# Cooperative Explanation as Rational Communication

**Kartik Chandra** (kach mit edu)
MIT CSAIL, Cambridge, MA 02139, USA

**Tony Chen** (thc mit edu)
MIT BCS, Cambridge, MA 02139, USA

**Tzu-Mao Li** (tzli ucsd edu)
UCSD ECE, San Diego, CA 92161, USA

**Jonathan Ragan-Kelley** (jrk mit edu)
MIT CSAIL, Cambridge, MA 02139, USA

**Joshua B. Tenenbaum** (jbt mit edu)
MIT BCS, Cambridge, MA 02139, USA

## Abstract

We offer a computational framework for modeling explanation as cooperative rational communication. Under our framework, when an explainer is faced with a "why?" question, they reason about the question-asker's current mental model, and intervene on that mental model in order to maximize the listener's future utility. We instantiate our framework in a planning domain, and show that our framework can model human explanations about plans across a wide variety of scenarios.

**Keywords:** explanation; theory of mind; Bayesian models; rational communication; social cognition

## Introduction

While wandering in Carroll's Wonderland, Alice asks for and receives dozens of explanations: some sensible, others less so. Consider the doorman in Chapter VI, who explains to Alice why there's no use in her knocking on a door: "First, because I'm on the same side of the door as you are, and second, because they're making such a noise inside, no one could possibly hear you."

No ordinary doorman would give that first explanation; if he did, we might call him uncooperative (or worse—Alice goes with "uncivil"). After all, explanations are typically cooperative social interactions: asking "why?" is a request for *help* in understanding something unexpected. The doorman's first explanation was not helpful—it failed to understand and support Alice's goals.

Decades of research into the nature of explanation (see Lombrozo (2006, 2012); Miller (2019) for reviews) has explored the many dimensions of good explanations: for example, the way good explanations are contrastive (Lipton, 1990; Riveiro & Thill, 2021), selective (Lombrozo, 2007; Gerstenberg & Icard, 2020; Poesia Reis e Silva & Goodman, 2022), and causal (Josephson & Josephson, 1996; Hilton, 1996; McClure, 2002) in nature.

In this paper, we seek to formalize the *social* dimension of explanation (Hilton, 1990; Kirfel, Icard, & Gerstenberg, 2022; Van Fraassen, 1988), approaching explanation as a kind of rational cooperative communication. We build on a long line of work that models pragmatic behavior in human communication (Grice, 1975) as recursive social reasoning. In particular, we build on the Rational Speech Acts (RSA) framework (Frank & Goodman, 2012; Goodman & Frank, 2016), which has previously been extended in a variety of ways (Sumers et al., 2023; Ho et al., 2022, 2021) to model behaviors like politeness (Yoon et al., 2018, 2017, 2016), punishment (Radkani, Tenenbaum, & Saxe, 2022), pedagogy (Shafto, Goodman, & Griffiths, 2014), and storytelling (Chandra, Li, Tenenbaum, & Ragan-Kelley, 2023a, 2023b). Under our computational framework, when faced with a "why?" question the explainer reasons recursively about the question-asker to *infer* and *correct* the question-asker's mental model. The explainer's goal is to help the question-asker now and in the future, without offending them as the doorman offended Alice. For this, we augment RSA with an additional "social" cost term.

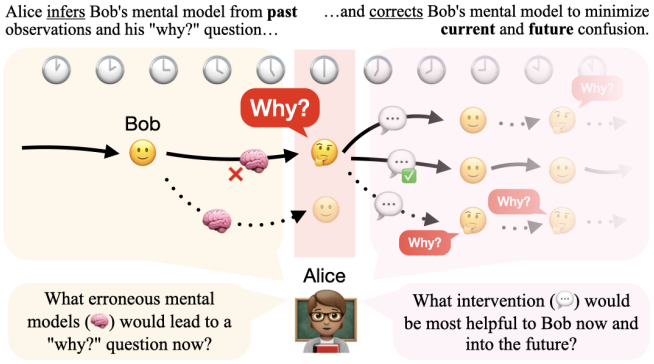



Figure 1: A computational framework for explanation. Alice, the explainer, first *infers* Bob's erroneous mental model from past observations and the fact that he is asking "why?" in the present. Then, she *corrects* Bob's mental model in order to minimize his future confusion.

In the next section, we will formalize this idea in Bayesian terms. Then, we instantiate our framework to model explanation in a planning domain, showing that our model produces human-like explanations across a variety of scenarios. Finally, we discuss scope for future work, including our ongoing efforts to apply this framework to build real-world explanation systems.

## Computational framework

In outlining our framework, we will let Alice be the explainer for a change. Our framework is centered around modeling her behavior—but we will start by describing a model of her friend Bob, who has just asked her a "why?" question.



In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.* 

Suppose Bob has some world model $m_{\text{Bob}}$, which, being a model, matches the real world in some ways but not in others. Over time, Bob observes some sequence of events $e_1, e_2, \ldots$. If at time $t$ he is surprised by $e_t$, i.e. if $p(e_t \mid m_{\text{Bob}}) \ll 1$, then Bob might suspect his world model is incorrect or incomplete. In that case, he might ask Alice "Why $e_t$?" The "why?" question is an admission that $m_{\text{Bob}}$ is missing something, and a request to Alice to fix it.

This is what Alice does when constructing an explanation. Based on Bob's past behavior from time 1 to $t$, as well as her prior belief about Bob's world model, she first *infers* $p(m_{\text{Bob}} \mid \text{Bob asked "why?" at time } t)$. Then, she chooses the best way to *correct* $m_{\text{Bob}}$, taking into account the costs and utilities associated with each potential intervention she could make (Figure 1). In the next two sections, we will formalize these two steps: *inferring* and *correcting* Bob's mental model.

## Inferring Bob's erroneous mental model

To infer $p(m_{\text{Bob}} \mid \text{Bob asked "why?" at time } t)$, we apply Bayes' rule. Let's say Alice's prior over Bob's world model is $p(m_{\text{Bob}})$. Then it remains to compute the likelihood, $p(\text{Bob asked "why?" at time } t \mid m_{\text{Bob}})$. We will assume that whenever Bob observes a surprising event (i.e. one which $m_{\text{Bob}}$ assigns probability lower than some threshold), he asks a "why?" question with probability $p_{\text{curious}}$. He might also ask spurious questions about non-surprising events with some tiny probability $p_{\text{spurious}} \ll p_{\text{curious}}$. Hence, upon observing event $e_t$ at time $t$, the likelihood of Bob asking "Why $e_t$?" is $p_{\text{curious}}$ if he found $e_t$ surprising according to $m_{\text{Bob}}$, and $p_{\text{spurious}}$ otherwise. That is,

$$p(\text{ask about } e_t \mid m_{\text{Bob}}) = \begin{cases} p_{\text{curious}} & \text{if } p(e_t \mid m_{\text{Bob}}) \ll 1 \\ p_{\text{spurious}} & \text{if } p(e_t \mid m_{\text{Bob}}) \not\ll 1. \end{cases}$$

However, Alice actually has a little more information than that: in addition to Bob's question about $e_t$, Alice observes Bob *not* ask "why?" about events $e_1, \ldots, e_{t-1}$. The likelihood of Bob *not* asking about a previously-observed event $e_{t'<t}$ is $1 - p(\text{asks about } e_{t'} \mid m_{\text{Bob}})$.

Hence, the overall likelihood of Alice's observations are given by the product of all of these terms:

$$p(\text{Bob asked "why?" at time } t \mid m_{\text{Bob}}) = \prod_{1 \le t' < t} (1 - p(\text{ask about } e_{t'} \mid m_{\text{Bob}})) \cdot p(\text{ask about } e_t \mid m_{\text{Bob}}).$$

Finally, by Bayes' rule, Alice's posterior belief about Bob's world model is proportional to the likelihood multiplied by the prior. That is, $p(m_{\text{Bob}} \mid \text{Bob asked "why?" at time } t) \propto p(\text{Bob asked "why?" at time } t \mid m_{\text{Bob}}) \cdot p(m_{\text{Bob}})$.

## Correcting Bob's mental model

At this point, Alice has *inferred* Bob's world model. Next, Alice decides how to *correct* Bob's world model.

Suppose Alice has some space $U$ of possible utterances. Let us say that utterance $u \in U$ causes a listener with world model $m$ to update their world model to $r(m,u)$. How should Alice value a given utterance $u$? This depends on the utilities and costs associated with $u$.

One natural factor to consider is whether or not $u$ resolves Bob's confusion, i.e. whether or not $p(e_t \mid r(m_{\text{Bob}}, u))$ has been raised (in expectation over the inferred $m_{\text{Bob}}$). But more generally, as noted by Tsvilodub et al. (2023), Alice might additionally wish to anticipate and (at lower priority) proactively address future confusion as well. That is, Alice might also wish to raise $p(e_{t'} \mid r(m_{\text{Bob}}, u))$ for $t' > t$.

With this in mind, we model Alice's utility of transmitting utterance $u$ as the sum

$$V(u) = \mathrm{E}_{m_{\text{Bob}}}\left[\sum_{t' \ge t} \gamma^{t'-t} p(e_{t'} \mid r(m_{\text{Bob}}, u))\right],$$

where the expectation is taken over Alice's posterior belief over $m_{\text{Bob}}$, which we computed in the previous section. Here, $0 \le \gamma < 1$ is a free parameter used to discount the value of resolving future confusion. The higher $\gamma$ is, the higher Alice values resolving Bob's potential future confusion beyond the present "why?" question.

Next, we have to consider the cost Alice incurs in choosing utterance $u$, which we will call $c(u)$. We consider two sources of costs that contribute terms to $c(u)$. First, there is a "transmission cost" proportional to the length of $u$. This models the time and energy needed for Alice to say $u$ to Bob.

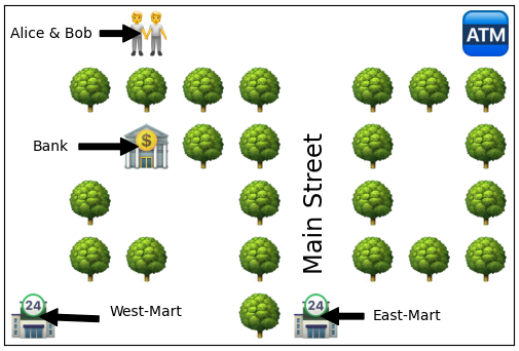
Second, we consider the "social cost" associated with telling Bob something that he *already knows*. There are many reasons why Alice might consider making such a redundant statement extra costly, on top of the transmission cost of that statement. For example, if Alice makes a redundant statement, then Bob might infer that she thinks he is ignorant, which could (a) embarrass him (e.g. in a classroom context), and (b) lower his trust that Alice has properly inferred his mental model (e.g. in a cooperative problem-solving context). Such redundancies in explanation are sometimes even perceived as rude, combative, condescending, or patronizing: for example, the pejorative term "man-splaining," and the suffix "–splaining" more generally, refer to this phenomenon (recall the example of the doorman from the introduction).

We do not explicitly model this recursive social reasoning on-line here: instead, in the spirit of resource-rational contractualism (Levine et al., 2023), we amortize that computation and reduce it to a social cost term in Alice's utility, which is proportional to the number of redundant statements Alice transmits to Bob (i.e. statements already known to be true in Bob's mental model). Because Alice is uncertain about Bob's mental model, the social cost is computed as an expectation over $m_{\text{Bob}}$. Note that the social cost is *not* subsumed by the transmission cost—we will see in the next section why it really is separately necessary.

Finally, putting everything together, we say that Alice selects utterance $u$ in a softmax-rational way, i.e. Alice selects utterance $u$ with probability $\propto \exp(\beta \cdot (V(u) - c(u))$, where the parameter $\beta$ controls the sharpness of the softmax.

## Instantiating the framework to explain surprising plans

Let us now use our framework to model how people intuitively give explanations in a simple planning domain. Consider a small town with two 24-hour convenience stores (East-Mart and West-Mart), a bank, and an ATM.




While taking a walk, Alice runs into her friend Bob and decides to take him to a convenience store to buy him a snack. Typically, Alice would take Bob directly to the nearest convenience store. However, in some scenarios, there could be unexpected exceptional circumstances that affect her plan. In particular, any (or all) of these conditions may be in effect:

("*street*") Main Street is closed for a parade.
("*cash*") Alice needs cash (has to stop at the bank or ATM).
("*mart*") West-Mart is closed for renovations.
("*strike*") The bank workers are on strike.

With this in mind, Alice plans a route and starts leading Bob. However, at some point along the way, Bob asks, "Wait, why are we going this way?" Alice then provides an explanation.

As an example, consider the scenario shown in Figure 2. Alice takes Bob south, then turns east. At that point Bob asks "Wait, why are we going east now?" Most people suggest that Alice explain "Because I need cash." We will use our computational framework to model such intuitions.

### Computational model

Recall that under our computational framework, Alice infers and corrects Bob's erroneous mental model. We modeled the space of Bob's possible mental models as the $2^4 = 16$ possible subsets of the 4 exceptional conditions. Note that this means Bob might erroneously think an exceptional condition was in effect (e.g. thinking Alice needs cash when in reality she has enough for snacks). We placed a uniform prior over these subsets, except for the empty subset (i.e. nothing unusual), which received a much higher prior. Following the principle of rational action (Jara-Ettinger et al., 2016), we assumed that Bob expects Alice to take the shortest path possible, and that he is "surprised" by an action taken by Alice (and moved to ask a "why?" question) if the action deviates from the optimal route under his mental model.

We modeled the space of Alice's possible utterances $u$ as follows: for each of the 4 exceptional conditions, Alice could either say it was in effect, say it was not in effect, or not comment on it at all. This gives $3^4 = 81$ possible utterances. Utterances denote interventions on possible world models: upon hearing $u$, Bob updates his world model to be consistent with the interventions in $u$. We set the transmission cost to the number of conditions Alice comments on (0–4), and we set the social cost to the number of statements Alice tells Bob which Bob already knows (in expectation, accounting for Alice's uncertainty over $m_{\text{Bob}}$). Finally, we restricted Alice to only say true statements.

Together, these pieces allow us to instantiate our framework in this domain.

### Alternate models

We additionally considered four alternate models:

- We lesioned Alice's inference of $m_{\text{Bob}}$. Instead, Alice always assumes that Bob is "naïve," i.e. that he believes no exceptional circumstances are in effect.
- We lesioned Alice's value of resolving Bob's future confusion. Instead, Alice seeks only to make Bob understand her *current* action. This can be implemented by fixing the discount factor $\gamma$ to be zero.
- We lesioned the social cost, so that Alice has no additional penalty for telling Bob something he already knows.
- As an alternative to the social cost, we instead added a

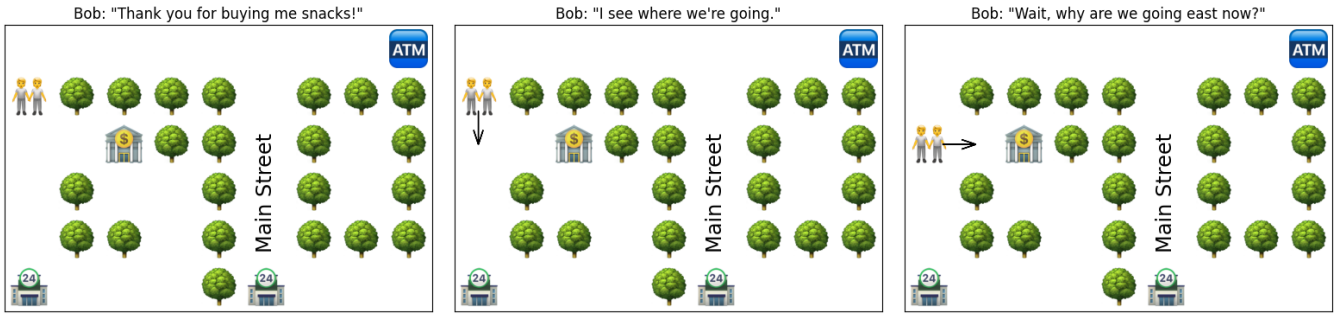



Figure 2: An example scenario in our planning domain, as shown to participants in our experiment. In this scenario, people typically say that Alice should explain "Because I need cash." Our computational model captures this intuition well—see Figure 3, where this scenario is shown as Scenario 1.

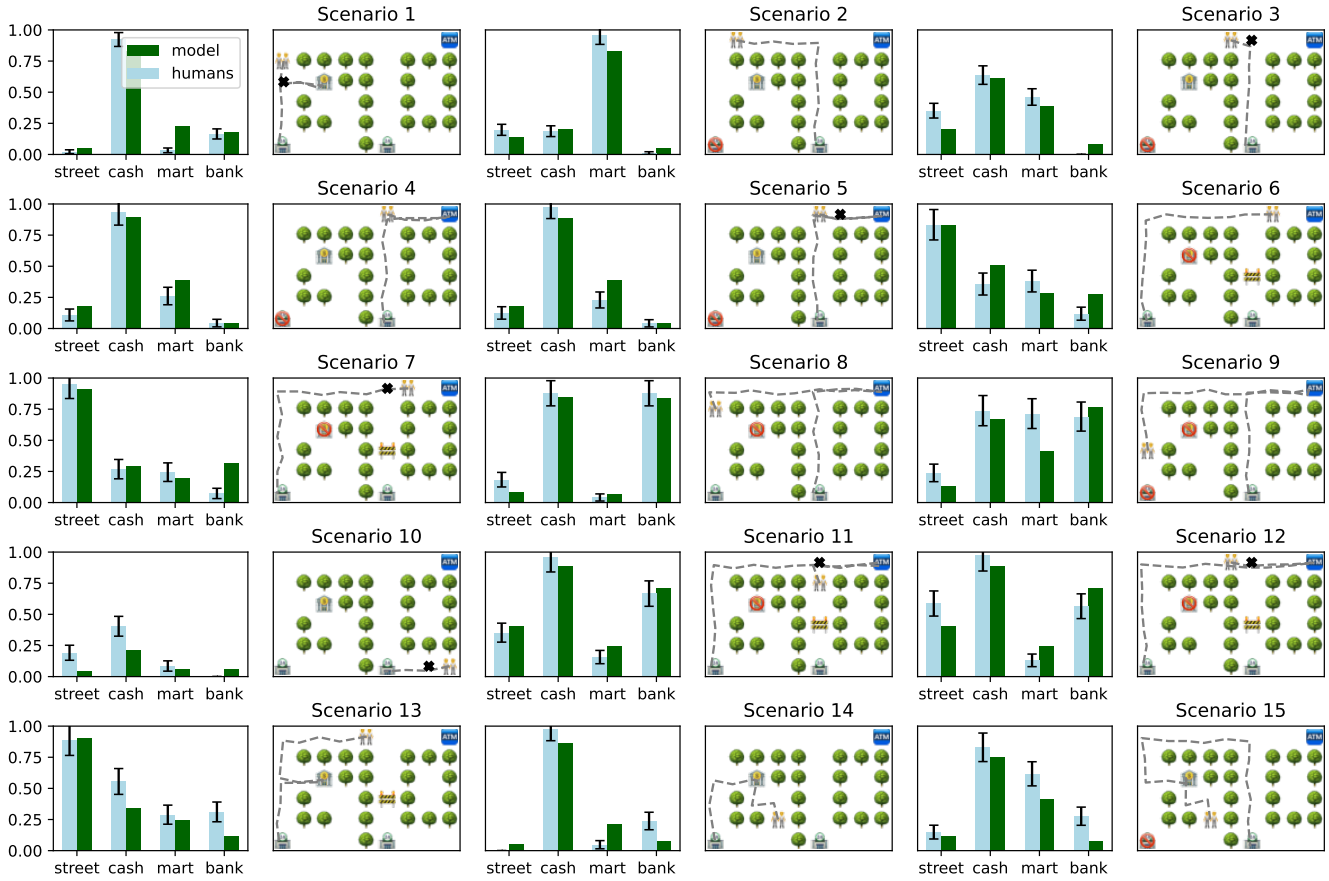



Figure 3: Our full model predicts human explanations well across a wide variety of scenarios. Here, we show a bar for the probability of including each statement in the explanation, marginalizing over the joint distribution of all possible utterances. Blue bars show human responses, while green bars show model predictions. See the text for discussion of specific scenarios. *Note: In each scenario, Bob asks "why?" at the location marked by "×," or at the beginning of the route if no × is shown.*

second layer of recursive reasoning under the Rational Speech Acts (RSA) framework (Frank & Goodman, 2012; Goodman & Frank, 2016) to see if the phenomenon of *pragmatic strengthening* (Goodman & Lassiter, 2015) provides an alternate account of the data. In RSA terms, we model Alice as an $S_2$ speaker without the social cost (as opposed to an $S_1$ speaker, as in all of the models above).

## Experiment

We recruited $n = 100$ participants on Prolific for an IRB-approved study. Each participant was paid $2.50 to complete 9 rounds, which took a median 10 minutes ($15/hour). In each round, they first read the conditions in effect for the current scenario and confirmed their understanding by selecting the best route for Alice among a set of distractors (given multiple chances). Then, they watched Alice lead Bob along the chosen path, and Bob ask a "why?" question at some point along that path (this was shown as in Figure 2). Finally, they designed an explanation for Alice to give Bob by choosing any (or all) of 4 checkboxes representing the exceptional circumstances (cash, strike, mart, parade). The checkboxes were labeled based on whether or not the exceptional circumstance was in effect in that scenario (e.g. "The bank workers are on strike." vs. "The bank is open."), and presented in randomized order. Participants were instructed that it was okay to check none of the boxes if they felt that none of the statements were a good explanation. In that case, they clicked a separate button to confirm that they intended to leave all boxes un-checked.

We collected data for 15 scenarios, chosen to elicit a variety of interesting explanations. Each scenario had the same map layout, but different starting locations for Alice and Bob, and different exceptional circumstances in effect. We excluded data from 5 participants who responded incorrectly to the understanding checks in a majority of scenarios. Finally, we independently fit our full model, as well as each of the four alternate models, to human data via black-box optimization.

## Results

Our full model fit human judgements well across a wide variety of scenarios ($r^2 = 0.88$). See Figure 3 for comparisons by scenario. Notice for example that our model captures human intuitions about how many statements to make, or how *selective* to be: compare Scenario 1, where both humans and

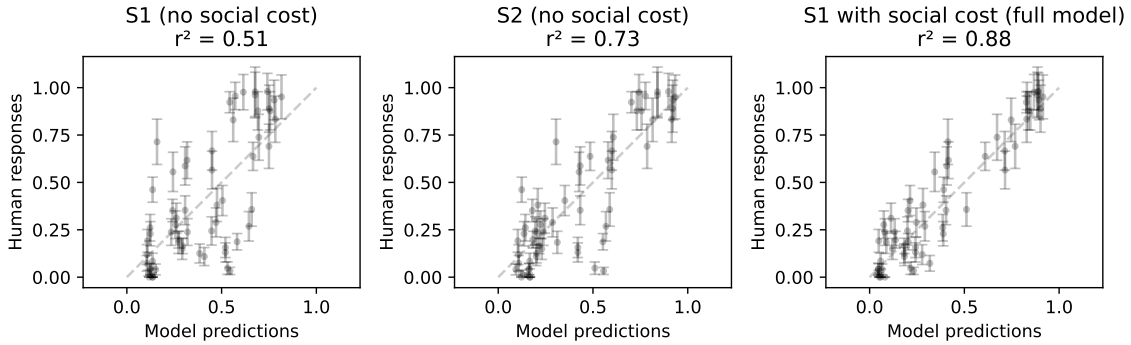



Figure 4: (**left**) Without the social cost, an $S_1$ model assigns nonzero probability to all statements, even statements that are very rarely selected by humans. (**center**) Adding another level of recursive social reasoning (i.e. an $S_2$ model) adds some pragmatic strengthening. (**right**) However, adding a social cost to the $S_1$ model captures human responses much better. (Each point in these plots represents one of the $15 \times 4$ bars in Figure 3.)

| **Model** | **Fit ($r^2$)** |
|---|---|
| Full model ($S_1$ with social cost) | **0.88** |
| No inference of $m_{\text{Bob}}$ | 0.86 |
| No future value ($\gamma = 0$) | 0.81 |
| $S_1$ (no social cost) | 0.51 |
| $S_2$ (no social cost) | 0.73 |

Table 1: A summary of the models we tested, and their fit to the data we collected. See the text for discussion.

the model choose only one statement, to Scenario 8, where they both choose two. It also captures graded uncertainty and even cases where there is no clear explanation—for example, in Scenarios 3 and 10 (which are further discussed below).

In the rest of this section we discuss our alternate models, the results of which are summarized in Table 1.

**Effect of the social cost** Lesioning the social cost from our model dramatically lowered the model fit ($r^2 = 0.51$). Without the social cost, the model assigns substantial nonzero probability to all statements, even ones that humans rarely or never select. This is visible in Figure 4 (left panel), where the model's predictions reach a floor well above zero.

To better understand this effect, consider Scenario 1 again, where the best explanation is that Alice needs cash. The lesioned model indeed assigns highest utility to that utterance. However, it also assigns some utility to Alice saying that she needs cash *and* that Main St. is open, because that utterance also corrects Bob's mental model (albeit at higher transmission cost). This raises the marginalized probability of selecting the statement about Main St.

Of course, people almost never select this statement when constructing their explanations. This suggests a unique aspect of explanation that sets it apart from other rational communication settings: the heightened cost of redundancy. In the signaling game setting described by Goodman and Frank (2016), for example, it would not be unusual for a speaker to redundantly refer to the person with "the glasses and the hat" even though "glasses" is redundant if "hat" is specified. Standard RSA models capture human intuitions well in such settings, especially when employing multiple levels of recursive reasoning. In our explanation setting here, however, redundant statements are almost never made (many bars in Figure 3 are at or near zero). Increasing the number of levels of RSA to an $S_2$ model (without a social cost) does improve the fit ($r^2 = 0.73$) compared to $S_1$, but the model still shows the same flooring effect (Figure 4, center).

As discussed earlier, we instead propose that speakers' reluctance to make extraneous statements is explained by the additional social cost to Alice of telling Bob something he already knows. Indeed, adding the social cost to $S_1$ does allow the model to predict that some statements are almost never chosen (Figure 4, right).

**Effect of reasoning about the future** Lesioning the contribution of Bob's future confusion to Alice's utility function also reduced the model fit ($r^2 = 0.81$). Consider Scenario 12, where Bob asks "why?" on the second eastward step (after crossing the corner of Main St). It is clear enough that Alice should say "because I need cash and the bank is closed." However, people and the model additionally say "because Main St. is closed," because they anticipate in the future that Bob will be surprised that they do not head down Main St. to East-Mart on the way back from the ATM. The lesioned model fails to predict this.

Another compelling example of this is Scenario 10, where Bob asks "why?" on the second westward step. This is an *odd* question—there is no reason for Bob to be confused, and indeed most people select zero statements. However, some suggest that Alice say that she does *not* need cash. Our full model captures this: the utterance lowers Bob's future surprise if he expects her to walk past East-Mart to the ATM and back. The lesioned model does not capture this effect, instead placing equal (low) weight on all four statements.

**Effect of reasoning about the past** Finally, lesioning Alice's inference of $m_{\text{Bob}}$ based on his past observations only slightly reduced the overall model fit ($r^2 = 0.86$), but its effect can still be seen in individual scenarios. Consider Scenario 3, where Bob asks "why?" when they turn south. People prefer that Alice say she does *not* need cash, than that she say West-Mart is closed. Our full model captures this: because Bob observed the earlier eastward step without asking "why?," Alice infers that it is likelier that he already knows West-Mart is closed than that he knows she has cash. In contrast, the lesioned model places precisely equal weight on "cash" and "mart."

## Discussion

### Potential extensions to the framework

Our framework currently makes several simplifying assumptions about Alice and Bob. We do not yet model uncertainty Bob might have over possible world models, or any on-line inferences Bob might make before asking his "why?" question. We also do not model Alice's uncertainty about future events. Finally, we do not model bounded rationality on Bob's part: Bob might have a correct world model, but he may have not performed enough computation to be able to predict Alice's actions. For example, if Alice were an expert chess player, Bob might ask why she sacrificed her queen for a pawn, and Alice might respond by telling him about a clever checkmate several moves ahead.

### *When* do people ask "why?" questions?

Our framework assumes that people ask "why?" questions when their expectations are violated, and our model operationalizes this assumption by modeling Bob as asking "why?" questions when he observes some event that (according to his world model) has probability less than some fixed threshold. This simplified account was enough to capture a variety of interesting effects in our model, but there are many reasons why this nonetheless remains dissatisfying.

First, people do not ask "why?" questions about *all* unlikely events—for example, to take an example from Griffiths and Tenenbaum (2007), people are likelier to ask "why did these coin tosses come up H T H T H T H T H T?" than to ask "why did these coin tosses come up H T T H H T H T H H?" even though both sequences of outcomes have the same probability of $\approx 0.1\%$. Second, people often *do* ask "why?" questions about phenomena that have relatively high probabilities and are not strong expectation violations (e.g. "why is the sky blue?" or "why do you prefer chocolate over vanilla?"). A richer model of Bob's why-question-asking might begin to address these issues by taking into account not only the likelihood Bob assigns to the explanandum, but also Bob's intuitions about what he might stand to learn or gain by understanding it.

### Practical applications of the framework

In concurrent work (Chandra, Li, Nigam, Tenenbaum, & Ragan-Kelley, 2024), we are studying ways in which we can use our framework to build tools that can automatically provide good, human-like explanations to human users—specifically, to programmers asking "why?" questions about surprising behaviors of their programs.

We were inspired by a famous software engineering talk by Bernhardt (2012), who demonstrated a variety of counter-intuitive behaviors of the programming language JavaScript. For example, he showed that in JavaScript, applying the addition operator (+) to two empty lists (`[]+[]`) produces the empty *string* (`""`) as output. The audience's nervous laughter shows that even expert programmers find such behaviors surprising (in most languages, we would expect this to give an error). Many cognitive scientists routinely encounter such surprises when using popular JavaScript-based tools like jsPsych and WebPPL.

Using our framework, we designed a tool that automatically answers questions like "Why does `[]+[]` return `""` in JavaScript?" Following our framework, we modeled Bob (the user) as having some potentially erroneous mental model of JavaScript's semantics. Inspired by VanLehn (1990) and Lu and Krishnamurthi (2024), we represented these erroneous mental models as buggy JavaScript interpreters, where "bugs" correspond to common misconceptions about the language. When a user asks "why?" about a program, our system uses modern program synthesis techniques (Torlak & Bodik, 2013, 2014; De Moura & Bjørner, 2008; Chandra & Bodik, 2017) to efficiently synthesize a buggy interpreter that would predict a different expected result from the true result of the program—that is, predict that the user would be surprised by the output. This corresponds to *inferring* misconceptions in the user's mental model. We then generate a succinct and helpful explanation of the program's behavior by *correcting* those misconceptions, i.e. by debugging the user's mental model. For example, our system explains the puzzle about JavaScript above by saying, "When the + operator is given non-numerical inputs, it converts them to strings and concatenates them. The empty list gets converted to the empty string. Hence, the result of `[]+[]` is the empty string." This closely resembles the kind of explanations humans give, for example on sites like Stack Overflow (Ventero, 2012).

## Conclusion

We presented a computational framework that formalizes explanation as rational communication: the explainer infers and debugs the question-asker's mental model. We used our framework to model explanation in a planning domain and showed that our model captures human intuitions well. Finally, we discussed a variety of future directions, including ongoing work on applying our framework to build a practical pedagogical tool for programmers. We hope these these ideas can be extended to build more tools that cooperatively help learners debug their mental models—not only of programming languages, but more broadly of any facet of our infinitely complex rabbit-hole of a world.

## Acknowledgements

We thank Tobias Gerstenberg and Thomas Icard III for thought-provoking conversations about explanation at CogSci 2023 and throughout the year since. This research was supported by NSF grants #CCF-1231216, #CCF-1723445 and #2238839, and ONR grant #00010803. Additionally, KC was supported by the Hertz Foundation and the NSF GRFP under grant #1745302, and TC was supported by an NDSEG fellowship.

## References

Bernhardt, G. (2012). *Wat.* Retrieved from https://www.destroyallsoftware.com/talks/wat

Chandra, K., & Bodik, R. (2017). Bonsai: synthesis-based reasoning for type systems. *Proceedings of the ACM on Programming Languages*, *2*(POPL), 1–34.

Chandra, K., Li, T.-M., Nigam, R., Tenenbaum, J., & Ragan-Kelley, J. (2024). Watchat: Explaining perplexing programs by debugging mental models. *arXiv preprint arXiv:2403.05334*.

Chandra, K., Li, T.-M., Tenenbaum, J., & Ragan-Kelley, J. (2023a). Acting as inverse inverse planning. In *Acm siggraph 2023 conference proceedings* (pp. 1–12).

Chandra, K., Li, T.-M., Tenenbaum, J. B., & Ragan-Kelley, J. (2023b). Storytelling as inverse inverse planning. *Topics in Cognitive Science*.

De Moura, L., & Bjørner, N. (2008). Z3: An efficient smt solver. In *International conference on tools and algorithms for the construction and analysis of systems* (pp. 337–340).

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Gerstenberg, T., & Icard, T. (2020). Expectations affect physical causation judgments. *Journal of Experimental Psychology: General*, *149*(3), 599.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, *20*(11), 818–829.

Goodman, N. D., & Lassiter, D. (2015). Probabilistic semantics and pragmatics uncertainty in language and thought. *The handbook of contemporary semantic theory*, 655–686.

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41–58). Brill.

Griffiths, T. L., & Tenenbaum, J. B. (2007). From mere coincidences to meaningful discoveries. *Cognition*, *103*(2), 180–226.

Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin*, *107*(1), 65.

Hilton, D. J. (1996). Mental models and causal explanation: Judgements of probable cause and explanatory relevance. *Thinking & Reasoning*, *2*(4), 273–308.

Ho, M. K., Cushman, F., Littman, M. L., & Austerweil, J. L. (2021). Communication in action: Planning and interpreting communicative demonstrations. *Journal of Experimental Psychology: General*. Retrieved from https://psyarxiv.com/a8sxk/

Ho, M. K., Saxe, R., & Cushman, F. (2022). Planning with theory of mind. *Trends in Cognitive Sciences*. Retrieved from https://saxelab.mit.edu/sites/default/files/publications/HoSaxeCushman2022.pdf

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in cognitive sciences*, *20*(8), 589–604. Retrieved from https://www.sciencedirect.com/science/article/pii/S1364661316300535

Josephson, J. R., & Josephson, S. G. (1996). *Abductive inference: Computation, philosophy, technology*. Cambridge University Press.

Kirfel, L., Icard, T., & Gerstenberg, T. (2022). Inference from explanation. *Journal of Experimental Psychology: General*, *151*(7), 1481.

Levine, S., Chater, N., Tenenbaum, J., & Cushman, F. (2023). Resource-rational contractualism: A triple theory of moral cognition.

Lipton, P. (1990). Contrastive explanation. *Royal Institute of Philosophy Supplements*, *27*, 247–266.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in cognitive sciences*, *10*(10), 464–470.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive psychology*, *55*(3), 232–257.

Lombrozo, T. (2012). Explanation and abductive inference. *Oxford handbook of thinking and reasoning*, 260–276.

Lu, K.-C., & Krishnamurthi, S. (2024). Identifying and correcting programming language behavior misconceptions. In *Oopsla.*

McClure, J. (2002). Goal-based explanations of actions and outcomes. *European review of social psychology*, *12*(1), 201–235.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, *267*, 1–38.

Poesia Reis e Silva, G., & Goodman, N. (2022). Left to the reader: Abstracting solutions in mathematical reasoning. , *44*(44). Retrieved from https://escholarship.org/uc/item/0j8753pd

Radkani, S., Tenenbaum, J., & Saxe, R. (2022). Modeling punishment as a rational communicative social action. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44). Retrieved from https://escholarship.org/content/qt47g8d89h/qt47g8d89h.pdf

Riveiro, M., & Thill, S. (2021). "that's (not) the output i expected!" on the role of end user expectations in creating explanations of ai systems. *Artificial Intelligence*, *298*, 103507.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive psychology*,

*71*, 55–89. Retrieved from https://www.sciencedirect.com/science/article/pii/S0010028514000024

Sumers, T. R., Ho, M. K., Griffiths, T. L., & Hawkins, R. D. (2023). Reconciling truthfulness and relevance as epistemic and decision-theoretic utility. *Psychological Review*.

Torlak, E., & Bodik, R. (2013). Growing solver-aided languages with rosette. In *Proceedings of the 2013 acm international symposium on new ideas, new paradigms, and reflections on programming & software* (pp. 135–152).

Torlak, E., & Bodik, R. (2014). A lightweight symbolic virtual machine for solver-aided host languages. *ACM SIGPLAN Notices*, *49*(6), 530–541.

Tsvilodub, P., Franke, M., Hawkins, R., & Goodman, N. (2023). Overinformative question answering by humans and machines. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 45).

Van Fraassen, B. (1988). The pragmatic theory of explanation. *Theories of explanation*, *8*, 135–155.

VanLehn, K. (1990). *Mind bugs: The origins of procedural misconceptions*. MIT press.

Ventero. (2012). *Answer to "What is the explanation for these bizarre JavaScript behaviours mentioned in the 'Wat' talk for CodeMash 2012?"*. Retrieved from https://stackoverflow.com/a/9033306

Yoon, E. J., MacDonald, K., Asaba, M., Gweon, H., & Frank, M. C. (2018). Balancing informational and social goals in active learning. In *Cogsci.* Retrieved from https://sll.stanford.edu/docs/2018_cogsci/Yoon_et_al_2018_cogsci.pdf

Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2016). Talking with tact: Polite language as a balance between kindness and informativity. In *Proceedings of the 38th annual conference of the cognitive science society* (pp. 2771–2776). Retrieved from http://socsci-dev.ss.uci.edu/~lpearl/courses/readings/YoonEtAl2016_Politeness.pdf

Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2017). "i won't lie, it wasn't amazing": Modeling polite indirect speech. In *Cogsci.* Retrieved from https://cogsci.mindmodeling.org/2017/papers/0679/paper0679.pdf