COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 47 (2023) e13257

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13257

When Naïve Pedagogy Breaks Down: Adults Rationally Decide How to Teach, but Misrepresent Learners' Beliefs

Rosie Aboody,[a] Joey Velez-Ginorio,[b] Laurie R. Santos,[a] Julian Jara-Ettinger[a]

[a]*Department of Psychology, Yale University*
[b]*Department of Computer and Information Science, University of Pennsylvania*

Received 2 February 2022; received in revised form 2 December 2022; accepted 22 January 2023

Abstract

From early in childhood, humans exhibit sophisticated intuitions about how to share knowledge efficiently in simple controlled studies. Yet, untrained adults often fail to teach effectively in real-world situations. Here, we explored what causes adults to struggle in informal pedagogical exchanges. In Experiment 1, we first showed evidence of this effect, finding that adult participants failed to communicate their knowledge to naïve learners in a simple teaching task, despite reporting high confidence that they taught effectively. Using a computational model of rational teaching, we found that adults assigned to our teaching condition provided highly informative examples but failed to teach effectively because their examples were tailored to learners who were only considering a small set of possible explanations. In Experiment 2, we then found experimental evidence for this possibility, showing that knowledgeable participants systematically misunderstand the beliefs of naïve participants. Specifically, knowledgeable participants assumed naïve agents would primarily consider hypotheses close to the correct one. Finally, in Experiment 3, we aligned learners' beliefs to knowledgeable agents' expectations and showed learners the same examples selected by participants assigned to teach in Experiment 1. We found that these same examples were significantly more informative once learners' hypothesis spaces were constrained to match teachers' expectations. Our findings show that, in informal settings, adult pedagogical failures result from an inaccurate representation of what naïve learners believe is plausible and not an inability to select informative data in a rational way.

Keywords: Pedagogy; Computational modeling; Social cognition; Theory of Mind

Correspondence should be sent to Rosie Aboody, Department of Psychology, Yale University, New Haven, CT 06520–8205, USA. E-mail: rosie.aboody@yale.edu

1. Introduction

From the first years of life, humans engage in informal pedagogy more flexibly and frequently than any other animal species (Gweon, 2019; Skerry, Lambert, Powell, & McAuliffe, 2013). Our propensity to share what we know allows us to compile extensive bodies of knowledge over time and underlies the development of human culture (Tennie, Call, & Tomasello, 2009; Tomasello, Kruger, & Ratner, 1993). However, despite the pervasiveness of informal pedagogy, sharing knowledge is far from straightforward. Explaining too much can be tedious and inefficient, while explaining too little can be outright ineffective. To share knowledge successfully, we must convey the right content in the right quantity, requiring us to reason about and track what others already know (Olson & Bruner, 1996; Tenenbaum & Griffiths, 2001). In this paper, we seek to better understand adults' capacity to engage in everyday informal pedagogy.

Informal pedagogy emerges early in the human lifecourse (see Gweon, 2021, for review). By the end of preschool, children can already infer other people's knowledge based on how they behave (Aboody, Zhou, Flowers, & Jara-Ettinger, 2019, Aboody, Zhou & Jara-Ettinger, 2021, Aboody, Huey & Jara-Ettinger, 2022 ; Einav & Robinson, 2011; Jara-Ettinger, Floyd, Tenenbaum, & Schulz, 2017; Wu & Schulz, 2018), use these inferences to decide what information to share (Baer & Friedman, 2018; Ronfard & Corriveau, 2016; Strauss, Ziv, & Stein, 2002), and expect others to do the same (Bonawitz et al., 2011; Bridgers, Jara-Ettinger, & Gweon, 2019; Gweon & Asaba, 2017; Gweon & Schulz, 2018; Gweon, Shafto, & Schulz, 2018; Rhodes, Bonawitz, Shafto, Chen, & Caglar, 2015). Given children's early understanding of pedagogical principles, one might expect adults to excel at informal pedagogy, but this is not the case. The literature on adult informal pedagogy shows a combination of clear successes (e.g., Ho et al., 2016; Shafto, Goodman, & Griffiths, 2014), as well as failures (Bromme, Brummernhenrich, Becker, & Jucks, 2012; Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001; Chi, Siler & Jeong 2004 ; Graesser, Person, & Magliano, 1995).

Naturally, teaching tasks where adults fail are more complex than those where they succeed. For example, in a now classical study, Chi, Siler, and Jeong (2004) showed that college students asked to tutor eighth graders in an open-dialogue session often failed to detect, diagnose, and correct misconceptions and tended to overestimate how much their students knew. By contrast, Shafto et al. (2014) showed that, in a constrained task (where participants could only place two dots on a screen to convey the location and size of a hidden rectangle), adults could easily generate the best possible teaching strategy (enabling learners to infer both relevant dimensions: size and location; and even six-year-olds succeed in conceptually similar tasks; Rhodes, Gelman, & Brickman, 2010).

The goal of our paper is to take a first step in investigating what causes untrained adults to struggle in more complex teaching tasks. One possibility is that the more complex a task, the more likely adults will be to misrepresent what learners know or consider plausible. If so, adults may be teaching effectively relative to what they think the learner knows but ultimately fail because this initial representation was wrong. Under this account, untrained adults succeed in simpler tasks because their representation of learners' knowledge is more accurate

but fail in more complex tasks because they do not consider the full space of possibilities that learners may be considering.

Alternatively, adults may struggle to teach in complex tasks not because they misrepresent learners' beliefs, but because they struggle to determine what kind of data would be most informative to share. That is, teaching in more complex tasks may require more complex planning because these tasks typically involve longer teaching events, where there are a greater number of possibilities about what information to share. Even if an untrained tutor understands what a learner initially knows or finds plausible, they may struggle to decide which pieces of information will be most useful to share. Under this account, adults may fail to teach effectively because deciding what to share and how to prioritize different pieces of information is challenging.

While both of these limitations likely play a role, identifying their individual contribution to informal pedagogical performance has been challenging. Tasks that elicit adult teaching failures are generally too complex for formal analysis: sessions are long and naturalistically unconstrained, making it difficult to identify the causes behind people's failures (and identifying the sources of failure is often not the main goal of the work; see VanLehn, 2011, for review). Conversely, tasks that are amenable to formal analysis are those where people generally succeed (e.g., Shafto et al., 2014), making them unsuitable for understanding the challenges that untrained tutors face in complex tasks.

1.1. The current paper

In the current paper, we seek to better understand what gives rise to teaching failures in untrained adults. Given that people excel at sharing knowledge in simple tasks (e.g., Shafto et al., 2014), but struggle in more complex ones (e.g., Chi et al., 2004), identifying the causes behind adults' successes and failures requires an intermediate task, richer than those where adults overwhelmingly succeed, but simple enough to enable us to identify the causes of variation.

To design such a task, we focused on the domain of causal reasoning, as we frequently learn and communicate cause-and-effect relations in our daily lives, and this process is well understood and has been formalized via computational modeling (see Holyoak & Cheng, 2011, for a review). Specifically, we used a "blicket-detector" paradigm, where a simple opaque box (known as the blicket detector) will activate whenever "blickets" are placed on top. Critically, blickets can be either individual objects or combinations of objects. In this paradigm, knowledgeable agents can choose different combinations of objects to place on the blicket detector, with the goal of getting a naïve learner to infer the underlying rule. Blicket-detector paradigms had their origins in developmental science (Gopnik, 2012; Griffiths, Sobel, Tenenbaum, & Gopnik, 2011; Kushnir, Gopnik, Lucas, & Schulz, 2010) but have since been used extensively with adults as well (e.g., Benton & Rakison, 2020; Gelpi, Prystawski, Lucas, & Buchsbaum, 2020; Griffiths et al., 2011; Herbst, Lucas, & Buchsbaum, 2017; Tenenbaum & Griffiths, 2003), as they provide a simple experimental paradigm to test causal learning without any interference from people's world knowledge (a potential limitation that we return to in the discussion). We implemented a causal structure we expected to be learnable but not trivially obvious, drawing on prior research showing that a causal relationship based upon the

logical operator "and"—where two blocks in conjunction are required to activate a machine—is not immediately obvious to adults (see Gopnik et al., 2017; Lucas, Bridgers, Griffiths, & Gopnik, 2014). We expected that communicating a causal structure might be more difficult for adults than previous teaching tasks (e.g., Shafto et al., 2014), while remaining simple enough to be analyzed via computational modeling.

Before proceeding, it is important to emphasize that this paper is not aimed at assessing the capacities of trained teachers and educators. Rather, we seek to understand adults' basic, everyday abilities to communicate our knowledge. For simplicity, in the remainder of the paper, we use the words "teach" and "teaching" to refer to the early-developing human capacity to share our knowledge informally and "teachers" to refer to participants assigned to an experimental teaching condition where they received privileged knowledge and were asked to communicate it to naïve learners. Because our goal is to understand the abilities of everyday adults, we recruited our sample from the general population.

In Experiment 1, we first confirm that our paradigm elicits teaching failures (despite people's confidence that they taught effectively), and, through a computational model of pedagogical reasoning (Shafto & Goodman, 2008; Shafto et al., 2014; Wang et al., 2020), we show that these failures are not due to adults generating poor examples. Instead, our computational model suggests that these failures emerge because naïve learners' hypothesis spaces are substantially larger than what knowledgeable adults assume. Experiment 2 directly assessed adults' expectations about learners' beliefs, finding that knowledgeable adults do not anticipate the breadth of hypotheses a naïve learner will find plausible. These results suggest that knowledgeable adults can provide useful data under the expectation that learners are only considering a constrained set of hypotheses, predicting that their examples should be fully informative when this assumption is correct. We test this prediction in Experiment 3 and find that aligning learners' hypothesis spaces to knowledgeable adults' beliefs improves learning, even from the same examples that had previously elicited learning failures. All data, analysis scripts, model code and stimuli are publicly available at the project OSF page: https://osf.io/pg9zy.

2. Computational framework

Our computational framework is inspired by previous research investigating how people share information (Frank & Goodman, 2012; Goodman & Frank, 2016; Shafto et al., 2014; Wang, et al., 2020). For consistency with literature in this field, we refer to untrained tutors as "teachers," in the sense that they are asked to engage in the act of teaching. Teachers can be formalized as generating data, given their knowledge of the correct hypothesis, and learners can be formalized as inferring which hypothesis best explains a teacher's decision to produce the observed data. The process of teachers tailoring their data to learners and learners reasoning about the pedagogical data can be formalized through a pair of recursive equations:

$$p_{teacher}(D|H) \propto p_{learner}(H|D) \tag{1}$$

and

$$p_{learner}(H|D) \propto p_{teacher}(D|H) \tag{2}$$

where $\boldsymbol{p_{teacher}}$ $\boldsymbol{(D|H)}$ is the probability that the teacher will generate certain data, D, given the true hypothesis H; and $\boldsymbol{p_{learner}}$ $\boldsymbol{(H|D)}$ is the probability that the learner will infer the correct hypothesis given the data that they observe. These equations formalize the idea that rational teachers select data that will allow rational learners to infer the right hypothesis and that rational learners infer this hypothesis by reasoning about why the teacher chose the data they did (Shafto et al., 2014).

The learner's success in recovering the right hypothesis (Eq. 2) depends on two factors: the teacher's data and the learner's hypothesis space. Here, our main interest is in using this model to evaluate teachers' data. Thus, $\boldsymbol{p_{teacher}}$ $\boldsymbol{(D|H)}$ is obtained from Experiment 1, and these data are evaluated using Eq. 2. To gauge the quality of the data, we designed a set of hypothesis spaces that sequentially increase in complexity, by combining basic primitive hypotheses using two logical operators: *AND* and *OR*.

In our task, participants were introduced to a novel machine and five blocks (lettered A–E). Participants assigned to our teaching condition learned that together, blocks B and E activated the machine (see the Method section of Experiment 1 for details). Here, the most basic "primitive" hypotheses correspond to a belief that a specific block must be on top of a machine for it to activate (e.g., that block A activates the machine, that block B activates the machine, and so on). There were five available blocks, and thus there are five primitive hypotheses. To scale hypothesis spaces up in complexity, we varied two dimensions.

First, we expanded our hypothesis spaces by increasing the number of primitive hypotheses (A, B, C, D, or E) that could be combined into a single hypothesis (called the *depth*). For example, at a depth of two, up to two primitives can be combined per hypothesis (yielding hypotheses such as: (B); *AND*(A, C); *OR*(B, E)).[1] These hypotheses correspond to simple beliefs participants might hold about the machine (e.g., that block B makes the machine activate by itself, that blocks A and C together are required to make the machine activate, or that either block B or E makes the machine activate). Increasing the depth allows for more complex hypotheses: At a depth of three, up to three primitives can be combined in a single hypothesis (yielding hypotheses such as *OR*(A, B, E), which expresses a hypothesis that either block A, B, or E makes the machine activate by itself). At a depth of four, up to four primitives can be combined in a single hypothesis and so on. This approach is consistent with related work in cognitive science that captures how hypothesis spaces scale in complexity (see Jin et al., 2018), finding that expression length is a useful proxy for complexity (Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Piantadosi, Tenenbaum, & Goodman, 2012; Velez-Ginorio, Siegel, Tenenbaum, & Jara-Ettinger, 2017).

In these hypothesis spaces, each set of primitives is combined by only one type of logical operator (i.e., primitives are either combined by *AND* or *OR* but not both). We refer to these as "single-operator" spaces. A second way to expand hypothesis spaces is by allowing individual hypotheses to combine both logical operators. We refer to these as "dual-operator" spaces. This adjustment introduces compositionality; a dual-operator hypothesis space at depth three could contain hypotheses such as: *OR*(*AND*(B, E), (E)) (which corresponds to the

hypothesis that either B and E together are required to make the machine activate, or that E may be sufficient by itself).[2] Although single-operator hypothesis spaces can only be scaled up to a ceiling of five (e.g., *AND*(A, B, C, D, E); *OR*(A, B, C, D, E)), dual-operator spaces can be scaled much higher (e.g., with a ceiling of seven, one possible hypothesis would be: *OR*(*AND*(A, B, E), *AND*(B, E), (B), (E), corresponding to the hypothesis that either A, B, and E together are required to make the machine activate, that B and E together are required, or that B or E may be sufficient on their own to make the machine activate).

3. Experiment 1

Prior research on teaching in naturalistic contexts has focused on situations where knowledgeable and naïve individuals directly interact (e.g., Chi et al., 2004). However, in these contexts, it can be difficult to determine to what extent pedagogical outcomes can be attributed to the teacher versus the learner. For instance, if some learners are more naturally communicative than others, their teachers may have an easier time selecting informative data. Our task was designed to enable independent evaluation of teacher and learner performance. Thus, knowledgeable participants assigned to a teaching condition first selected examples to reveal how a blicket detector worked, and these examples were then shown to multiple naïve learners (allowing us to measure how effective a teacher's examples are without concern that these results will be unduly influenced by a single inattentive learner).

Specifically, Experiment 1 assigned participants to either a teaching or a learning task. Participants in the teaching task learned how to activate a machine and then generated examples to demonstrate how the machine worked. Participants in the learning task saw one teacher's examples and then explained how the machine worked. Teachers' performance was assessed based on the proportion of naïve participants who learned how the machine worked. The sources of any failures were analyzed by providing teacher-selected data as an input to our model.

3.1. Methods

3.1.1. Participants

Two hundred and twenty participants were recruited from Amazon's Mechanical Turk platform. The first 20 participants (mean age = 35.4; range = 21–70) were assigned to the teacher condition, and the last 200 participants (mean age = 34.4; range = 19–68) were assigned to the learner condition. Five additional participants were recruited but not included in the experiment because they failed an inclusion question (teacher $n = 1$; learner $n = 2$) or because they did not follow task instructions and provided fewer than three unique examples (teacher $n = 2$).

3.1.2. Stimuli

Stimuli consisted of images of a "light-up" machine with a triangle on the front and five blocks lettered A–E. The color of the triangle signaled whether the machine was on or off (see

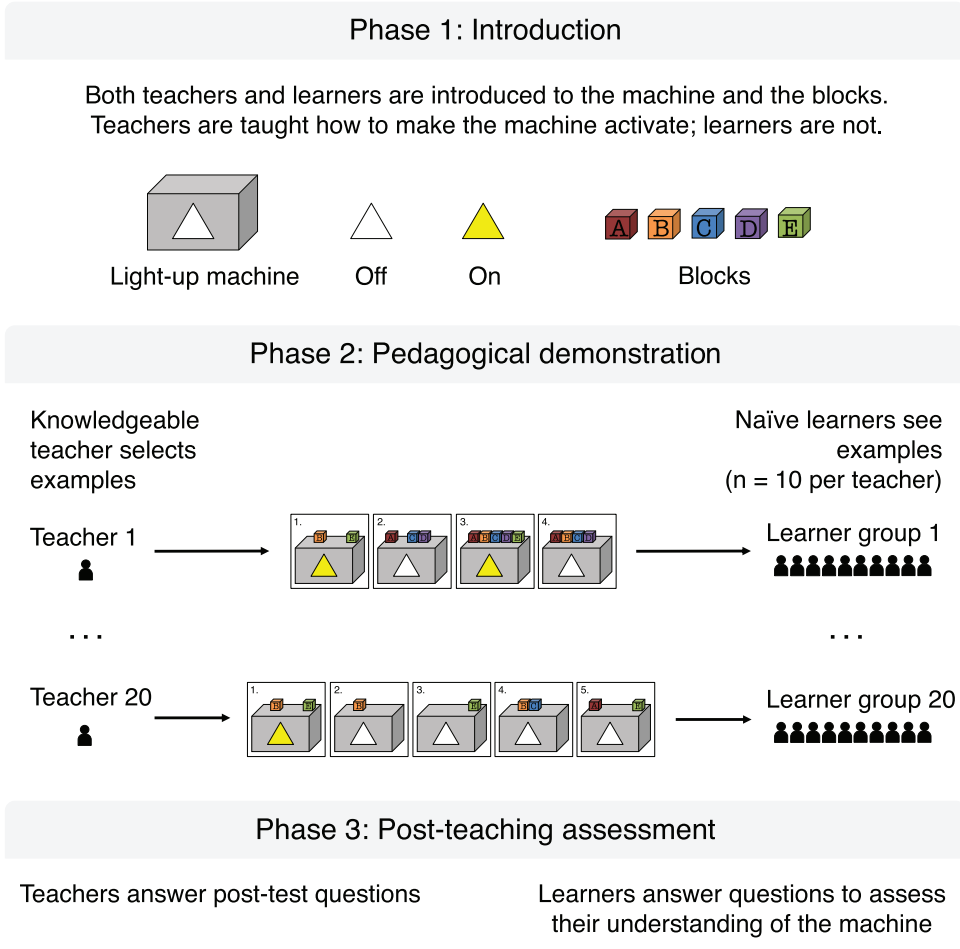

Fig. 1. Stimuli and experimental procedure. "Teacher" and "Learner" refer to the experimental conditions that participants were assigned to. All participants were introduced to the machine and the blocks and learned how to identify whether the machine was on or off, but only teachers were taught how the machine worked. Teachers then selected examples to communicate how the machine worked; learners observed these examples ($n = 10$ learners assigned to each teacher), and their understanding of the machine was assessed.

Fig. 1). When two particular blocks (B and E) were placed on top of the machine together, it activated (henceforth, referred to as the "B&E" rule). The presence or absence of other blocks did not affect the outcome. We chose a machine with a novel causal structure for the teaching task because prior research demonstrates that children and adults can often learn how such machines work from limited data (Gopnik, 2012; Griffiths et al., 2011; Kushnir et al., 2010). We selected the B&E rule because prior work shows that adults do not find a conjunctive activation rule trivially obvious (Gopnik et al., 2017; Lucas et al., 2014).

3.1.3. Procedure

Informed consent was obtained from all participants for all experiments reported in the manuscript and associated Supporting Information.

3.1.3.1. Teachersb: Participants assigned to the teacher condition were introduced to the light-up machine and were told how it worked (learning that both blocks B and E were necessary to activate the machine and that other blocks did not matter). To be included in the task, participants needed to pass an attention check question (selecting the number 2 on a Likert scale) and correctly identify which blocks were necessary to activate the machine.

To reinforce teachers' understanding of the machine, participants were asked to identify whether the machine was on or off in three scenarios (see OSF page for the full survey), and inaccurate responses were corrected. Participants then rated their confidence (on a Likert scale) that they understood how to activate the machine and could teach another person how it worked. To ensure that teachers understood the task, participants were asked to identify what their learners would know prior to seeing any examples (the correct answer was that learners would only know how to tell whether the machine was on or off), what the learners' goal was (the correct answer was that the learners' goal was to understand the machine as well as they did), and we ensured that teachers understood how to indicate whether they wanted to provide another example or whether they wanted to end the task. As before, incorrect responses were corrected.

Finally, teachers were asked to generate between three and 20 unique examples that would teach a naïve learner how the machine worked. Critically, teachers were explicitly told that learners would know nothing about how the machine worked but that their understanding of the machine would be tested after they saw teachers' examples. After selecting examples, participants were asked to rate their confidence that these examples would be effective in teaching another person how the machine worked. For completeness, we also asked teachers whether there were any examples they wanted to show but were not able to; and vice versa, whether they felt that they had provided more examples than necessary.

3.1.3.2. Learners: Participants assigned to the learner condition first learned that they were about to see a new machine and that another person had chosen examples to teach them how the machine worked (see Fig. 1). To be included in the task, participants needed to pass an attention check question (selecting the number 2 on a Likert scale, presented near the midpoint of the task; see OSF page for the full survey).

Learners first saw the light-up machine and the five blocks and learned that when the machine activated, the white triangle on the front of the machine lit up and turned yellow. Ten learners were assigned to view each teacher's examples; so after learning what the machine looked like when it activated, learners saw the set of examples their assigned teacher had generated. Each example was shown as an image. The blocks the teacher had chosen appeared on top of the machine, and the triangle indicated whether the machine was on or off (see Fig. 1). Accompanying text also described whether or not the blocks in each example caused the machine to turn on.

Finally, learners' understanding was assessed through a qualitative and a quantitative measure. For our qualitative measure, participants were asked to provide a written explanation of how the machine worked. For our quantitative measure, participants were asked to predict which block combinations would activate the machine (out of the 31 possible combinations). To produce a quantitative score for each participant, we sum the number of correct responses. The order of the 31 possible block combinations was initially randomized to ensure that trials did not appear as a structured sequence, and then presented to all participants in the same order. Learners also rated their confidence (on a Likert scale) that they understood how to activate the machine and for completeness indicated whether they thought their teacher had provided more examples than necessary (see Supporting Information).

3.2. Results

After the task introduction, teachers were both confident they understood how to activate the machine ($M = 6.75$, $SD = 0.44$, range = 6–7 on a 7-point scale), and that they could teach another person how the machine worked ($M = 6.6$, $SD = 0.68$, range = 5–7 on a 7-point scale), suggesting that the task introduction was clear. Although teachers were only required to provide a minimum of three unique examples, they produced an average of 7.5 examples ($SD = 4.39$, range = 3–20), with only two participants providing the minimum. This demonstrates that participants were motivated to teach and did not just put in the minimum effort required. Furthermore, participants felt that they were able to teach effectively within the constraints of the task, with only one participant indicating that there was an example they had been unable to show (this participant wanted to show blocks B and E arranged in a different order).

After selecting their examples, participants were confident they had taught well, judging that a naïve participant would successfully learn from their examples ($M = 6.05$, $SD = 0.76$, range = 5–7 on a 7-point scale). Seventy percent of teachers even thought they had provided more examples than necessary, judging that they had provided an average of 2.8 "extra" examples ($SD = 2.8$, range = 1–12).

Despite teachers' confidence that they had taught effectively and learners' confidence that they had figured out how the machine worked ($M = 5.3$, $SD = 1.57$, range = 1–7 on a 7-point scale), only 50% of learners ($n = 100$) performed at or near ceiling in the quantitative task, making no more than one mistake when identifying which block combinations activated the machine ($n = 88$ performing perfectly). While the remaining 50% of participants failed to learn the exact activation rule, they nonetheless showed evidence that they had partially learned how the machine worked. On average, these participants correctly predicted whether the machine would be activated in 70.6% of trials, performing significantly above chance ($t(99) = 17.58$; $p < .001$; see Fig. 2). Although it is possible that learner performance reflects only differences in individuals' motivation or attention, a regression revealed that this is not the case. Learners' performance is significantly predicted by the teacher they were assigned to learn from ($p = .03$; note that throughout, all regression p-values are obtained via permutation tests).

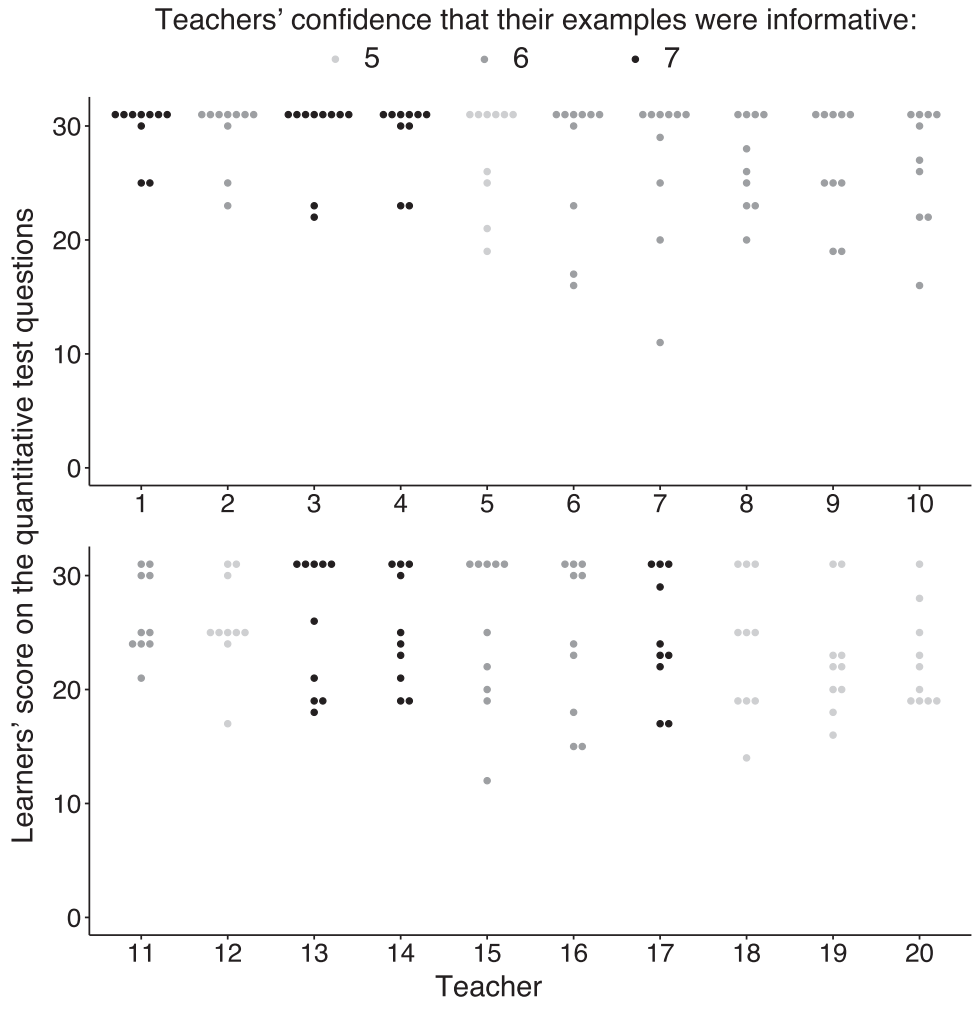

Fig. 2. Distribution of participants' quantitative scores (out of 31 questions), grouped by the teacher they learned from. Teachers are arranged by the mean number of questions their learners answered correctly, from highest (Teacher 1) to lowest (Teacher 20). Each point represents one learner. The color of each point indicates how confident teachers were that a naïve learner would understand how the machine worked from their examples, from lightest (least confident) to darkest (most confident). While learners performed above chance overall, there is still variation in their scores, demonstrating that teachers' data were not maximally informative for every learner.

To uncover the types of errors learners made, we examined their qualitative explanations. These explanations were coded into the following four categories (set a priori; see OSF page for the coding scheme): uninformative "other" explanations (e.g., "When the machine is on, the triangle turns to the yellow color"), correct explanations, incorrect explanations that focused on the "right kinds" of hypotheses (realizing that the machine was activated by certain blocks but failing to correctly identify the activation rule; e.g., "Certain letter combinations cause the machine to turn on"; "It's possible the presence of E turns it on and

all other letters do nothing [off]"), and incorrect explanations that focused on the "wrong kinds" of hypotheses (failing to identify the relevant feature of the task and referencing incorrect activation mechanisms; e.g., "If two or more blocks are placed on the machine, it turns on"; "The machine turns on when the blocks are places on opposite sides"). The first and second authors independently coded participants' explanations according to this coding scheme; inter-rater reliability was high (Cohen's $\kappa = 0.85$; $p < .001$).

Nine learners (4.5%) provided uninformative explanations, and 104 learners (52%) correctly described how to activate the machine. These learners were largely the ones who performed well on the quantitative task (86 performed perfectly, and six made one error). The remaining 87 participants (43.5%) produced incorrect explanations: 42 learners (21%) gave the "right kind" of explanation, understanding that the blocks mattered but failing to identify the right ones, and 45 (22.5%) gave the "wrong kind" of explanation, focusing on task-irrelevant features. For completeness, we also asked learners whether they felt their teacher had provided extra examples; see Supporting Information.

3.3. Model-based analysis of performance in the teaching condition

Despite teachers' confidence, many learners struggled to learn how the machine worked. Why were teachers' examples often ineffective? Did participants assigned to teach underestimate the breadth of possibilities learners considered (providing data that were effective only if a learner was considering a narrow hypothesis space)? Or did teachers form an accurate epistemic representation, but fail to effectively plan over it (struggling to identify useful examples)?

We analyzed teachers' examples by providing them as inputs to our computational model, using a uniform prior over all hypotheses.[3] Given a hypothesis space, the model computes how teachers' examples should affect learners' beliefs. If learners struggled because teachers provided confusing data, then the model should fail to infer how the machine works given any hypothesis space. However, if the model succeeds in inferring how the machine works given simple hypothesis spaces (but not complex ones), this would suggest that teachers misunderstood learners' hypothesis spaces and provided under-informative data.

We first evaluated teacher examples[4] using the simplest hypothesis space: a single-operator hypothesis space with a depth of 2 (see Computational Framework for an explanation of the parameters). Hypotheses contained in this space were either single blocks (e.g., B), or comprised two blocks, combined with an *AND* or an *OR* operator, (e.g., *AND*(B, E); 25 hypotheses total). The model inferred the correct rule for 75% of teachers ($n = 15$), placing over 95% of the posterior probability mass on the correct hypothesis (these results are identical when the mass threshold is decreased to 50%). For the remaining five participants, the model continued to place the highest posterior probability on the correct hypothesis (B & E; on average 27%), but other hypotheses were rated as equally or similarly plausible, preventing the right hypothesis from accruing a probability mass above 50%.

The model's success shows that teachers' data were informative given a constrained hypothesis space. This suggests that many learners did not consider a hypothesis space this constrained because their performance was substantially poorer than our model's: Only 50%

of learners performed at or near ceiling on the quantitative task, and only 52% were able to correctly explain how the machine worked.

To gauge the efficacy of teachers' data as hypothesis space complexity increased, we repeated the same analysis with each of the remaining hypothesis spaces. The model continued to succeed for the same 75% of teachers in all of the simple single-operator hypothesis spaces, for all depth values (max depth = 5). By contrast, in the smallest dual-operator space (a depth of 3 or 95 hypotheses total), the model inferred the activation rule for only 55% of teachers ($n = 11$). The remaining teachers produced examples that failed to rule out alternative hypotheses. At a depth of four (300 hypotheses), the model inferred the activation rule for only 25% of teachers ($n = 5$). At a depth of five and six (852 and 2222 hypotheses, respectively), the model did not infer the activation rule for any teachers, suggesting that learners did not consider hypothesis spaces this complex, as many learners did succeed in learning about the machine from teachers' examples. As before, while no other hypotheses were ever rated as *more* likely than the B&E rule, there continued to be too many hypotheses left that were consistent with teachers' data. Thus, while teachers' examples were sufficient for a modeled learner to infer the activation rule in simple hypothesis spaces, they were under-informative in more complex ones.

Importantly, our model inferences are insensitive to the size of a hypothesis space: Bayesian updating applies the same rule to all hypotheses without problem in finite hypothesis spaces like the ones we consider here, with size affecting only the time it takes to do so. Thus, according to our model, teachers' data were under-informative in complex hypothesis spaces not because they contained a greater number of hypotheses but rather because they contained a greater variety. This suggests that teachers may have misunderstood the breadth of learners' hypothesis spaces, selecting examples under an expectation that learners considered a simple set of hypotheses. However, it is also possible that our smaller hypothesis spaces were so simple that any set of examples would have enabled the model to learn how the machine worked. If this is the case, this would suggest that teachers' examples were not uniquely informative, and any random set of examples would have sufficed.

To test whether teachers' data were uniquely informative, we generated 10,000 random sequences of 20 examples (the maximum number teachers could provide in our task). To test whether teachers' examples were not only helpful, but given in a particularly informative order, we randomly reordered each teacher's examples (generating all possible permutations if there were less than 10,000 possible permutations, $n = 7$ examples or less, and sampling 10,000 permutations otherwise). We compared the randomly sampled and reordered examples to Because our model already failed to infer the true hypothesis from any teacher's examples given a dual-operator hypothesis space at a depth of five, we do not consider more complex hypothesis spaces.

Consistent with the possibility that teachers provided uniquely informative data, in the simplest hypothesis space, the model converged on the true hypothesis most quickly given teachers' original data as compared to the same data presented in a shuffled order. In turn, shuffled data caused the model to converge on the true hypothesis more quickly than randomly sampled data (see Fig. 3). The same pattern held for the rest of the single-operator hypothesis spaces. However, in our smallest dual-operator space (at a depth of three), the difference

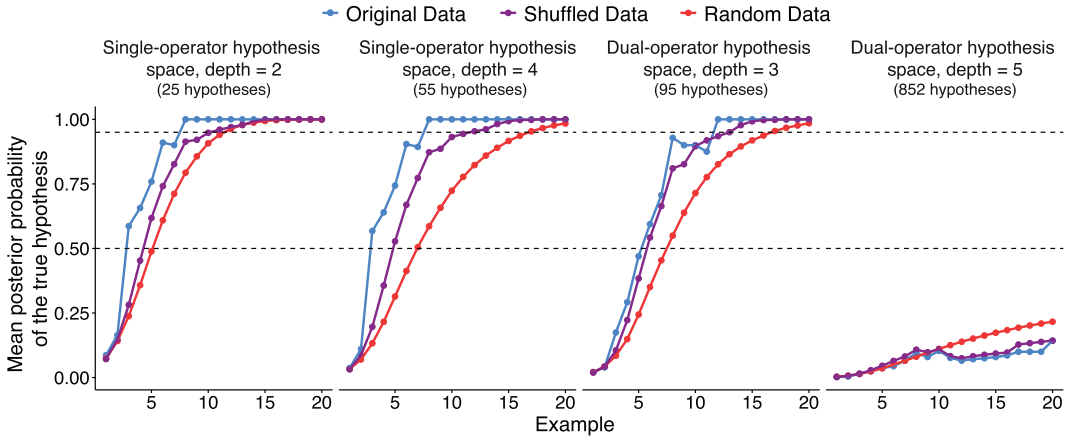

Fig. 3. Posterior probability mass placed on the true hypothesis (B&E) by the model as a function of teachers' examples. The posterior of the true hypothesis given teachers' original examples is plotted in blue; the posterior-given shuffled data are plotted in purple; and the posterior-given randomly generated examples are plotted in red. The dashed lines mark the 50% and 95% probability thresholds. In simpler hypothesis spaces, teachers' original examples are more informative than the same examples presented in a random order, which is more informative in turn than randomly generated examples. But as hypothesis spaces grow in complexity, teachers' examples break down, eventually becoming less informative than randomly generated examples (for a plot including intermediate hypothesis spaces not depicted here, see Supporting Information, Fig. S1).

between teachers' data in its original and shuffled order shrank; and as hypothesis spaces continued to increase in complexity, teachers' data ceased to be useful, ultimately becoming less informative than randomly sampled data.

These results support two conclusions: First, teachers not only provided informative examples but also gave them in a particularly informative order. Second, these results suggest that teachers generated their examples under an assumption that learners had a constrained hypothesis space: While teachers' data were uniquely well-suited when considered under simple, constrained hypothesis spaces, it became less informative than randomly generated examples when considered under richer hypothesis spaces.

3.4. Experiment 1 replications

We conducted two replications. First, to ensure that the results of Experiment 1 were robust, we conducted a direct replication. The results of our direct replication were qualitatively identical to that of the original experiment, showing that our results are reliable (see Supporting Information for details).

Second, although most teachers in Experiment 1 were satisfied with their teaching performance, one participant reported that there were examples they had wanted to show but had been unable to share due to the constraints of the online task. Did the online nature of the teaching task hinder participants' teaching performance? To test whether this is the case, we conducted a pre-registered replication where we recruited in-lab participants to complete the teaching task. In our original online teaching task, participants could select which blocks

they wanted to show but could not manipulate variables like order, orientation, or location. In our in-lab teacher replication, participants were introduced to an actual machine they could touch and interact with and provided examples by placing blocks directly on the machine. Participants were able to manipulate every feature of their demonstrations; each demonstration was then photographed and shown to learners (recruited from Amazon Mechanical Turk, as before).

If teachers recruited from Amazon Mechanical Turk were inattentive or were unable to show the kinds of examples they wanted, then teachers who participated in the lab should provide much more informative examples. Participants in the in-lab teacher condition were recruited from the campus community, and unlike the online participants, we were able to ensure they participated in a controlled environment (in a quiet room with no distractions). These participants did not provide a greater number of unique examples than those who participated online, giving an average of 7.45 examples ($SD = 3.71$, range = 3–16) with only one participant producing the minimum. Teachers were again confident that a naïve participant would successfully learn from their examples, with a mean confidence rating of 5.35 ($SD = 1.04$, range = 4–7 on a 7-point scale). Additionally, when asked why they had stopped providing examples, 19 of 20 participants explicitly said they had stopped when they felt their examples were sufficient (e.g., "I felt like I would have understood how the machine works at that point"; "More examples weren't necessary. It would have led to the learner getting confused"; "I thought I had covered all the important information"; see Supporting Information for full explanations).

Despite teachers' confidence, learners again struggled. Only 26.5% of learners ($n = 53$) performed at or near ceiling in the quantitative task ($n = 49$ performing perfectly), and only 29.5% ($n = 59$) correctly explained how the machine worked (see Supporting Information for full results and model-based analyses). This suggests that our results are not due to teaching constraints imposed by running the experiment online—and in fact, constraining the kinds of examples teachers could provide may have actually improved pedagogical outcomes.

3.5. Discussion

Experiment 1 aimed to distinguish between two potential explanations for why untrained adults fail in complex pedagogical tasks, testing whether adults fail to provide informative data, or whether they struggle to grasp the kinds of hypotheses learners could be considering. To do so, we introduced participants to a simple machine and taught them how it worked. Participants then chose examples to teach a naïve learner how the machine worked. After seeing examples generated by participants in the teaching condition, learners explained how they thought the machine worked and judged whether the machine was on or off for all possible block combinations. Across both our original experiment and two replications, participants in the teacher condition reported high confidence that naïve learners would understand how the machine worked after seeing their examples. But despite this confidence, many learners still failed to infer how the machine worked. Why might this be? According to our model, participants in the teaching condition produced data that were highly informative given simple hypothesis spaces but became progressively less helpful as hypothesis spaces increased in complexity. These results suggest that knowledgeable participants may have generated

examples under an assumption that learners would consider only a constrained set of hypotheses, whereas actual learners may have considered a broader spread of possibilities, thus rendering teachers' examples under-informative.

However, it is also worth considering whether our results could be explained by simpler task- or motivation-based alternatives. A first possibility is that participants in the teacher condition simply were not motivated to succeed in our task. Our data suggest that this was not the case: Most participants provided more examples than the minimum required, and participants recruited in the lab did not provide more examples than those who participated online. A second possibility is that the task design constrained participants in a way that prevented them from teaching successfully. This could explain teaching failures in our online tasks, where participants could only choose the identity of the blocks (but were unable to manipulate dimensions like order, position, or orientation). If this were the case, participants should have then succeeded in our in-lab replication, where participants could place the blocks in any way that they wanted. However, in-lab participants were actually less effective than our online teachers (only 26.5% of participants who learned from in-lab teachers performed at or near ceiling in the quantitative task). This suggests that our online task design cannot explain the failures of knowledgeable participants to teach well. In Experiment 2, we elicit knowledgeable agents' explicit judgments over naïve learners' beliefs, testing whether teachers' failures in Experiment 1 genuinely arose from a misunderstanding of learners' hypothesis spaces.

4. Experiment 2

Experiment 1 suggests that, when teaching, adults may misrepresent learners' beliefs, assuming that learners consider a more constrained hypothesis space than they actually do. To test this possibility, Experiment 2 directly asked knowledgeable and ignorant participants to rate the plausibility of different hypotheses about how the machine from Experiment 1 works. Knowledgeable participants received the same introduction as participants in Experiment 1's teaching condition, being told exactly how the machine worked. Naïve participants received the same introduction as learners in Experiment 1, learning what the machine looked like when it was on but not what made it turn on. To assess the impact of privileged knowledge, participants in both groups rated hypotheses about how the machine worked. Knowledgeable participants rated how likely a naïve agent would find each possibility; naïve participants rated how likely they themselves found each possibility. If participants in Experiment 1's teaching condition misrepresented the range of hypotheses naïve learners could find plausible, then knowledgeable and naïve participants' ratings should diverge in a systematic way.

4.1. Methods

4.1.1. Participants

Forty participants were recruited from Amazon's Mechanical Turk platform. The first 20 participants (mean age = 38.3; range = 22–64) were assigned to the knowledgeable condition, and the last 20 participants (mean age = 35.4; range = 23–60) were assigned to the naïve condition. Six additional participants were recruited in the knowledgeable condition but

were not included because they failed the inclusion question (identifying which blocks made the machine go).

4.1.2. Stimuli

Stimuli consisted of a list of hypotheses for how the machine works. We began with the true hypothesis (that B and E were both required to make the machine activate) and built the rest of the list based on incorrect explanations from the learner qualitative task in Experiment 1. A qualitative analysis of these responses suggested that explanations fell into seven categories: (a) placing certain blocks on the machine made it go, even if other blocks were also present; (b) placing certain blocks on the machine made it go only if other blocks were not also present; (c) placing certain blocks on the machine, in a specific order, made it go; (d) placing certain blocks on the machine made it go, but specific blocks could inhibit the machine from activating if also placed on top; (e) the machine only activated when the right number of blocks were placed on top; (f) the machine only activated when the right letters were on top (e.g., maybe the blocks needed to spell out a word) or when the blocks were in the right location on top; and (g) explanations coded as "other."

We generated at least one exemplar from each category, and generated more hypotheses from categories that were more common or with greater intrinsic variability, forming a list of 26 possible hypotheses (this list also contained the true hypothesis; see Supporting Information for full list).

4.1.3. Procedure

Participants in the knowledgeable condition saw the same introduction as participants in the teaching condition in Experiment 1, learning exactly how the light-up machine worked. Demonstrating that they had indeed learned how the machine worked, participants in the knowledgeable condition were confident they understood ($M = 6.75$, $SD = 0.44$, range = 6–7) and could teach how the machine worked ($M = 6.75$, $SD = 0.55$, range = 5–7). Participants in the naïve condition saw the same introduction as the learners in Experiment 1, learning how to identify whether the light-up machine was on or off but not learning which blocks made it go. As in Experiment 1, participants in the knowledgeable condition who could not identify which blocks made the machine go were excluded; participants in both conditions were also excluded if they did not pass our attention check measure (selecting the number 2 on a Likert scale).

After completing the introduction, participants from both conditions were presented with all 26 possible hypotheses, one at a time. The hypotheses were initially randomized so that they did not appear grouped by category but were presented to all participants in the same order. Participants in the knowledgeable condition were explicitly told "Before they see your examples, the other worker has no idea how the machine works," and were asked to rate the hypotheses from the perspective of a naïve learner, indicating how likely a naïve agent would find each hypothesis from 0 (very unlikely) to 100 (very likely). Participants in the naïve condition were asked to rate how likely they themselves found each possibility, along the same scale.

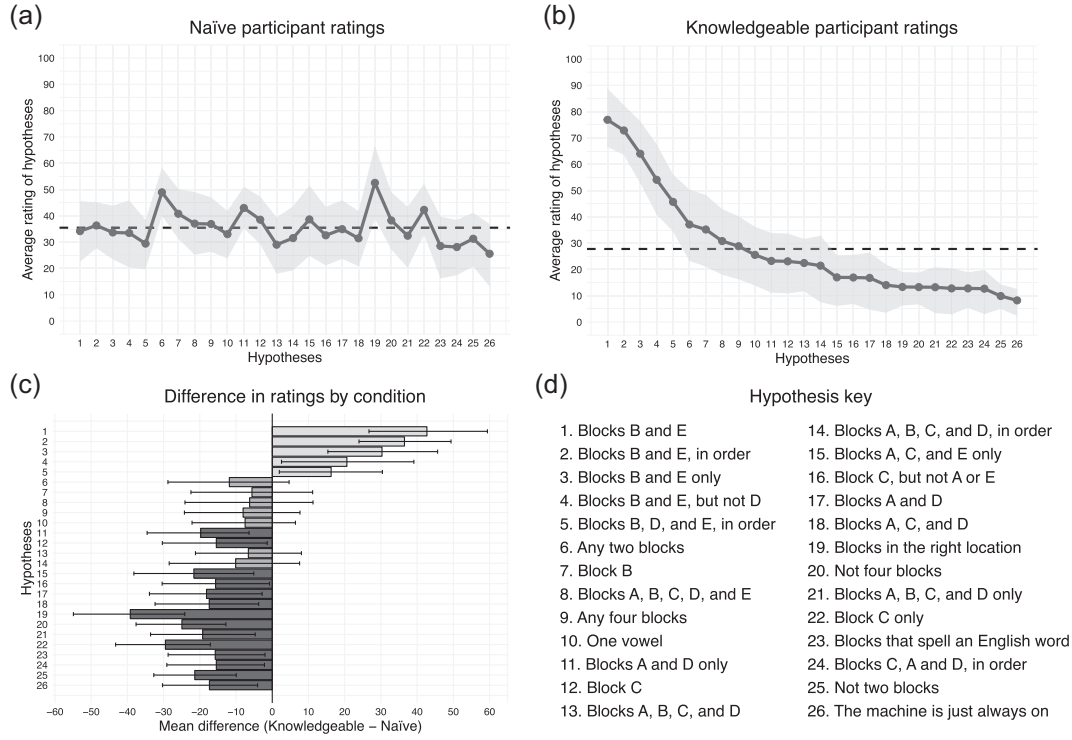

Fig. 4. (a-b) Participants' mean ratings of each hypothesis in the naïve and knowledgeable conditions, respectively (ordered by knowledgeable participants' ratings). The shaded areas indicate 95% bootstrapped confidence intervals. The dotted lines indicate the mean rating across all hypotheses by condition. While naïve participants rated all hypotheses save two as being equally likely, knowledgeable agents' ratings generally diverged from the condition mean. (c) Difference in mean rating of each hypothesis by condition. Positive values indicate that knowledgeable participants found a hypothesis more likely, and negative values indicate that naïve participants found a hypothesis more likely. The error bars are bootstrapped 95% confidence intervals. Bars are shaded according to whether knowledgeable participants found a hypothesis more likely (lightest values), whether naïve participants found a hypothesis more likely (darkest values), or whether there was no reliable difference between conditions. (d) Key for the 26 hypotheses presented in panels (a–c).

4.2. Results

Overall, participants in the naïve condition found all 26 hypotheses relatively unlikely ($M = 35.5$, $SD = 25.5$; see Fig. 4a). While participants in the knowledgeable condition correctly judged that naïve agents would find the hypotheses presented unlikely overall ($M = 27.8$, $SD = 30.3$; see Fig. 4b), the two rating distributions significantly differed ($p < .001$ by computing a permutation test over the between-condition sum of squared errors). Specifically, comparing the difference in participants' ratings by condition reveals that knowledgeable agents overestimated how likely naïve ones were to find the true hypothesis and closely related hypotheses, and underestimated how likely they were to find possibilities

farther from the truth (assessed by computing a 95% bootstrapped confidence interval over the difference between knowledgeable and naïve agents' ratings; see Fig. 4c).

4.3. Discussion

In Experiment 2, we tested if knowledgeable agents could correctly predict how likely naïve agents would find different hypotheses about a light-up machine. Our results show that, even when provided with a set of hypotheses to rate, knowledgeable agents could not evaluate them from the perspective of a naïve agent. Participants in the naïve condition found 24 of the 26 hypotheses equally plausible, while participants in the knowledgeable condition overestimated naive agents' belief in hypotheses that were similar to the true one and underestimated their belief in hypotheses distant from the truth.

Building upon our findings from Experiment 1, these results further suggest that participants in the teaching condition failed to realize that naïve learners were likely to consider hypotheses further from the truth. While these results do not reveal the exact hypotheses that each individual learner actually considered in Experiment 1, the hypotheses participants rated were generated based upon learners' incorrect explanations ($n = 96$ explanations in Experiment 1). Therefore, they are likely representative of other hypotheses that learners might have considered plausible (and that participants in the teaching condition dismissed, or even failed to consider in the first place). These results are consistent with our modeling analyses from Experiment 1, suggesting that knowledgeable participants did not anticipate the breadth of hypotheses learners might reasonably consider—and thus that their failure to teach was, in part, caused by a failure to consider or address these alternate possibilities when providing examples.

5. Experiment 3

Experiments 1–2 suggest that untrained adults tend to fail at teaching causal relations because they provide examples under an (erroneous) expectation that learners consider only the right kinds of hypotheses. If this is the case, then learning performance should improve if learners' hypothesis spaces are constrained to match the expectations of participants assigned to teach. Experiment 3 tests this possibility, constraining learners' beliefs by first showing them how to activate two other light-up machines. We hypothesized that this would help learners realize that only block identity (and not other features such as order or orientation) determined whether a machine would activate, helping them align their hypothesis space to knowledgeable participants' expectations.

5.1. Methods

5.1.1. Participants

Two hundred participants were recruited from Amazon Mechanical Turk (mean age = 34.4; range = 18–66). One additional participant was recruited and replaced because they failed an attention check.

5.1.2. Stimuli

Stimuli consisted of three light-up machines: the original machine and two additional training machines. Each machine was of a different size and color to the original but clearly belonged to the same category (for pictures see Supporting Information, Fig. 2). The first training machine was presented with three blocks, L, M, and N. Only block M was required to make the machine activate. The second training machine was presented with six blocks, Q, R, S, T, U, and V. Blocks Q, T, and V were all required to make the machine activate. As with the original machine, the presence of additional blocks did not affect whether the training machines activated, and neither did the order of the blocks.

5.1.3. Procedure

Participants were shown an image of the two training machines and the original light-up machine side by side. Participants were told that they would read explanations about how the first two machines worked, and they then would learn how the third (original) machine worked by seeing examples another participant had selected for them.

As in Experiment 1, participants first learned that when a light-up machine activated, the white triangle on the front of the machine lit up and turned yellow. Then participants were introduced to the first training light-up machine and its three corresponding blocks, L, M, and N. They were explicitly told that the machine only turned on when block M was placed on top, that other blocks did not matter, and that the order of the blocks did not matter. To illustrate, they saw two examples of the machine activating: in the first example, only block M was on top, and in the second example, all three blocks were on top. To assess whether participants understood how the machine worked, they were asked to identify which block made the machine activate, and whether the machine was on or off in two examples (note that these and the following memory check questions did not serve as inclusion criteria).

Next, participants were introduced to the second training light-up machine and its six corresponding blocks, Q, R, S, T, U, and V. They were told that this machine only turned on when blocks Q, T, and V were all on top and that other blocks and the order of blocks did not matter. To illustrate, they saw two examples of the machine activating. In the first example, only blocks Q, T, and V were on top; in the second, all six blocks were on top. Participants again answered three questions about the machine, identifying which blocks made the machine activate and whether the machine was on or off in two examples.

Finally, participants were shown the third (original) light-up machine and its five corresponding blocks, A, B, C, D, and E. By introducing all three machines as exemplars of the same "light-up machine" category, we expected participants to infer that all of the machines worked the same way (because adults readily use category knowledge to infer properties of new exemplars; Osherson, Smith, Wilkie, Lopez, & Shafir, 1990). Additionally, after introducing the third light-up machine, participants were explicitly told that their goal was to figure out which block(s) were responsible for making the machine activate, and which block(s) did not matter. After being reminded again that they were about to see examples selected to teach them how the machine works, participants proceeded to see a set of examples selected by a participant assigned to teach in Experiment 1.

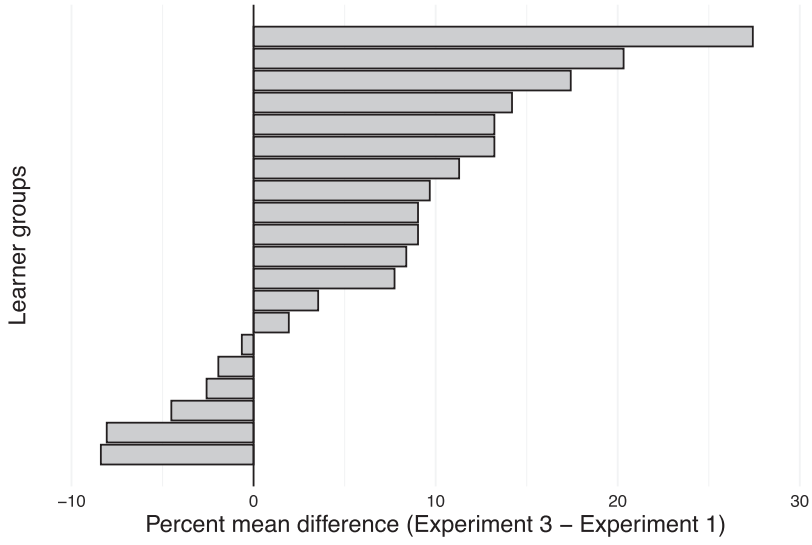

Fig. 5. Difference in learners' performance on the quantitative test questions between Experiments 1 and 3. Each bar represents the learners assigned to one participant from the teaching condition in Experiment 1. The bars are arranged in order from greatest to least improvement in Experiment 3. Positive values indicate that learners in Experiment 3 performed better on the quantitative measure; negative values indicate that learners in Experiment 1 performed better.

After seeing these examples, participants' understanding was assessed as in Experiment 1: They were asked to explain how the machine worked, indicate their confidence that they understood how to activate the machine on a Likert scale, and answered the same 31 quantitative test questions (indicating whether the machine was on or off for all 31 possible combinations of blocks). As before, participants were excluded if they did not pass our attention check measure (selecting the number 2 on a Likert scale).

5.2. Results

Despite receiving the exact same teacher-generated examples as learners in Experiment 1, learners in the current experiment were significantly more confident they understood how the machine worked ($M = 6.2$, $SD = 1.09$, range = 1–7 on a 7-point scale; $\beta = 0.87$, $p < .001$). This confidence was justified: They also performed significantly better on the quantitative test questions ($M = 92.1\%$, $SD = 15.7$; $\beta = 2.18$, $p < .001$; see Fig. 5). These results confirm our model-based analyses, suggesting that teachers' data were highly informative for learners with a constrained hypothesis space. Thus, previous learning failures may have stemmed from teachers' poor representation of learners' initial beliefs.

To isolate the source of learners' improvements, we next analyzed their qualitative explanations. As before, the first and second authors independently coded learners' qualitative explanations (Cohen's $\kappa = 0.95$; $p < .001$). Four participants provided uninformative explanations and were therefore excluded from qualitative analyses. Confirming that our

manipulation was successful, fewer participants provided explanations focusing on the wrong kinds of hypotheses (Experiment 1, $n = 45/200$; Experiment 3, $n = 7/200$). A two-proportion z-test demonstrated that this difference was significant, $p < .001$. Instead, validating our model-based analyses, more participants provided correct explanations (Experiment 1, $n = 104/200$; Experiment 3, $n = 150/200$). Again, a two-proportion z-test demonstrated that this difference was significant, $p < .001$). Last, there was no difference in the number of participants who produced the "right kind" of explanation, understanding that the identity of the blocks mattered but failing to identify the right blocks in their explanations (Experiment 1, $n = 42/200$; Experiment 3, $n = 39/200$; two-proportion z-test, $p = .80$). This result suggests that our manipulation was effective in constraining learners' hypothesis spaces, leading to improvement in learners' performance.

Finally, consistent with the predictions of our account, the efficacy of a teacher's examples (calculated as the mean number of test questions their learners answered correctly in Experiment 3 minus Experiment 1) is marginally predicted by the proportion of their learners whose explanations referenced the wrong kinds of hypotheses in Experiment 1 ($\beta = 0.77$; $p = .0545$). This suggests that even teachers whose learners were most off-base in Experiment 1 provided data that were informative to learners with constrained hypothesis spaces.

5.3. Discussion

Taken together, Experiments 1 and 2 suggest that participants in the teaching condition provided data under an assumption that learners consider constrained hypothesis spaces (prioritizing hypotheses similar to the truth). Thus, knowledgeable participants' data may have been informative for learners considering a more constrained hypothesis space but uninformative otherwise. In the current experiment, we directly test whether constraining learners' hypothesis spaces render teachers' data more informative. Consistent with the results of Experiments 1 and 2, we find that this is the case: Learners were both more accurate on our quantitative test questions and more likely to produce correct qualitative explanations of how the machine worked.

While we chose to manipulate learners' beliefs to match those of participants assigned to teach, our hypothesis predicts that learning outcomes should also improve if we aligned teachers' beliefs to match those of their learners. Doing so is difficult for two main reasons. First, to obtain a stable measure of teaching efficacy, each teacher's data had to be evaluated against multiple learners; it would have been difficult to cleanly ensure that each participant in the teaching condition had an accurate grasp of all 10 learners' hypothesis spaces. Second, learners likely did not come into the task with a pre-specified hypothesis space but rather generated and discarded hypotheses as the task progressed (and we return to this point in the General Discussion). This makes it difficult to ensure that teachers always have an accurate grasp of learners' expectations. Nonetheless, the fact that aligning learners' expectations to those of teachers significantly improved performance suggests that teachers' data were indeed informative but not adequately tailored to fully naïve learners' beliefs.

6. General discussion

Despite the ubiquity and early developmental origins of informal pedagogy (Gweon, 2021), adults often struggle to share information effectively (e.g., Chi et al., 2004; Graesser et al., 1995; Hinds, 1999; Siler & VanLehn, 2015). In this paper, we sought to better understand adults' abilities to informally share information. Across three experiments (and two replications), we investigated why untrained adults teach well in some tasks but not others, testing whether adults struggle to grasp the breadth of hypotheses naïve learners could be considering, or whether they struggle to decide what data will be most informative to share. In Experiment 1, we introduced a teaching paradigm that was simple enough to be analyzed quantitatively but complex enough that it elicited teaching failures. Adults assigned to the teaching condition were given an opportunity to select a sequence of examples to reveal how a novel machine worked. Despite these participants' confidence that their examples would enable a naïve participant to learn how the machine worked, participants presented with these examples often failed to uncover how the machine worked. To investigate the source of these failures, we implemented a computational model of rational teaching (Shafto et al., 2014) that evaluated the quality of untrained teachers' examples under different learner hypothesis spaces.

By varying the complexity of potential learner hypothesis spaces, our model revealed that knowledgeable adults can generate highly informative examples tailored to a learner considering a simple hypothesis space. In this case, participants' examples revealed how the machine worked more effectively than a random series of examples, and most impressively, better than the same examples presented in a random order. These results suggest that adults are proficient at sharing information in a way that most effectively scaffolds learning. At the same time, our model also revealed that knowledgeable participants' examples ceased to be useful under more complex hypothesis spaces. Thus, participants might have failed to teach effectively because their learners were considering a richer set of hypotheses than what they expected.

In Experiment 2, we found a direct mismatch between naïve agents' beliefs and knowledgeable agents' representations of those beliefs. Consistent with Experiment 1, knowledgeable participants assumed that naïve agents would find hypotheses close to the truth to be more plausible, whereas, in reality, naïve participants found all hypotheses to be comparably likely. Critically, the hypotheses that participants rated represented the types of explanations that learners considered in the actual task (as they were generated based upon learners' explanations in Experiment 1). These results further suggest that people's difficulty teaching in Experiment 1 arose from a failure to consider the wider range of hypotheses that learners considered plausible—and not due to an inability to generate informative examples.

Finally, Experiment 3 confirmed that knowledgeable participants' examples were informative when given to a naïve learner whose beliefs matched their expectations. Specifically, when learners' hypothesis spaces were constrained, they learned more effectively from examples generated by adults in the teacher condition. Taken together, these results suggest that adults' failures to effectively teach in our task were not due to a poor capacity to generate

informative data but rather due to a poor understanding of the kinds of beliefs learners find likely.

Combined, our work has four main findings. First, our work shows that adults engaged in informal teaching do not struggle to select helpful pieces of information, even in situations where the space of possible examples is large. Participants assigned to the teaching condition were highly informative (both in the examples they selected and the order in which they presented them), and their examples were useful for learners with constrained hypothesis spaces. These results demonstrate that failures to teach in our task emerged because knowledgeable participants assigned to teach underestimated the richness of learners' hypothesis spaces. Although humans share our knowledge ubiquitously (and generally effectively), our results suggest that we may sometimes struggle to gauge what naïve others could believe—and thus fail to successfully communicate our knowledge.

Second, our results shed light on how privileged knowledge affects pedagogical performance. While prior research has found that knowledgeable agents are often affected by a *curse of knowledge* (wherein their ability to reason about naïve minds is impaired by their own privileged knowledge; Camerer, Loewenstein, & Weber, 1989; Hinds, 1999; Nickerson, 1999; Nickerson, Baddeley, & Freeman, 1987), this phenomenon is broad in scope (from formal pedagogy to simple everyday communication) and can occur in one of two ways. A first possibility is that the curse of knowledge is a form of *representational leak*, whereby agents struggle to separate their own knowledge from their representations of other people's knowledge (i.e., assuming that others know what they know). A second possibility is that the curse of knowledge impairs agents' ability to predict and track what others believe, without necessarily confusing others' knowledge for their own (i.e., assuming their instruction helps learners rule out alternatives more effectively than it actually does). Our results point toward the first possibility: Participants assigned to teach had privileged knowledge, and this affected their ability to estimate what a naïve person would find plausible. Because knowledgeable participants based their examples upon inaccurate representations of learners' beliefs, most teachers' examples were appropriately informative as long as learners' beliefs were also constrained (e.g., in Experiment 3). This suggests that participants assigned to teach were biased by their own knowledge when initially estimating what hypotheses a learner might find plausible but were able to accurately track how their examples would affect a learner's beliefs.

Third, our work suggests that adults engaging in informal pedagogy can track how their examples affect learners' beliefs. Participants in Experiment 1 generated highly informative pedagogical data with an informative temporal structure (i.e., providing the right examples in the right order). This ability is particularly impressive when considering that people had to select multiple examples sequentially, could not review past examples, and were unable to engage with learners in real time to better understand their knowledge state.

Finally, our results also highlight the importance of teaching as a professional practice that requires years of training. Although even young children have remarkably sophisticated intuitions about pedagogical principles (see Gweon, 2019, for review), informal pedagogy appears to be most effective in situations where learners have a relatively constrained hypothesis space. Therefore, as society builds increasingly complex bodies of knowledge, the value of trained educators becomes even more important. For instance, trained teachers may have

greater experience engaging in metacognition (thinking about thinking; Flavell, 1979) over their own knowledge and pedagogical process and in considering how to encourage the same in learners (e.g., Blakey & Spence, 1990; Pintrich, 2002; Wilson & Bai, 2010). In particular, repeated and extensive experience with students might help professional educators gain more precise representations of naïve learners' beliefs (e.g., Herppich, Wittwer, Nückles, & Renkl, 2013).

6.1. Study limitations

Our work has several limitations. First, our study explored informal pedagogy using *blicket detectors*—opaque machines used to study how people learn causal relations from statistical information (Bonawitz & Lombrozo, 2012; Gopnik, Sobel, Schulz, & Glymour, 2001; Sobel & Kirkham, 2007). Critically, blicket detectors isolate causal reasoning by making the underlying mechanism unknown. We therefore do not know how our findings would change when people teach about a system with a non-opaque mechanism. It is possible that learners in our task had a surprisingly broad hypothesis space because they lacked mechanistic information to constrain it with. If so, then people might succeed in informal pedagogy more often than our task suggests because learners may often have narrow hypothesis spaces constrained by the underlying mechanism. Nonetheless, our work still shows that, in the absence of mechanistic information, people have trouble representing what a naïve learner might consider plausible, therefore failing to share information effectively.

A related limitation is that our task focused on a single learning rule, where the relevant causal relation relied upon the logical operator "AND": Both blocks B *and* E were needed to make the machine activate. Thus, we do not know to what extent pedagogical success might change under different activation rules. Note however, that participants failed to teach because they did not realize naïve learners might consider radically different types of hypotheses (such as "a combination of blocks that could form a word" or "blocks [that] are far apart"; see Experiments 1 and 2). Therefore, our results would only change if a different activation rule helped untrained teachers become aware of the broader hypothesis space that learners might consider. This is a direction we hope future work will explore.

A third limitation is that our work required us to formalize the relative complexity of different hypothesis spaces. To do so, we used expression length as a proxy for complexity—a common approach in computational cognitive science (e.g., Goodman et al., 2008; Piantadosi et al., 2012; Velez-Ginorio et al., 2017) However, the complexity depends on the system's compositional primitives: For instance, in our model, expressing the hypothesis that any one block might make the machine activate requires one operator and five elements (*OR*(A, B, C, D, E)). This hypothesis is considered more complex than related hypotheses (such as the hypothesis that either B or E make the toy go; *OR*(B, E)). But if our model included the operator "any," then the reverse would be true (*ANY*() contains fewer elements and would be simpler than *OR*(B, E)). While this is a fundamental problem in cognitive science (Goodman, 1983), future work should explore the relative perceived complexity of different hypothesis spaces in these learner contexts. Nevertheless, because our hypothesis spaces scale logically, more complex hypothesis spaces contain all hypotheses from the preceding spaces. Thus, no

matter which hypotheses we code as "simple" or "complex," our key result—that knowledgeable adults engaged in pedagogy provide insufficient evidence for learners considering varied hypotheses—still stands.

One final limitation is that our work did not study how the processes behind people's failures to teach might appear in other domains. Specifically, people's failure to select disconfirmatory evidence for others is reminiscent of a failure to select disconfirmatory evidence in first-person tasks (where adults often explore examples that confirm their hypotheses but not those that have the potential to disconfirm them; e.g., Wason, 1960). One critical difference, however, is that people in first-person tasks have access to their own beliefs, while in pedagogy, people must act under a representation of what they think others think. It is unclear to what extent shared computations underlie an ability to: (a) discover a rule firsthand and (b) communicate a rule to others (without first having to discover it oneself and without having firsthand access to learners' beliefs). Future research should investigate these questions.

6.2. Open questions

Our findings open several additional questions for future work. First, we do not know how knowledgeable agents form their beliefs about a naïve agent's hypothesis spaceosf. One possibility is that knowledgeable agents construct a space of possibilities by taking the right answer (in our case the B and E rule), and then modifying it to generate a set of plausible alternatives. This would explain why participants assigned to teach in our task expected learners to prioritize hypotheses close to the true one. If so, this would imply that our grasp of less informed minds could be fundamentally skewed by the contents of our own. By better understanding how knowledgeable agents come to represent naïve minds, future work can investigate how to remedy this curse.

However, it is an open question how to best correct the beliefs of knowledgeable adults asked to informally teach. Perhaps it is enough to simply tell knowledgeable adults the hypotheses a learner is considering; or perhaps adults also need to be explicitly told how a learner weights these hypotheses, being shown that a learner still places considerable weight on hypotheses "farther" from the truth. It is also unclear whether individual-level learner assessments are needed. Perhaps, simply being told what learners believe in general could help teachers understand the breadth of hypotheses learners may consider, or perhaps only individual-level information is helpful. This could vary as a function of domain: in cases where learners' hypothesis spaces are largely unique, individual-level information might be required, and in cases where learners tend to consider similar hypotheses, information about the average learner could be equally helpful. These are questions we hope to address in future work.

A related open question is how knowledgeable and naïve agents' hypothesis spaces change throughout a pedagogical interaction. Our computational model was designed to capture pedagogical interactions at a computational level of analysis (Marr, 1982). As such, our model used static and pre-determined hypothesis spaces that represented the full space of alternatives learners could consider throughout a pedagogical interaction. Intuitively, however, learners are likely to actively generate hypotheses in response to the data they receive from teachers (Schulz, 2012). The question of how learners actively build their hypothesis spaces is

beyond the scope of this work but is an important open question for understanding how adults can sometimes not only fail to teach effectively but inadvertently mislead learners (by guiding them to generate incorrect hypotheses that they did not have prior to the interaction). Nonetheless, our findings still demonstrate that even in minimally complex tasks, untrained teachers can generate informative data but fail to consider hypotheses learners find plausible.

Finally, we manipulated hypothesis space complexity in our model by increasing the number and kind of hypotheses a space could contain. However, it is also possible to scale complexity by varying not the specific hypotheses included in a space but rather the degree of plausibility placed upon these hypotheses. Such hypothesis spaces could better capture our finding that knowledgeable adults are willing to believe learners might consider hypotheses further from the truth, but simply expect learners to find these hypotheses unlikely (Experiment 2). Future work should address this possibility, implementing models where the prior distribution over a space of hypotheses is manipulated, rather than the nature of the hypotheses themselves (or a combination of the two).

6.3. Conclusion

Even from the first years of life, humans teach effectively, considering what others know or want and sharing the information we think they need (Gweon, 2021). Yet, while we excel in constrained teaching tasks from childhood (Rhodes et al., 2010; Shafto et al., 2014), even adults fail to effectively share information in more naturalistic tasks (e.g., Chi et al., 2004). Across three studies, we find that when knowledgeable adults teach, they often fail to consider the breadth of hypotheses a naïve learner may be considering. Although knowledgeable adults provide informative data when they teach (as assessed by a computational model of a rational teacher), they do not provide enough information for some naïve participants to learn. Our results unify prior findings, suggesting that adults should successfully share their knowledge in tasks where the kinds of hypotheses learners can consider are relatively constrained, but struggle in more naturalistic tasks where learners can consider varied possibilities. Because most real-world pedagogy occurs in naturalistic, unconstrained settings, these findings suggest that when sharing knowledge, we could all benefit from putting more effort into gauging learners' beliefs—or constraining them.

Acknowledgments

We thank the members of the Yale Psychology Department for helpful discussion and feedback. We thank Victor Hunt for help with stimuli design and data collection. This material is based upon work supported by the Center for Brains, Minds, and Machines (CBMM), funded by NSF-STC award CCF-1231216.

Open Research Badges

 This article has earned Open Data and Open Materials badges. Data are available at osf.io/pg9zy/ and materials are available at https://osf.io/pg9zy/.

Notes

1 We do not consider the hypothesis space that consists only of hypotheses with no logical primitives (A, B, C, D, E), because it does not contain the true hypothesis *AND*(B,E) used in our task.
2 Note that at a depth of 2, hypotheses can contain only up to two primitives. Two primitives can only be combined in one way: with either an *AND* or an *OR* operator. That is, participants can hypothesize that either A or B make the machine activate (*OR*(A, B)), or that both A and B are required (*AND*(A, B)), but there is no way to add a second logical operator to either of these hypotheses. Only at a depth of 3 can multiple operators be combined in one hypothesis. Thus, at a depth of 2, there is no difference between the single-operator and dual-operator hypotheses spaces.
3 A reasonable alternative would have been to implement a "simplicity prior" to penalize complex hypotheses. We opted not to do this because in our approach, each hypothesis space is meant to represent what learners might consider plausible (testing data quality across a range of nested hypothesis spaces). A simplicity prior would, a priori, deem some hypotheses to be implausible.
4 Teachers were instructed not to provide duplicate examples, but despite this, some participants produced duplicates (generally, by providing the example "B&E" more than once). Any duplicates were removed before providing the data as an input to the model, but the full data (duplicates included) are available in the OSF project page.

References

Aboody, R., Zhou, C., Flowers, M., & Jara-Ettinger, J. (2019). Ignorance = doing what is reasonable: Children expect ignorant agents to act based on prior knowledge. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, Montreal, Canada.

Aboody, R., Huey, H., & Jara-Ettinger, J. (2022). Preschoolers decide who is knowledgeable, who to inform, and who to trust via a causal understanding of how knowledge relates to action. Cognition, 228, 105212.

Aboody, R., Zhou, C., & Jara-Ettinger, J. (2021). In pursuit of knowledge: Preschoolers expect agents to weigh information gain and information cost when deciding whether to explore. *Child Development*, *92*(5), 1919–1931.

Baer, C., & Friedman, O. (2018). Fitting the message to the listener: Children selectively mention general and specific facts. *Child Development*, *89*(2), 461–475.

Benton, D. T., & Rakison, D. H. (2020). Computational modeling of backwards-blocking reasoning in human adults. Accessed date 11-1-2022. Available at: https://doi.org/10.31234/osf.io/xq8ws

Bonawitz, E. B., & Lombrozo, T. (2012). Occam's rattle: Children's use of simplicity and probability to constrain inference. *Developmental Psychology*, *48*(4), 1156.

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120*(3), 322–330.

Blakey, E., & Spence, S. (1990). *Developing metacognition*. Syracuse, NY: ERIC Clearinghouse on Information and Technology.

Bridgers, S., Jara-Ettinger, J., & Gweon, H. (2019). Young children consider the expected utility of others' learning to decide what to teach. *Nature Human Behaviour*, *4*(2), 144–152.

Bromme, R., Brummernhenrich, B., Becker, B. M., & Jucks, R. (2012). The effects of politeness-related instruction on medical tutoring. *Communication Education*, *61*(4), 358–379.

Camerer, C., Loewenstein, G., & Weber, M. (1989). The curse of knowledge in economic settings: An experimental analysis. *Journal of Political Economy*, *97*(5), 1232–1254.

Chi, M. T., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science*, *25*(4), 471–533.

Chi, M. T., Siler, S. A., & Jeong, H. (2004). Can tutors monitor students' understanding accurately? *Cognition and Instruction*, *22*(3), 363–387.

Einav, S., & Robinson, E. J. (2011). When being right is not enough: Four-year-olds distinguish knowledgeable informants from merely accurate informants. *Psychological Science*, *22*(10), 1250–1253.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, *34*(10), 906.

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

Gelpi, R., Prystawski, B., Lucas, C. G., & Buchsbaum, D. (2020). Incremental hypothesis revision in causal reasoning across development. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 974–980). Cognitive Science Society.

Goodman, N. (1983). *Fact, fiction, and forecast*. Cambridge, MA: Harvard University Press.

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

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

Gopnik, A. (2012). Scientific thinking in young children: Theoretical advances, empirical research, and policy implications. *Science*, *337*(6102), 1623–1627.

Gopnik, A., O'Grady, S., Lucas, C. G., Griffiths, T. L., Wente, A., Bridgers, S., Bridgers, S., & Dahl, R. E. (2017). Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. *Proceedings of the National Academy of Sciences*, *114*(30), 7892–7899.

Gopnik, A., Sobel, D. M., Schulz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: Two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology*, *37*(5), 620.

Graesser, A. C., Person, N. K., & Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology*, *9*(6), 495–522.

Griffiths, T. L., Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2011). Bayes and blickets: Effects of knowledge on causal induction in children and adults. *Cognitive Science*, *35*(8), 1407–1455.

Gweon, H. (2021). Inferential social learning: Cognitive foundations of human social learning and teaching. *Trends in Cognitive Sciences*, *25*(10), 896–910.

Gweon, H. (2019). Understanding others to learn and to inform: Foundations of distinctively human social learning. In S. Grimm (Ed.), *Varieties of understanding: New perspectives from philosophy, psychology, and theology*. Oxford, England: Oxford University Press.

Gweon, H., & Asaba, M. (2017). Order matters: Children's evaluation of underinformative teachers depends on context. *Child Development*, *89*(3), e278–e292.

Gweon, H., & Schulz, L. (2018). From exploration to instruction: Children learn from exploration and tailor their demonstrations to observers' goals and competence. *Child Development*, *90*(1), e148–e164.https://doi.org/10.1111/cdev.13059

Gweon, H., Shafto, P., & Schulz, L. (2018). Development of children's sensitivity to overinformativeness in learning and teaching. *Developmental Psychology*, *54*(11), 2113.

Herbst, E., Lucas, C. G., & Buchsbaum, D. (2017). Investigating the explore/exploit trade-off in adult causal inferences. In G. Gunzelman, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th annual conference of the cognitive science society* (pp. 501–506). London, UK: Cognitive Science Society.

Herppich, S., Wittwer, J., Nückles, M., & Renkl, A. (2013). Does it make a difference? Investigating the assessment accuracy of teacher tutors and student tutors. *The Journal of Experimental Education*, *81*(2), 242–260.

Hinds, P. J. (1999). The curse of expertise: The effects of expertise and debiasing methods on prediction of novice performance. *Journal of Experimental Psychology: Applied*, *5*(2), 205.

Ho, M. K., Littman, M., MacGlashan, J., Cushman, F., & Austerweil, J. L. (2016). Showing versus doing: Teaching by demonstration. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., & Garnett, R., editors, Advances in Neural Information Processing Systems 29, 3027–3035. Curran Associates,Inc.

Holyoak, K. J., & Cheng, P. W. (2011). Causal learning and inference as a rational process: The new synthesis. *Annual Review of Psychology*, *62*(1), 135–163.

Jara-Ettinger, J., Floyd, S., Tenenbaum, J. B., & Schulz, L. E. (2017). Children understand that agents maximize expected utilities. *Journal of Experimental Psychology: General*, *146*(11), 1574.

Jin, L., Schwartz, L., Doshi-Velez, F., Miller, T., & Schuler, W. (2021). Depth-bounding statistical PCFG induction as a model of human grammar acquisition. *Computational Linguistics*, *47*(1), 181–218.

Kushnir, T., Gopnik, A., Lucas, C., & Schulz, L. (2010). Inferring hidden causal structure. *Cognitive Science*, *34*(1), 148–160.

Lucas, C. G., Bridgers, S., Griffiths, T. L., & Gopnik, A. (2014). When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. *Cognition*, *131*(2), 284–299.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Cambridge, MA: MIT Press.

Nickerson, R. S. (1999). How we know—and sometimes misjudge—what others know: Imputing one's own knowledge to others. *Psychological Bulletin*, *125*(6), 737.

Nickerson, R. S., Baddeley, A., & Freeman, B. (1987). Are people's estimates of what other people know influenced by what they themselves know? *Acta Psychologica*, *64*(3), 245–259.

Olson, D. R., & Bruner, J. S. (1996). Folk psychology and folk pedagogy. In D. R. Olson & N. Torrance (Eds.), *The handbook of education and human development*. Cambridge, MA: Blackwell Publishers Ltd.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, *97*(2), 185.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2012). Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition*, *123*(2), 199–217.

Pintrich, P. R. (2002). The role of metacognitive knowledge in learning, teaching, and assessing. *Theory into Practice*, *41*(4), 219–225.

Rhodes, M., Bonawitz, E., Shafto, P., Chen, A., & Caglar, L. (2015). Controlling the message: Preschoolers' use of information to teach and deceive others. *Frontiers in Psychology*, *6*, 867.

Rhodes, M., Gelman, S. A., & Brickman, D. (2010). Children's attention to sample composition in learning, teaching and discovery. *Developmental Science*, *13*(3), 421–429.

Ronfard, S., & Corriveau, K. H. (2016). Teaching and preschoolers' ability to infer knowledge from mistakes. *Journal of Experimental Child Psychology*, *150*, 87–98.

Schulz, L. (2012). Finding new facts; thinking new thoughts. *Advances in Child Development and Behavior*, *43*, 269–294.

Shafto, P., & Goodman, N. (2008). Teaching games: Statistical sampling assumptions for learning in pedagogical situations. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 64–70). Austin, TX: Cognitive Science Society.

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.

Siler, S. A., & VanLehn, K. (2015). Investigating microadaptation in one-to-one human tutoring. *The Journal of Experimental Education*, *83*(3), 344–367.

Skerry, A. E., Lambert, E., Powell, L. J., & McAuliffe, K. (2013). The origins of pedagogy: Developmental and evolutionary perspectives. *Evolutionary Psychology*, *11*(3). https://doi.org/10.1177/147470491301100306

Sobel, D. M., & Kirkham, N. Z. (2007). Interactions between causal and statistical learning. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, Philosophy, and Computation*, (pp. 139–153. Oxford, England: Oxford University Press.

Strauss, S., Ziv, M., & Stein, A. (2002). Teaching as a natural cognition and its relations to preschoolers' developing theory of mind. *Cognitive Development*, *17*(3-4), 1473–1487.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629.

Tenenbaum, J. B., & Griffiths, T. L. (2003). Theory-based causal induction. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems* (Vol. 15, pp. 35–42). Cambridge, MA: MIT Press.

Tennie, C., Call, J., & Tomasello, M. (2009). Ratcheting up the ratchet: On the evolution of cumulative culture. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *364*(1528), 2405–2415.

Tomasello, M., Kruger, A. C., & Ratner, H. H. (1993). Cultural learning. *Behavioral and Brain Sciences*, *16*(3), 495–511.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, *46*(4), 197–221.

Velez-Ginorio, J., Siegel, M. S., Tenenbaum, J. B., & Jara-Ettinger, J. (2017). Interpreting actions by attributing compositional desires. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th annual conference of the cognitive science society* (pp. 1284–1289). Austin, TX: Cognitive Science Society.

Wang, P., Wang, J., Paranamana, P., & Shafto, P. (2020). A mathematical theory of cooperative communication. *Advances in Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. Quarterly journal of experimental psychology, 12(3), 129-140.

Wilson, N. S., & Bai, H. (2010). The relationships and impact of teachers' metacognitive knowledge and pedagogical understandings of metacognition. *Metacognition and Learning*, *53*, 269–288.

Wu, Y., & Schulz, L. (2018). Inferring beliefs and desires from emotional reactions to anticipated and observed events. *Child Development*, *89*(2), 649–662.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supplementary Information

Supplementary Fig. S1. Posterior probability mass placed on the true hypothesis (B&E) by the model as a function of teachers' examples. The posterior of the true hypothesis given teachers' original examples is plotted in blue; the posterior-given shuffled data are plotted in purple; and the posterior-given randomly generated examples are plotted in red. The dashed lines mark the 50% and 95% probability thresholds. An abridged version of

this plot can be found in the main manuscript; we provide the full version here.

Supplementary Fig. 2. The additional light-up machines used in Experiment 3, depicted with their corresponding blocks. Block M is required to make the first machine go; blocks Q, T, and V are all required to make the second machine go.