Question asking as program generation

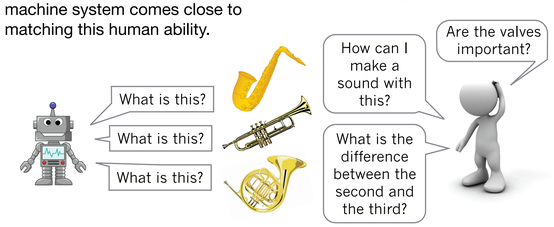
computation and cognition lab // new york university

Anselm Rothe[1], Brenden Lake[1,2], & Todd Gureckis[1]

[1]Department of Psychology, [2]Center for Data Science, New York University

IF YOU COULD ASK ANYTHING, WHAT WOULD YOU ASK? 1

People ask **rich and creative questions** when seeking information, yet no machine system comes close to matching this human ability.



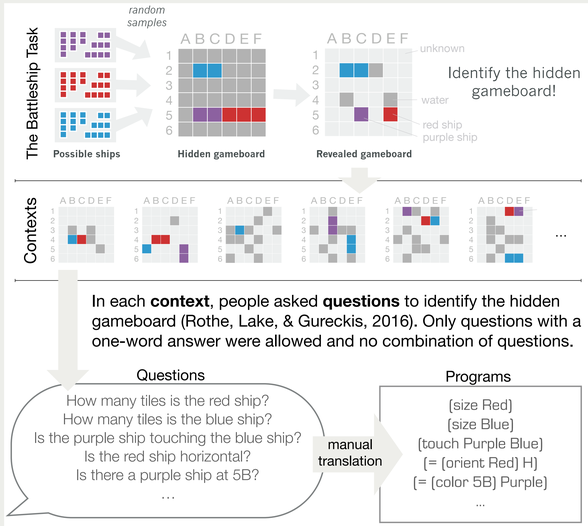
We propose a new **computational framework** that explains how people construct rich questions, treating question asking as program synthesis.

Related work
Goal-directed **dialog systems** typically only choose between a small set of canned questions ("What type of food are you looking for?").

Recent **deep learning systems** have shown interesting results but require large datasets of images paired with human questions.

However, **people** can, with virtually no practice, ask intelligent questions in novel scenarios, and can flexibly adapt to changes in task or goals.

QUESTION DATA SET 2



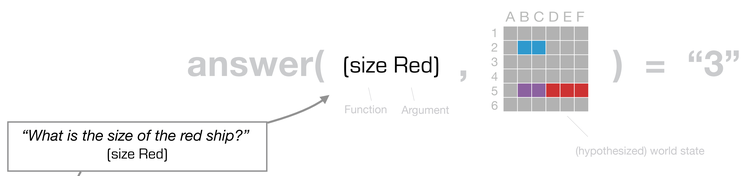
In each **context**, people asked **questions** to identify the hidden gameboard (Rothe, Lake, & Gureckis, 2016). Only questions with a one-word answer were allowed and no combination of questions.

Questions		Programs
How many tiles is the red ship? How many tiles is the blue ship? Is the purple ship touching the blue ship? Is the red ship horizontal? Is there a purple ship at 5B? …	manual translation	(size Red) (size Blue) (touch Purple Blue) (= (orient Red) H) (= (color 5B) Purple) …

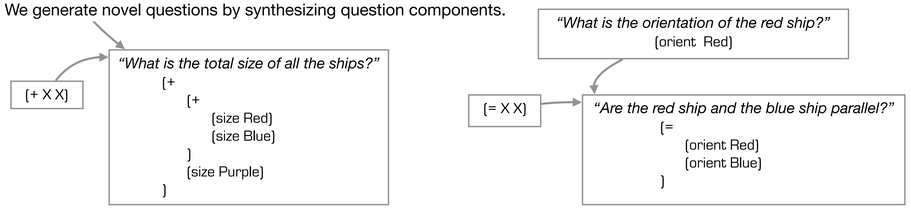
This research was supported by NSF grant BCS-1255538, the John Templeton Foundation Varieties of Understanding project, a John S. McDonnell Foundation Scholar Award to TG, and the Moore-Sloan Data Science Environment at NYU.

QUESTIONS AS PROGRAMS 3

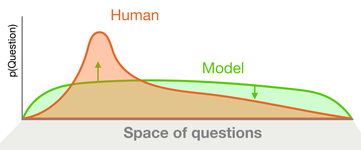
We represent questions as programs that, when executed on the state of the world output an answer.



We generate novel questions by synthesizing question components.



PROBABILISTIC GENERATIVE MODEL 4



We fit a **log-linear model** over semantic expressions, in order to estimate the latent probabilities of asking different questions given the current context.

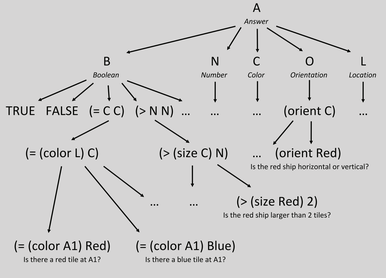
This model can be used to ask **novel questions** (i.e., plausible questions that no human asked) and to predict **what questions people will ask** in novel (unfitted) contexts.

⬅ The **space of questions** X is defined by a grammar.

GRAMMAR FOR QUESTIONS

Rules (subset)

A → B	B → TRUE	N → (size C)	C → Blue
A → N	B → FALSE	N → 1	C → Red
A → C	B → (> N N)	N → 2	C → Purple
…	…	N → 3	…
		…	



Question features
- f_1 **Informativeness** Expected Information Gain
- f_2 **Complexity** log probability under the probabilistic grammar
- f_3 **Answer type** Boolean, Number, Color, Location
- f_4 **Relevance** Auxiliary feature to filter out questions that do not address the game board

The **energy** $\mathcal{E}$ of question x is a weighted sum of its features and is related to the **probability of asking** x.

$$\mathcal{E}(x) = \theta_1 f_1(x) + \theta_2 f_2(x) + \ldots + \theta_K f_K(x)$$

$$p(x;\boldsymbol{\theta}) = \frac{\exp(-\mathcal{E}(x))}{\sum_{x \in X} \exp(-\mathcal{E}(x))}$$

Maximum likelihood estimation with question d from the human question data set.

$$\arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \log p(d^{(i)};\boldsymbol{\theta})$$

Optimization We had to approximate the gradient (via importance sampling) since the set of all questions X is intractably large.

RESULTS 5

ASKING NOVEL QUESTIONS

The model can generate **novel, "human-like" questions** that no human had asked, for new contexts that the model was not trained on.

Context	Question (manual translation)	Program (samples from model)
	What is the column of the bottom right water tile?	(rowL (bottomright (coloredTiles Water)))
	What is the row of the top left purple tile?	(rowL (topleft (coloredTiles Purple)))
	Are all the ships horizontal?	(all (map (lambda x (== H (orient x))) (set Blue Red Purple)))
	What is the column of the bottom right of the tiles with the same color as tile 3E?	(colL (bottomright (coloredTiles (color 3E))))
	Are any of the ship sizes greater than 2?	(any (map (lambda x (> (size x) 2)) (set Blue Red Purple)))
	What is the column of the bottom right blue tile?	(colL (bottomright (coloredTiles Blue)))
	How many tiles have the same color as tile 4A?	(setSize (coloredTiles (color 4A)))
	What is the top left of all the ship tiles?	(topleft (setDifference (set 1A … 6F) (coloredTiles Water)))
	What is the color of the top left of the tiles that have the same color as 5C?	(color (topleft (coloredTiles (color 5C))))
	Are blue and purple ships touching and red and purple not touching (or vice versa)?	(== (touch Blue Purple) (not (touch Red Purple)))

We removed duplicate questions that were equivalent to the human questions (determined by the mutual information of their answer vectors, as well as functional equivalence up to a swapping of the arguments, e.g. *(size Blue)* is form equivalent to *(size Red)*).

PREDICTING WHAT PEOPLE WILL ASK

A We fit 4 **model variants** to all but one context and let it predict the remaining one. The log-likelihoods of the human questions were averaged across **held out contexts**.

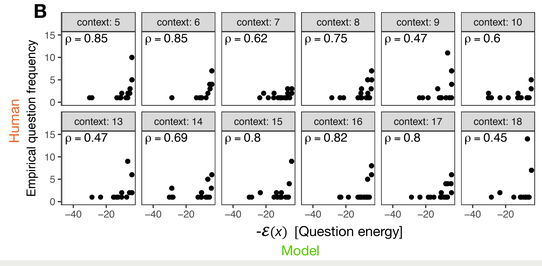
Models that had one key feature lesioned achieved a lower log-likelihood than the full model using all features.

B The predictions of the **full model** showed strong alignment with the question frequencies in the data set for some contexts and more modest alignment for others (average correlation ρ = .64).

A

Model	Features	Log-likelihood
Full	**all**	**-1400.06**
Information-agnostic	not f_1	-1464.65
Complexity-agnostic	not f_2	-22993.38
Type-agnostic	not f_3	-1419.26

B



TAKE HOME POINTS 6

By treating **questions** as **symbolic programs**, our model
- can **produce** interesting and informative, "human-like" semantic questions
- can **predict** what questions people will ask in a given game context
- can **learn** from provided answers

This also represents a new approach to query synthesis in active learning.

FUTURE DIRECTIONS 7

- Generalization to more domains beyond Battleship
- Turing test of human vs machine questions
- Automatic translation between question programs and natural language