Rational Teachers Should 'Lie' to Bounded Students

Huang Ham (hamhuang@princeton.edu)[1], Dilip Arumugam[2], Carlos G. Correa[3], Bonan Zhao[4], Thomas L. Griffiths[1,2] & Natalia Vélez[1]

[1]Department of Psychology, Princeton University
[2]Department of Computer Science, Princeton University
[3]Department of Psychology, New York University
[4]School of Informatics, University of Edinburgh

Abstract

Educators often build up complex concepts by teaching simplified versions that are not quite accurate, such as Bohr's model of the atom or Newtonian mechanics. "Lying to children", while ubiquitous in STEM teaching, poses a challenge to existing cognitive models of pedagogy, which assume that teachers select evidence that truthfully represents a target concept. Why would helpful, knowledgeable teachers lie? We present a theoretical framework that addresses this puzzle by reinterpreting optimal pedagogy through the lens of bounded rationality. When learners face cognitive constraints on belief updating, our model predicts that teachers should prioritize examples that will bring the learner closest to the target concept—even if they do not represent the target concept truthfully; by contrast, classic pedagogy models fail to make this prediction. Our work formalizes an insight that educators have long understood: pedagogical "lies" are not meant to mislead learners, but to meet them where they are.

Keywords: Bayesian Pedagogy; Bounded Rationality; Information Theory; Optimal Transport; Rate-Distortion Theory; Social Cognition

Introduction

STEM educators often build up complex concepts by teaching students simplified versions that are, strictly speaking, wrong. Physics students learn Newtonian mechanics before discovering that it fails at high velocities; chemistry students memorize orbital diagrams that are superseded by quantum mechanics; and geometry students master Euclidean proofs before learning that parallel lines can meet. This practice—colloquially termed "lies-to-children" (Pratchett et al., 1999)—poses a challenge for Bayesian models of pedagogy, which assume that teachers pass down concepts to learners by selecting among examples that are consistent with the target concept (Shafto et al., 2021; Shafto et al., 2014). However, this assumption is often at odds with educational practice. Some prior research even suggest that learners fail to extract helpful information if the teacher is being deceptive (Alister et al., 2023). Why would, then, teachers deliberately select examples that are inconsistent with the concept they intend to convey?

Existing work from developmental and educational psychology offers a potential resolution: these "lies" are not strictly-speaking deception. They are *simplifications* or *approximations* designed to help teachers meet learners where they are. Good teaching cannot be defined in absolute terms; it depends on the learner's current knowledge state. This idea traces back to Vygotsky (1978) who argued that teaching should be situated within the learner's zone of proximal development, defined as the gap between what a learner can do independently and what they can achieve with guidance. Empirical work on the expertise reversal effect suggests that properly calibrating teaching to learners' expertise is critical: instructional techniques that are effective with beginners can harm more advanced learners, and vice versa (Kalyuga et al., 2003). Strikingly, even preschool-aged children provide information appropriately calibrated to the learner's knowledge (Gweon et al., 2018). In this light, simplified theories like Newtonian mechanics may act as scaffolds that enable progression towards more advanced concepts (Corcoran et al., 2009; Smith et al., 1993). Yet while prior work establishes the importance to calibrate to the learner, it does not explain when or why effective pedagogy should involve teaching content that is strictly false.

We address this gap by developing a computational framework that formalizes how teachers should calibrate to learners' cognitive constraints. First, we reformulate Bayesian pedagogy models to allow for an interpretation through bounded rationality, where learners have limited capacity to update their beliefs and where teachers aim to provide information that brings the learner *close enough* to the target hypotheses. Unlike classic models, our framework relaxes the assumption that teachers should only select among examples that accurately instantiate the target concept; instead, our model prioritizes examples that provide the most effective bridge between learners' initial beliefs and the truth. Second, through simulation studies, we establish that our framework identifies specific conditions where teachers should teach the best approximation of the target concept; by contrast, Bayesian pedagogy models do not predict that teachers should ever teach approximations. We conclude by discussing how this framework could be extended to capture when and how pedagogical approximations enable progression towards more sophisticated concepts across longer curricula. Simulation code and mathematical de-

tails are at https://github.com/HuangHam/CogSci2026-pedagogical-lies.

Background

Our framework builds on Bayesian pedagogy models, which formalize teaching as a series of recursive inferences between a teacher and a learner. More recent formulations based on optimal transport theory extend this approach to characterize teaching as a special case of cooperative communication.

Bayesian Pedagogy

Bayesian pedagogy models characterize teaching as a set of recursive inferences between a teacher and a learner, where the teacher selects examples that will lead the learner toward a target concept and the learner works backward from the examples provided to infer what the teacher is trying to tell them (Shafto et al., 2014). Formally, these models assume that teachers and learners share the following information:

- Hypothesis space $\mathcal{H}$: The set of possible concepts or hypotheses the teacher may wish to convey.
- Data space $\mathcal{D}$: The set of possible data or examples the teacher can provide.
- $P_L(h) \in \Delta(\mathcal{H})$: The learner's prior belief over the hypotheses, often assumed to be uniform.
- $P_T(d) \in \Delta(\mathcal{D})$: The teacher's prior belief over data, which is often assumed to be uniform as well.
- $P(d|h)_{init}$: An initial likelihood function, often defined by a consistency matrix that assigns equal probability to all examples d consistent with h ($P(d|h)_{init} \propto 1$), and 0 otherwise.

Within this context, teachers select evidence to teach ($P_T(d|h)$) that maximizes the learner's belief in the target hypothesis ($P_L(h|d)$). The solution can be derived through the following system of equations:

$$P_L(h|d) = \frac{P_T(d|h)P_L(h)}{\sum_{h' \in \mathcal{H}} P_T(d|h')P_L(h')} \tag{1}$$

$$P_T(d|h) = \frac{P_L(h|d)^{\alpha} P_T(d)}{\sum_{d' \in \mathcal{D}} P_L(h|d')^{\alpha} P_T(d')} \tag{2}$$

where $\alpha \in [0, \infty)$ is a temperature parameter that controls the extent to which the teacher chooses data that maximizes the learner's posterior belief in the true hypothesis.

This system of equations is typically solved using fixed-point iteration. The process begins with the teacher's initial likelihood function, $P(d|h)_{init}$, which assigns probability only to examples that are strictly consistent with the target hypothesis. The teacher then considers how likely the learner is to infer the correct hypothesis from these examples, assigning greater probability to examples that maximize the learner's posterior belief. This recursive process iterates until it converges to a fixed point. Importantly, because classic pedagogy models ground teaching in the consistency matrix $P(d|h)_{init}$, they can only select examples that are strictly true. Our theoretical framework relaxes this assumption to take learners' cognitive constraints into account.

Optimal Pedagogy as Belief Transport

A key insight from Bayesian pedagogy models is that the principles that underlie effective teaching support effective communication more broadly. Thus, these models capture teaching behaviors in a variety of modalities, including verbal descriptions (Sumers et al., 2022), demonstrations (Ho et al., 2016), and examples (Shafto et al., 2014), and they share deep commonalities with models of cooperative communication in language (Frank & Goodman, 2012; Goodman & Frank, 2016). Shafto et al. (2021) recently formalized this insight by proving that Bayesian pedagogy models can be derived from a more general theory of cooperative communication.

Specifically, this formalization is grounded in a well-studied optimization problem in probability theory known as optimal transport (Peyré, Cuturi, et al., 2019; Villani, 2008). While originally developed to find optimal plans to transport resources, Shafto et al. (2021) demonstrate how this framework can also be used to understand how teachers effectively "transport" a target concept or belief into the learner's mind. The teacher aims to minimize the expected cost of teaching, which is measured as: (1) how well the learner understands the target concept and (2) how easily the teacher selects examples. For pedagogical communication, the cost of teaching example d to convey hypothesis h is thus defined as:

$$C(d, h) = -\log P_L(h|d) - \log P_T(d) \tag{3}$$

We call this cost function the *Log-Posterior Cost*. The first term, $-\log P_L(h|d)$, captures the learning outcome: it is lower when the example d leads the learner to strongly believe in the target hypothesis h. The second term, $-\log P_T(d)$, reflects the teacher's prior preference or availability of examples. Notably, the optimal transport problem involves constraints on individual marginal distributions, requiring the provision of both $P_L(h)$ and $P_T(d)$. Relaxing one of these constraints constitutes an important step towards designing our new model.

By casting Bayesian pedagogy as an optimal transport problem, Shafto et al. (2021) outline a specific optimization problem teachers aim to solve for balancing pedagogical effectiveness with practical constraints on example selection. However, this optimization does not account for cognitive limitations on learners' belief-updating. Our bounded rationality framework builds on this notion of cost by using tools from information theory (Shannon, 1948) to formalize the relationship between pedagogical accuracy and the cognitive effort of learning.

Rate-distortion Theory

Rate-distortion theory (RDT) is the sub-area of information theory dedicated to lossy compression (Shannon, 1959). It formalizes the tradeoff between compression and fidelity: how much information can be discarded while keeping expected errors within acceptable bounds.

This framework maps onto pedagogy when we recognize that learners face analogous constraints: they cannot fully absorb all information a teacher provides, forcing them to balance the cognitive effort of updating their beliefs (rate) against how accurately those beliefs reflect the target concept (distortion). Critically, RDT provides mathematical tools to formalize this tradeoff. The cognitive effort of belief updating can be quantified using the Kullback–Leibler (KL) divergence (Kullback & Leibler, 1951) between the learner's prior and posterior beliefs,

$$D_{\mathrm{KL}}(P_L(h|d) \;||\; P_L(h)) = \sum_{h \in \mathcal{H}} P_L(h|d) \log \left(\frac{P_L(h|d)}{P_L(h)} \right)$$

which is equal to the mutual information (Cover & Thomas, 2012) between the teacher's provided data and learner's beliefs. Imposing a rate constraint is commensurate with a learner having limited capacity for learning. Under such bounded rationality, we can leverage RDT to derive optimal teaching strategies that minimize expected cost. Crucially, as we show later, this framework predicts that teachers should sometimes select examples that are strictly false but cognitively accessible.

Optimal Pedagogy for Bounded Learners

We can now formalize optimal pedagogy under cognitive constraints. From prior work, we begin with the optimal transport formulation of pedagogy, where teachers select examples that balance pedagogical effectiveness against practical constraints. Our key innovation, motivated by rate-distortion theory, is to relax the constraint on the teacher's marginal $P_T(d)$ imposed by optimal transport. Rather than fixing how teachers distribute probability over examples in advance, we allow teachers to select examples based on the learner's prior beliefs, balancing cognitive effort against expected cost. This small change fundamentally alters what counts as optimal teaching.

Formally, the teacher solves this optimization problem:

$$\min_{P_T(d|h)} \lambda \underbrace{\mathbb{E}\left[C(D,H)\right]}_{\text{Distortion}} + \underbrace{\mathbb{I}(D;H)}_{\text{Rate}}. \tag{4}$$

This objective balances two competing goals. The first term, $\mathbb{E}\left[C(D,H)\right]$, represents the expected cost of the learner's errors—that is, how much the learner's beliefs deviate from the target concept. Minimizing this term alone would lead teachers to select examples that maximize accuracy, as in Bayesian pedagogy models. This choice of cost function need not be the log-posterior cost (Equation 3) typically used in Bayesian pedagogy; in fact, we later entertain with an alternative "expected-mistake" cost function that is pivotal to our model of pedagogical approximation. The second term,

$$\begin{aligned} \mathbb{I}(D;H) &= \sum_{d \in \mathcal{D}} P_T(d) \cdot D_{\mathrm{KL}}(P_L(h|d) \;||\; P_L(h)) \\ &= \sum_{h \in \mathcal{H}} P_L(h) \cdot D_{\mathrm{KL}}(P_T(d|h) \;||\; P_T(d)) \end{aligned} \tag{5}$$

formalizes the cognitive effort of learning as the mutual information between the learner's beliefs and the teacher's examples—a standard measure of informational cost in cognitive science (Quillien & Taylor-Davies, 2026; Taylor-Davies & Quillien, 2025; Zhu & Griffiths, 2025). Intuitively, this measures how much the learner's beliefs change in response to teaching. Higher values indicate that the examples require the learner to substantially revise what they already believe; lower values indicate examples that align closely with the learner's prior beliefs. Minimizing this term corresponds to selecting examples that require minimal cognitive effort—that is, examples that "meet learners where they are."

The parameter $\lambda \in \mathbb{R}_{\geq 0}$ controls the tradeoff between these two objectives. When λ is large, the model prioritizes accuracy over cognitive effort, implicitly presuming that learners have arbitrarily high capacity. This setting recovers the behavior of Bayesian pedagogy models where teachers always select strictly true examples. When λ is small, learners face tight cognitive constraints, and the model prioritizes cognitive accessibility. This allows teachers to select examples that may be strictly false but easier for learners to absorb.

This formalization offers a potential explanation for pedagogical approximations: when cognitive resources are limited, teachers must balance bringing the learner *as close as possible* to the target concept against staying close to what learners already know and believe. Under sufficiently tight constraints, this tradeoff may favor selecting examples that are strictly inconsistent with the target concept but cognitively accessible. We test this prediction using a toy example inspired by real-world mathematics curricula, systematically comparing our bounded rationality model against classical Bayesian pedagogy.

Approximations about Geometry

The following toy scenario is inspired by the teaching of geometry. To the best of our knowledge, space is not flat, and thus Euclidean geometry is an over-simplified model of physical reality. Depending on the gravity field, hyperbolic geometry best describes regions with negative curvature (a saddle shape), while Riemannian geometry best describes regions with positive curvature (the surface of a ball). Thus, the system that best describes physical reality encompasses all of these alternative geometric systems—in other words, a general geometry. However,

Euclidean geometry is still widely taught in math and physics classes as the only geometric truth.

We formalize this pedagogical scenario as follows. The hypothesis space includes three geometric systems: h_1 = Euclidean geometry, h_2 = Riemannian geometry, and h_3 = general geometry. The data space includes three statements about triangle angle sums: d_1 = "the sum of interior angles of triangles is 180 degrees", d_2 = "the sum of interior angles of triangles is *greater* than 180 degrees", d_3 = "the sum of interior angles of triangles is *less* than 180 degrees". The consistency matrix below shows which statements are true under each geometry:

	h_1	h_2	h_3
d_1	1	0	1
d_2	0	1	1
d_3	0	0	1

Note that Euclidean geometry (h_1) is only consistent with d_1, Riemannian geometry (h_2) is only consistent with d_2, and general geometry (h_3) is consistent with all.

Model setup

Cost functions. We compare two cost functions to determine which give rise to pedagogical approximations. The first is the *log-posterior cost* commonly used in Bayesian pedagogy models (Equation 3). This cost is minimized when the learner strongly believes in the target hypothesis. Importantly, this cost function only allows teachers to select examples that are strictly consistent with the target hypothesis ($\log(0) = \infty$ for inconsistent examples).

The second cost function addresses a key limitation: existing models lack a notion that *some wrong beliefs are more useful than others* because they provide a bridge between the learner's initial beliefs and the truth. To capture this, we define a graded distance matrix M_H that describes how costly each mistaken belief is:

	h'_1	h'_2	h'_3
h_1	0	2	2
h_2	0.5	0	2
h_3	0.5	0.6	0

Rows represent the true concept; columns represent what the learner believes. These values capture the intuition that if reality is Euclidean (h_1), then believing otherwise is very costly. However, if reality is non-Euclidean (h_2 or h_3), believing it to be Euclidean is less costly—oversimplifying is better than overgeneralizing.

Using this matrix, we define the *expected-mistake cost*:

$$C(d, h) = \sum_{h'} P_L(h'|d) M_H(h, h') \tag{6}$$

This measures the expected cost of the learner's mistakes under the posterior shaped by example d. Unlike log-posterior cost, this allows teachers to select examples that are strictly inconsistent with the target concept, as long as the resulting mistakes are less costly than the alternatives.

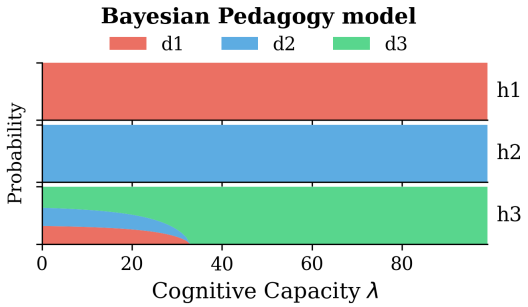

Figure 1: Bayesian pedagogy model. Rows correspond to different true hypotheses, varying the learner's cognitive capacity. Colors indicate the probability mass allocated to each datum.

Learner priors. We assume learners have uniform priors over hypotheses, consistent with standard Bayesian pedagogy models. However, this framework can accommodate non-uniform priors as well.

Aligning model parameters for comparison. Using the mathematical connection between optimal transport and rate-distortion theory, it can be shown that Bayesian pedagogy models are a special case of our framework. For a given α value, the Bayesian pedagogy model optimizes the bounded-rational objective (Equation 4) within a *constrained* subspace of all possible teaching policies, due to the marginal constraint imposed by the teacher's prior $P_T(d)$. This means we can directly compare how the models behave as we vary cognitive constraints by setting $\lambda = \alpha$. We simulate both models across λ values ranging from 0.1 to 100 (evenly spaced in log space), where $\lambda = 0.1$ represents the tightest cognitive constraints and $\lambda = 100$ represents learners with minimal constraints.

Bounded rationality predicts approximations

Simulation of the Bayesian pedagogy model (Figure 1) shows that it never predicts that teachers should teach approximations. The model consistently uses d_1 to teach h_1, d_2 to teach h_2, and d_3 to teach h_3, regardless of cognitive capacity λ. While there is more stochasticity at small λ, the model assigns zero probability to any inconsistent statement. Critically, it fails to capture real educational practice, where student are taught "the sum of interior angles is 180 degrees" (d_1), though general geometry best captures reality (h_3).

By contrast, our bounded rationality model successfully predicts pedagogical approximations (Figure 2). We modified the Bayesian pedagogy model in two ways: first, by replacing the log-posterior cost function with expected-mistake cost (Equation 6) and, second, by relaxing the marginal constraint on $P_T(d)$ to obtain the

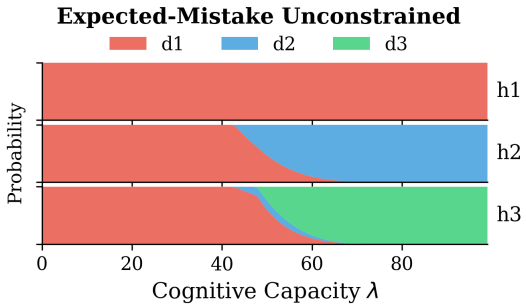

Figure 2: Our model. Rows correspond to different true hypotheses, varying the learner's cognitive capacity. Colors indicate the probability mass allocated to examples.

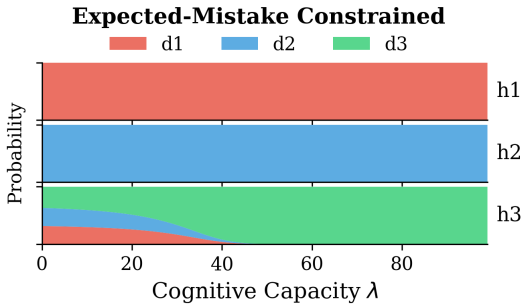

Figure 3: Only changing the cost function does not produce behavior of teaching approximations. Rows correspond to different true hypotheses, varying the learner's cognitive capacity. Colors indicate the probability mass allocated to each datum.

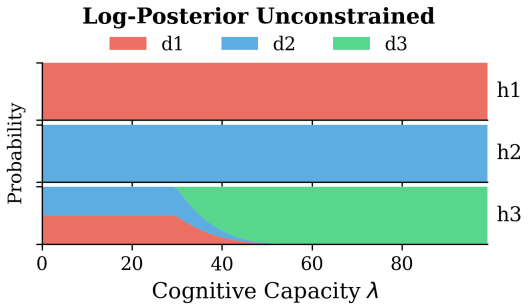

Figure 4: Only relaxing the constraint does not produce behavior of teaching approximations. Rows correspond to different true hypotheses, varying the learner's cognitive capacity. Colors indicate the probability mass allocated to each datum.

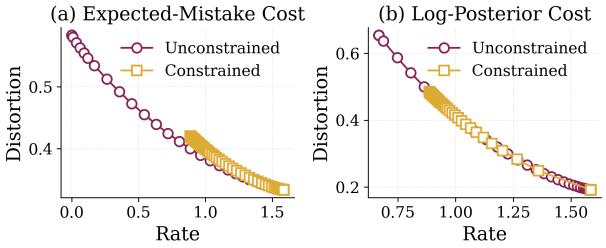

Figure 5: Rate-distortion curves: The dots correspond to the 100 λ values we specified to simulate the models. The curves show the trade-off achieved by each of our four models in solving the bounded-rational teaching objective (Equation 4). For a given rate, the lower the distortion, the more optimal is the model's solution.

unconstrained optimal solution to the bounded-rational objective (Equation 4). When cognitive constraints are tight (λ is low), the model predicts teachers should use d_1 ("the sum of interior angle of a triangle is 180 degrees") to teach all three geometric systems. This is a genuine approximation: d_1 is strictly inconsistent with h_2, yet optimal teaching still selects it under tight constraints. As constraints relax (λ increases), the model transitions through a period where d_2 becomes more probable for teaching h_2 and h_3, mirroring how Riemannian geometry is commonly introduced first to illustrate that geometry need not be Euclidean. When constraints are minimal (λ is large), the model converges to the same predictions as Bayesian pedagogy, selecting only consistent examples.

Localizing pedagogical approximations

We next explore whether the two modifications our model made are both necessary to induce approximations. First, we explore whether changing the cost function into the expected mistake cost (Equation 6) alone suffices. We simulated a model where we changed the cost function but only optimized within the same marginal constraint as the Bayesian pedagogy model by also using a optimal transport solver. Interestingly, this new model looks qualitatively the same as the Bayesian pedagogy model (Figure 3). This suggests that changing the cost function alone does not suffice. One hypothesis is that optimizing within a further constrained subspace aggressively prunes otherwise optimal solutions to the teaching problem from consideration. Support for this is visible in the rate-distortion curve of the two models (Figure 5A). When using the same cost function, the optimal transport solver (which is constrained by the marginal) only yields a sub-optimal solution compared to the RDT solver; at each rate, it incurs a higher expected distortion.

For completeness, we next show that the change in cost function is also necessary for the qualitative behavioral pattern. In other words, simply relaxing the marginal constraint assumed by the Bayesian pedagogy model is not enough. We simulated another model where we kept the log-posterior cost function but relax the marginal constraint (Figure 4). This is mathematically equivalent to the RD-RSA model (Zaslavsky et al., 2021; Zhou et al.,

2021). Even though the model simulation shows a qualitative difference from the regular pedagogy model, it still does not predict approximations when it is optimal to teach inconsistent examples sometimes. We plotted the rate-distortion curve of this model and contrasted it with the Bayesian pedagogy model (Figure 5B), which shows that the unconstrained model achieved a broader range of optimal teaching strategies (more achievable rates).

We have simulated four models to solve the bounded-rational teaching problem. Two models used the log-posterior cost function, the other two used the expected-mistake cost function, which does not assume all mistakes are equally bad. Two models enforced the teacher to be constrained by their prior distribution $P_T(d)$ by using an optimal transport solver, whereas the other two relaxed such a constraint by using an RDT solver. We show that only the model using expected-mistake cost function *and* relaxed the constraint on teacher successfully simulated the behavior of teaching approximations to cognitively limited learner. Results are summarized in Table 1.

Table 1: Summary of bounded-rational teaching models and their ability to simulate behavior of teaching approximations when learners are cognitively limited.

	Cost Function	
Solution	Log-posterior	Expected-mistake
Constrained	No	No
Unconstrained	No	Yes

Discussion

Real-world teachers often teach knowledge that is, strictly-speaking, incorrect but useful as approximation to truth. Although this does not constitute a deliberate attempt to mislead the learner, people colloquially coined the term "lying-to-children" while referring to this pedagogical phenomenon. Research from educational and developmental sciences argues that this is a good pedagogical practice that accommodates learners' cognitive limitations (Corcoran et al., 2009; Kalyuga et al., 2003; Smith et al., 1993). However, the Bayesian pedagogy model does not predict that lying to learners should ever be optimal, regardless of learner's cognitive limitations. By reinterpreting and generalizing the Bayesian pedagogy model using tools from rate-distortion theory, we found that it makes two overly-restrictive assumptions—one about the objective of teaching and another about the constraints from teacher's prior. Only by changing these two assumptions did we show that lying to learners can emerge as the optimal pedagogical practice when learners are cognitively limited. More broadly, our formulation extends the line of work on rate-distortion theory as a formal tool for modeling resource-limited cognition (Arumugam et al., 2024; Bhui et al., 2021; Gershman, 2020; Lai & Gershman, 2021; Prystawski et al., 2023, 2025; Sims, 2016; Turner et al., 2025; Zaslavsky et al., 2021; Zhao et al., 2025). However, our findings so far are only theoretical. Future studies need to test the theory against human behavior to see whether learner's cognitive limitations drive teacher's decisions about whether to teach things inconsistent with the true hypothesis.

Our theoretic results demonstrate the fruitfulness of considering optimal pedagogy through the lens of bounded rationality. Beyond explaining "lying-to-children", our framework affords a more general view of optimal pedagogy that can be used to better explain the richness of real-world teaching. For example, a key insight of our framework is that approximations constitute optimal teaching when learners face cognitive constraints on belief updating—instead of being restricted to providing information that is strictly true, our model prioritizes examples that provide a bridge between learners' initial beliefs and the truth. One implication of this result is that approximations are useful because they may aid progression towards more complex concepts. If this is the case, then a natural next step is to model not just the optimal teacher for a single data point, but to design a full curriculum. Information geometry (Amari, 2016) may contain useful insights, providing the formal tools to reimagine the probability distribution as a dot on a surface and belief shaping as movement along that surface. It allows us to convert an abstract problem of teaching into a more intuitive problem of designing the most efficient path.

Another exciting direction is to formalize alternative forms of teaching. Our current framework assumes that learners update beliefs through Bayesian inference over a fixed hypothesis space, which constrains the teacher to shaping beliefs *within* that space. However, human teachers often aim to restructure the hypothesis space itself—for instance, by conveying useful abstractions or reasoning strategies that learners can later apply across domains (Ham et al., 2025). Extending the information geometry perspective, we might conceptualize such interventions further as transformations of the manifold's geometry, effectively changing the representational space over which future learning occurs. This connects to longstanding questions in cognitive science about how instruction facilitates representational change, schema acquisition, and learning to learn (Carey, 2009; Gentner et al., 2003). Formalizing these richer pedagogical strategies would bring our framework closer to capturing the full scope of human teaching, where the goal is often not merely to transmit specific knowledge but to equip learners with cognitive tools that accelerate learning.

Acknowledgments

We thank Patrick Shafto for pointing us to the connections between EOT and RDT. We thank Jian-qiao Zhu

for helpful discussions regarding the application of KL divergence in cognitive science.

References

Alister, M., Ransom, K. J., & Perfors, A. (2023). Inferring the truth from deception: What can people learn from helpful and unhelpful information providers? *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45).

Amari, S.-i. (2016). Information geometry and its applications (Vol. 194). Springer.

Amari, S.-i., Karakida, R., & Oizumi, M. (2018). Information geometry connecting Wasserstein distance and Kullback–Leibler divergence via the entropy-relaxed transportation problem. *Information Geometry*, *1*(1), 13–37.

Arimoto, S. (1972). An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, *18*(1), 14–20.

Arumugam, D., Ho, M. K., Goodman, N. D., & Van Roy, B. (2024). Bayesian reinforcement learning with limited cognitive load. *Open Mind*, *8*, 395–438.

Berger, T. (1971). Rate distortion theory: A mathematical basis for data compression. Prentice-Hall.

Bhui, R., Lai, L., & Gershman, S. J. (2021). Resource-rational decision making. *Current Opinion in Behavioral Sciences*, *41*, 15–21.

Blahut, R. (1972). Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory*, *18*(4), 460–473.

Boyd, S. P., & Vandenberghe, L. (2004). Convex optimization. Cambridge University Press.

Carey, S. (2009). The origin of concepts. Oxford University Press.

Chiang, M., & Boyd, S. (2004). Geometric programming duals of channel capacity and rate distortion. *IEEE Transactions on Information Theory*, *50*(2), 245–258.

Corcoran, T., Mosher, F. A., & Rogat, A. (2009, May). *Learning progressions in science: An evidence-based approach to reform* (CPRE Research Report No. RR-63). Consortium for Policy Research in Education. Philadelphia, PA.

Cover, T. M., & Thomas, J. A. (2012). Elements of information theory. John Wiley & Sons.

Csiszár, I. (1974a). On an extremum problem of information theory. *Studia Scientiarum Mathematicarum Hungarica*, *9*.

Csiszár, I. (1974b). On the computation of rate-distortion functions. *IEEE Transactions on Information Theory*, *20*(1), 122–124.

Duchi, J. C. (2025). Lecture notes for statistics 311/electrical engineering 377: Information theory and statistics. Stanford University.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998. https://doi.org/10.1126/science.1218633

Gentner, D., Loewenstein, J., & Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology*, *95*(2), 393–408. https://doi.org/10.1037/0022-0663.95.2.393

Gershman, S. J. (2020). Origin of perseveration in the trade-off between reward and complexity. *Cognition*, *204*, 104394.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829. https://doi.org/10.1016/j.tics.2016.08.005

Gray, R. M. (2011). Entropy and information theory. Springer.

Gweon, H., Shafto, P., & Schulz, L. (2018). Development of children's sensitivity to overinformativeness in learning and teaching. *Developmental Psychology*, *54*(11), 2113–2125. https://doi.org/10.1037/dev0000580

Ham, H., Zhao, B., Griffiths, T. L., & Vélez, N. (2025). Teaching recombinable motifs through simple examples. *Cognitive Science*, *49*(8), e70103.

Ho, M. K., Littman, M. L., MacGlashan, J., Cushman, F., & Austerweil, J. L. (2016). Showing versus doing: Teaching by demonstration. *Advances in Neural Information Processing Systems*, *29*, 3027–3035.

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist*, *38*(1), 23–31. https://doi.org/10.1207/S15326985EP3801_4

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86.

Lai, L., & Gershman, S. J. (2021). Policy compression: An information bottleneck in action selection. In K. D. Federmeier (Ed.), *The psychology of learning and motivation* (pp. 195–232, Vol. 74). Academic Press.

Lei, E., Hassani, H., & Bidokhti, S. S. (2022). Neural estimation of the rate-distortion function for massive datasets. *2022 IEEE International Symposium on Information Theory (ISIT)*, 608–613.

Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, *11*(5-6), 355–607.

Polyanskiy, Y., & Wu, Y. (2024). Information theory: From learning to coding. Cambridge University Press.

Pratchett, T., Stewart, I., & Cohen, J. (1999). The science of Discworld. Ebury Press.

Prystawski, B., Arumugam, D., & Goodman, N. (2023). Cultural reinforcement learning: A framework for mod-

eling cumulative culture on a limited channel. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45).

Prystawski, B., Arumugam, D., & Goodman, N. D. (2025). Lossy communication constrains iterated learning. *arXiv preprint arXiv:2511.18220*.

Quillien, T., & Taylor-Davies, M. (2026). Factive mindreading reflects the optimal use of limited cognitive resources. *Proceedings of the Royal Society B*, *293*(20251852). https://doi.org/10.1098/rspb.2025.1852

Shafto, P., Wang, J., & Wang, P. (2021). Cooperative communication as belief transport. *Trends in Cognitive Sciences*, *25*(10), 826–828. https://doi.org/10.1016/j.tics.2021.07.012

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. https://doi.org/https://doi.org/10.1016/j.cogpsych.2013.12.004

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Shannon, C. E. (1959). Coding theorems for a discrete source with a fidelity criterion. *Institute of Radio Engineers National Convention Record*, *4*, 142–163.

Sims, C. R. (2016). Rate–distortion theory and human perception. *Cognition*, *152*, 181–198.

Smith, J. P., diSessa, A. A., & Roschelle, J. (1993). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *The Journal of the Learning Sciences*, *3*(2), 115–163. https://doi.org/10.1207/s15327809jls0302_1

Sumers, T. R., Hawkins, R. D., Ho, M. K., Griffiths, T. L., & Hadfield-Menell, D. (2022). How to talk so AI will learn: Instructions, descriptions, and autonomy. *Advances in Neural Information Processing Systems*, *35*.

Taylor-Davies, M., & Quillien, T. (2025). An information-bottleneck view of social stereotype use. *Proceedings of the 47th Annual Meeting of the Cognitive Science Society*.

Turner, C. R., Arumugam, D., Nelson, L. R., & Griffiths, T. (2025). Trade-offs between tasks induced by capacity constraints bound the scope of intelligence. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*.

Villani, C. (2008). Optimal transport: Old and new (Vol. 338). Springer Science & Business Media.

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds.). Harvard University Press.

Zaslavsky, N., Hu, J., & Levy, R. (2021). A rate–distortion view of human pragmatic reasoning. *Proceedings of the Society for Computation in Linguistics 2021*, 347–348.

Zhao, B., Mieczkowski, E., Arumugam, D., Vélez, N., & Griffiths, T. (2025). Discovering hidden laws in innovation by recombination. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*.

Zhou, I., Hu, J., Levy, R. P., & Zaslavsky, N. (2021). Empirical support for a rate–distortion account of pragmatic reasoning. *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*, 528–534.

Zhu, J.-Q., & Griffiths, T. L. (2025). Computation-limited Bayesian updating: A resource-rational analysis of approximate Bayesian inference. *Psychological Review*. https://doi.org/10.1037/rev0000514

Appendix

Preliminaries

Probability Theory & Information Theory Here, we introduce core concepts in probability theory and information theory (Shannon, 1948) used throughout this paper. See Cover and Thomas (2012), Gray (2011), and Polyanskiy and Wu (2024) and Duchi (2025) for more background.

Our probability theory notation aligns closely with that of Polyanskiy and Wu (2024). For any set $\mathcal{X}$, $\Delta(\mathcal{X})$ denotes the set of all probability distributions with support on $\mathcal{X}$. For any two sets $\mathcal{X}$ and $\mathcal{Y}$, we denote the class of all functions mapping from $\mathcal{X}$ to $\mathcal{Y}$ as $\{\mathcal{X} \to \mathcal{Y}\} \triangleq \{f \mid f : \mathcal{X} \to \mathcal{Y}\}$. For any random variable X taking values in $\mathcal{X}$, we will denote the law or distribution of X as $P_X \in \Delta(\mathcal{X})$. Similarly, the joint distribution of random variables X and Y — taking values in $\mathcal{X}$ and $\mathcal{Y}$, respectively — will be written as $P_{X,Y} \in \Delta(\mathcal{X} \times \mathcal{Y})$; we denote the conditional probability distribution of Y given X as $P_{Y|X} \in \{\mathcal{X} \to \Delta(\mathcal{Y})\}$. For any two distributions $P, Q \in \Delta(\mathcal{X})$, the Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951) between P and Q is defined as

$$D_{\mathrm{KL}}(P \mid\mid Q) = \begin{cases} \mathbb{E}_P\left[\log\left(\frac{P(X)}{Q(X)}\right)\right] & P \ll Q \\ +\infty & P \not\ll Q \end{cases},$$

where $P \ll Q$ denotes the absolute continuity of P with respect to Q.

We define the mutual information between any two random variables X, Y as the KL-divergence between the joint distribution $P_{X,Y}$ and the product of their respective marginal distributions, P_X and P_Y:

$$\mathbb{I}(X;Y) = D_{\mathrm{KL}}(P_{X,Y} \mid\mid P_X \times P_Y).$$

We define the entropy and conditional entropy for any two random variables X, Y as

$$\mathbb{H}(X) = \mathbb{I}(X;X), \qquad \mathbb{H}(Y \mid X) = \mathbb{H}(Y) - \mathbb{I}(X;Y).$$

This yields the following identity for mutual information that applies to any arbitrary random variables X and Y:

$$\mathbb{I}(X;Y) = \mathbb{H}(X) - \mathbb{H}(X \mid Y).$$

We may also define the joint entropy for two random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ as the (traditional) entropy of the corresponding $(\mathcal{X} \times \mathcal{Y})$-valued random variable: $\mathbb{H}(X,Y) = \mathbb{H}((X,Y))$. Through the chain rule of the KL-divergence and the fact that $D_{\mathrm{KL}}(P \mid\mid P) = 0$ for any probability distribution P, we obtain another equivalent definition of mutual information,

$$\mathbb{I}(X;Y) = \mathbb{E}\left[D_{\mathrm{KL}}(P_{Y|X} \mid\mid P_Y)\right] = \mathbb{E}\left[D_{\mathrm{KL}}(P_{X|Y} \mid\mid P_X)\right].$$

Fact 1 (Theorem 2.4.1 of Cover and Thomas (2012))**.** *For any two random variables X and Y,*

$$\mathbb{I}(X;Y) = \mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(X,Y).$$

Rate-Distortion Theory Rate-distortion theory (Berger, 1971; Shannon, 1959) is the sub-area of information theory dedicated to lossy compression problems. Tools from rate-distortion theory help to establish fundamental limits on how much information must be retained in a lossy compression in order to achieve a bounded degree of expected error. Moreover, there are algorithms from rate-distortion theory that facilitate the computation of such optimal compressions.

A lossy compression problem requires two inputs: an information source and a distortion function. An information source is a probability distribution $P_X \in \Delta(\mathcal{X})$ representing uncompressed data. For a given information source P_X, one looks to compress data X via a channel. A channel $Q_{Y|X} : \mathcal{X} \to \Delta(\mathcal{Y})$ is a conditional probability distribution over compressed outputs $Y \in \mathcal{Y}$ to each input $x \in \mathcal{X}$. A distortion function $d : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ measures the degradation or loss of fidelity when assigning compressed values to uncompressed data. To any channel, there is an associated rate defined as the mutual information $\mathbb{I}(X;Y)$ between the original, uncompressed data $X \sim P_X$ (channel input) and the compressed data $Y \sim Q_{Y|X}$ (channel output). This rate is a measure of how much information (on average) from the original data is retained by the compression. Meanwhile, we are often equally interested in the expected distortion incurred by a channel:

$$\mathbb{E}\left[d(X,Y)\right] = \mathbb{E}_{Q_{X,Y}}\left[d(X,Y)\right] = \mathbb{E}_{P_X}\left[\mathbb{E}_{Q_{Y|X}}\left[d(X,Y)\right]\right].$$

The distortion-rate function quantifies the fundamental limit on the minimum amount of expected distortion incurred when the channel is constrained to an upper bound on rate. While the distortion-rate function may normally be written

as an optimization function over channels $Q_{Y|X}$, it is useful for this paper to equivalently express it as an optimization over joint distributions $Q_{X,Y}$ constrained to preserve the X marginal, P_X:

$$\mathcal{D}(R) = \min_{Q_{X,Y}:Q_X=P_X} \mathbb{E}\left[d(X,Y)\right] \text{ such that } \mathbb{I}(X;Y) \leq R.$$

Here, $R \in \mathbb{R}_{\geq 0}$ is a rate constraint that a channel achieving the distortion-rate limit must adhere to. In the case of discrete random variables, such a channel can be computed directly via the Blahut-Arimoto algorithm (Arimoto, 1972; Blahut, 1972; Csiszár, 1974b), which is an alternating minimization algorithm that minimizes the (unconstrained) Lagrangian (Boyd & Vandenberghe, 2004) of the constrained optimization problem:

$$\min_{Q_{X,Y}:Q_X=P_X} \mathbb{E}\left[d(X,Y)\right] + \beta \cdot \mathbb{I}(X,Y).$$

The Lagrange multiplier $\beta \in \mathbb{R}_{\geq 0}$ serves as a hyperparameter negotiating between the relative importance of minimizing rate versus expected distortion. Relating a particular rate limit $R \in \mathbb{R}_{\geq 0}$ back to a concrete value of β follows from the fact that the distortion-rate function is itself a convex optimization problem (Chiang & Boyd, 2004) where strong duality holds (Csiszár, 1974a):

$$\mathcal{D}(R) = \sup_{\beta \in \mathbb{R}_{\geq 0}} \left[\left(\min_{Q_{X,Y}:Q_X=P_X} \mathbb{E}\left[d(X,Y)\right] + \beta \cdot \mathbb{I}(X,Y) \right) - \beta \cdot R \right].$$

From a computational perspective, the Blahut-Arimoto algorithm operates under the assumption that all random variables involved in the lossy compression problem are discrete, however the procedure holds in full generality for abstract random variables (Csiszár, 1974a, 1974b).

Optimal Pedagogy as Bounded Rationality

Recall that optimal pedagogy can be expressed as an instance of an entropic optimal transport (EOT) problem (Shafto et al., 2021) defined as

$$\min_{Q_{X,Y}:Q_X=P_X,Q_Y=P_Y} \mathbb{E}\left[C(X,Y)\right] - \lambda^{-1} \cdot \mathbb{H}(X,Y),$$

where $P_X \in \Delta(\mathcal{X})$ is a given X marginal that the transport plan must adhere to, $P_Y \in \Delta(\mathcal{Y})$ is a given Y marginal that the transport plan must adhere to, $C : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ is a given cost function, and $\lambda \in \mathbb{R}_{\geq 0}$ is a Lagrange multiplier weighting the relative importance of maximizing joint entropy and minimizing expected cost.

To draw an explicit connection between optimal pedagogy as entropic optimal transport and bounded rationality, we offer the following proposition:

Proposition 0.1. *Let $P_X \in \Delta(\mathcal{X})$ and $P_Y \in \Delta(\mathcal{Y})$ be two known marginal distributions for random variables X and Y, respectively, Then, for any joint distribution $Q_{X,Y} \in \Delta(\mathcal{X} \times \mathcal{Y})$ such that both marginals are preserved — $Q_X = P_X$ and $Q_Y = P_Y$ — we have that*

$$\operatorname*{arg\,min}_{Q_{X,Y}:Q_X=P_X,Q_Y=P_Y} \mathbb{E}\left[C(X,Y)\right] - \lambda^{-1} \cdot \mathbb{H}(X,Y) = \operatorname*{arg\,min}_{Q_{X,Y}:Q_X=P_X,Q_Y=P_Y} \mathbb{E}\left[C(X,Y)\right] + \lambda^{-1} \cdot \mathbb{I}(X;Y).$$

Proof. The proof follows by a direct application of Fact 1 and the marginal constraints imposed by the EOT problem itself, which render the marginal entropy terms constant with respect to the optimization over joint distributions $Q_{X,Y}$:

$$\begin{aligned}\operatorname*{arg\,min}_{Q_{X,Y}:Q_X=P_X,Q_Y=P_Y} \mathbb{E}\left[C(X,Y)\right] - \lambda^{-1} \cdot \mathbb{H}(X,Y) &= \operatorname*{arg\,min}_{Q_{X,Y}:Q_X=P_X,Q_Y=P_Y} \mathbb{E}\left[C(X,Y)\right] - \lambda^{-1} \cdot (\mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{I}(X;Y)) \\ &= \operatorname*{arg\,min}_{Q_{X,Y}:Q_X=P_X,Q_Y=P_Y} \mathbb{E}\left[C(X,Y)\right] + \lambda^{-1} \cdot \mathbb{I}(X;Y).\end{aligned}$$

□

To give an interpretation of Proposition 0.1 and its significance, recall that one definition of mutual information is

$$\mathbb{I}(X;Y) = \mathbb{E}\left[D_{\mathrm{KL}}(P_{X|Y} \,||\, P_X)\right].$$

Adopting a Bayesian interpretation, P_X represents a prior distribution over X and $P_{X|Y}$ represents the corresponding posterior distribution over X after observing Y. Thus, the mutual information quantifies the expected information gain about X from Y. This interpretation takes on the explicit form of quantifying how much an individual observation shifts a learner's prior to a corresponding posterior, on average. One may interpret minimizing this term as being commensurate with minimizing cognitive effort or the degree to which learning disrupts prior beliefs. The form of KL-divergence has been used widely in machine learning and cognitive science to measure informational cost to a cognitive system (Quillien & Taylor-Davies, 2026; Taylor-Davies & Quillien, 2025; Zhu & Griffiths, 2025).

From Bounded Rationality to Rate-Distortion Theory

If we consider the objective function associated with the minimizer identified by Proposition 0.1,

$$\min_{Q_{X,Y}:Q_X=P_X,Q_Y=P_Y} \mathbb{E}\left[C(X,Y)\right] + \lambda^{-1} \cdot \mathbb{I}(X;Y),$$

then we may also draw a direct line from the aforementioned bounded-rationality perspective on optimal pedagogy to rate-distortion theory. In particular, observe that a relaxation of the Y marginal constraint immediately yields the Lagrangian associated with the distortion-rate function:

$$\min_{Q_{X,Y}:Q_X=P_X,Q_Y=P_Y} \mathbb{E}\left[C(X,Y)\right] + \lambda^{-1} \cdot \mathbb{I}(X;Y) \geq \min_{Q_{X,Y}:Q_X=P_X} \mathbb{E}\left[C(X,Y)\right] + \lambda^{-1} \cdot \mathbb{I}(X;Y),$$

where the cost function C is the distortion function and λ^{-1} acts as the corresponding Lagrange multiplier. Note that we are not the first to observe an upper bound between the distortion-rate function and EOT objective (Amari et al., 2018; Lei et al., 2022).

By relaxing the teacher's marginal constraint, we relax the assumption that the teacher is performing Bayesian inference. Psychologically, this implies the teacher need not be constrained by a certain pattern of target-independent "cost" of teaching each example. We find this to be a reasonable for the common scenarios wherein the teacher does not predetermine how often each example must be shown. At this point, note that the retention of the X marginal constraint is tantamount to maintaining the assumption that a learner will update their beliefs according to Bayes' rule.