothesis Testing. *Journal of Experimental Psychology: General*, *143*(1), 94–122.

Mills, C. M., Legare, C. H., Grant, M. G., & Landrum, A. R. (2011). Determining who to question, what to ask, and how much information to ask for: The development of inquiry in young children. *Journal of Experimental Child Psychology*, *110*(4), 539–560.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608–631.

Ruggeri, A., & Feufel, M. A. (2015). How basic-level objects facilitate question-asking in a categorization task. *Frontiers in Psychology*, *6*. doi: 10.3389/fpsyg.2015.00918

Ruggeri, A., Lombrozo, T., Griffiths, T. L., & Xu, F. (2015). Children search for information as efficiently as adults, but seek additional confirmatory evidence. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society.*

Settles, B. (2009). *Active Learning Literature Survey* (Tech. Rep.). University of Wisconsin-Madison.

Settles, B. (2011). Closing the Loop: Fast, Interactive Semi-Supervised Annotation With Queries on Features and Instances. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 1467–1478).