

Toward a Formalization of Human Intuitive Theories of Bodily Pain

Marlene D. Berke¹, Katherine M. Collins^{2,3,4}, Joshua B. Tenenbaum², and Rebecca Saxe^{1,2}

¹McGovern Institute for Brain Research, Massachusetts Institute of Technology

²Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

³Princeton AI Lab, Princeton University

⁴Cambridge Centre for Human-Inspired AI, Cambridge University

Author Note

Correspondence concerning this article should be addressed to Marlene D. Berke, McGovern Institute for Brain Research, Massachusetts Institute of Technology. Email: berkem@mit.edu. The authors used Claude (Anthropic, 2026) and Codex (OpenAI, 2026b) to assist with copyediting portions of the manuscript and supplementary materials. The authors reviewed and take responsibility for all content.

Abstract

Pain is a private, subjective experience, yet observers routinely make consequential judgments about another person's pain (e.g., whether it warrants treatment, accommodation, or whether it's real at all). How do observers make these judgments? We propose that people attribute pain through an intuitive theory—the same kind of theory they use to infer other hidden mental and physical states. We explore the structure of people's intuitive theory of pain by formalizing competing conceptions of pain: from pain resulting from bodily damage observable through physical evidence to pain as a subjective experience known only via self-report. Each theory is formalized as a causal model that predicts how people interpret naturalistic vignettes drawn from Reddit posts and medical training materials. Three experiments probe classical signatures of Bayesian causal reasoning: predicting pain from physical evidence (Experiment 1), inferring and treating pain (Experiment 2), and inferring a hidden cause (ulterior motives) that explains away inflated pain reports (Experiment 3). Across all three, a causal model that integrates over physical and mental states best captures human behavior and generalizes across eight categories of bodily pain, from headaches to finger burns. These results are consistent with our proposal that people have an intuitive theory of pain arising from problems in the body, and this theory supports quantitative causal reasoning. As a direct consequence of this intuitive theory, lay people view pain reports that lack corroborating physical evidence with suspicion. This work is a first step toward a quantitative, scientific model of folk intuitive theories of pain.

Keywords: social cognition; intuitive theories; computational modeling; intuitive medicine; Bayesian modeling; mind-body problem; theory of mind; pain

Toward a Formalization of Human Intuitive Theories of Bodily Pain

Pain is a universal human experience, and yet, each person’s pain is private and inaccessible to outsiders. At the same time, people need outsiders to understand their pain experience in order to access needed resources like medical care or help with daily tasks. This makes the challenge of communicating and understanding pain a very important and high-stakes problem for social cognition to solve.

Despite the challenge, observers make pain attributions all the time: parents judge whether their kid is ‘really’ sick or just doesn’t want to go to school, clinicians estimate how much pain their patient is in, and juries determine responsibility and damages for pain and suffering. A rich descriptive literature has characterized what kinds of evidence contribute to these judgments and when people choose to rely on which (for review, see Tait et al., 2009). Our goal is to develop a formal cognitive model of *how* observers actually do this—a quantitative account of how different sources of evidence are combined to ascribe pain to another person, what predictions follow when they conflict, and how the same cognitive machinery might support a wide repertoire of reasoning about pain: predicting pain, treating pain, suspecting the truthfulness of a pain report.

We hypothesize that people solve the pain-attribution problem the way they solve inference problems about other kinds of hidden physical or mental states: by reasoning over an *intuitive theory*. The framework of intuitive theories has been used to formalize lay reasoning in related domains like intuitive physics (Battaglia et al., 2013; McCloskey, 1983), intuitive psychology (Baker et al., 2009; Gopnik & Wellman, 1992), and intuitive biology (Atran, 1990; Tenenbaum et al., 2006). Under this view, each domain supports causal models over which people perform causal reasoning, enabling a variety of abilities, including predicting effects from causes, inferring hidden causes from observable effects, intervening on a cause to produce a desired effect, and explaining effects by appealing to causes (Gerstenberg & Tenenbaum, 2017; Gopnik et al., 2004; Pearl, 2000). Pain seems to straddle intuitive physics, biology, and psychology; for example, people believe that physical

events like kicking a rock or eating a berry can cause pain (Outa & Liu, 2023), even from a young age (Schulz et al., 2007). Furthermore, the ability to experience pain occupies a uniquely important position in folk taxonomies of mental life (Weisman et al., 2017) and morality (Gray et al., 2007). Despite its unique importance and its direct links to intuitive physics and intuitive psychology, folk intuitions about pain have rarely been positioned within the intuitive theory framework. Doing so opens a path toward formalization.

Describing pain attribution as inference over a causal model gives us a language for formalizing verbal theories of what pain is and how people think it works. This allows us to build cognitive models representing these different theories and test them quantitatively. Drawing inspiration from philosophical and scientific theories of pain, we develop competing causal models instantiating different theoretical commitments about the nature of pain and its evidence. We make these models *stimulus-computable* so that they produce judgments directly from the same stimuli that humans see (Raz et al., 2025), enabling a direct and quantitative comparison between model and human behavior.

We test human intuitions about pain in three experiments designed to probe for classical signatures of causal reasoning. Experiment 1 tests *forward prediction*: Can people predict pain from physical events and evidence alone, without any report of experience? Experiment 2 tests *inverse inference* and *intervention*: Do people treat pain experience like a hidden state that must be inferred, or do they take pain reports at face value? Do they prescribe treatment accordingly? Experiment 3 tests *inferring a hidden cause* and *explaining away*: Do people infer the presence of a hidden cause (e.g., ulterior motives) to explain elevated pain reports, and does this hidden cause explain away pain reports? Across all three experiments, the same causal model with fixed structure and fixed parameters captures human judgments better than alternative models lacking its causal structure. This model generalizes across eight different categories of bodily pain. These results provide a formal description of a hybrid mind-body folk theory of pain: neither purely subjective nor purely physical, but integrated through a causal model that links

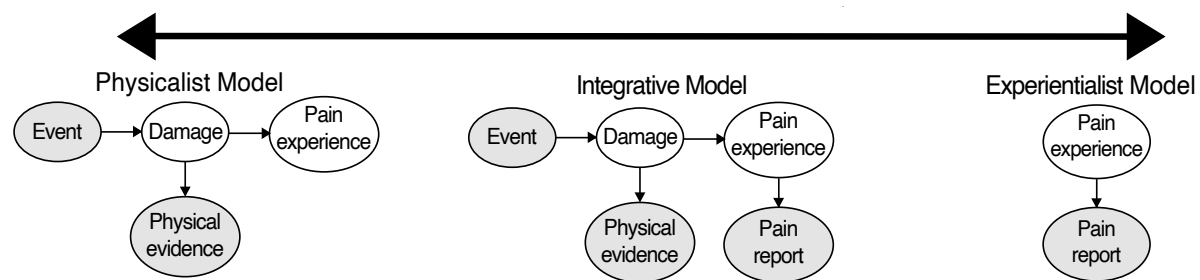
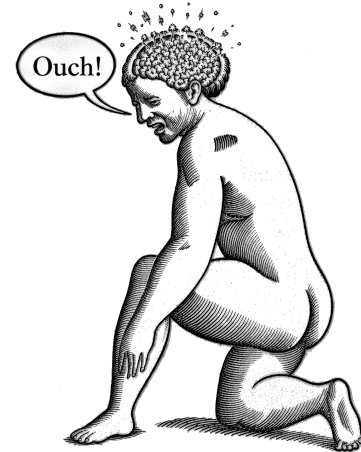
physical body states to subjective experience.

1 Pain: a body state or mental state?

What might an intuitive theory of pain look like? Folk intuitive theories resemble early scientific theories in many domains. For example, adult intuitive physics resembles pre-Newtonian impetus theory (McCloskey, 1983), and adult folk biology recapitulates pre-Darwinian taxonomies (Atran, 1990). We therefore look to historical, scientific, and philosophical theories of pain as a source of hypotheses about what intuitive theories of pain might look like.

Theories of pain can be arranged along a continuum from *purely bodily* (pain is a state of the body, located in a body part, fully determined by tissue damage and therefore potentially observable to outsiders) to *purely mental* (pain is a state of the mind and a subjective experience that is fundamentally private). There are two main reasons to organize around this axis. First, the major scientific and philosophical theories of pain disagree precisely on this dimension. Second, lay people’s views of pain also vary along this dimension (both as stable individual differences in tendencies toward body- vs. mind-centric positions; Borg et al., 2021; and within a person across contexts; Borg et al., 2020; Salomons et al., 2022). Figure 1 visualizes this axis and previews where each of the three causal models we formalize in § 2 sits.

Consider first the leftmost pole in Figure 1, the purely body-centric view. This view is motivated by three threads of evidence: historical philosophical theories of pain, historical scientific and clinical theories of pain, and empirical evidence that laypeople might hold this kind of view. In the 17th century, the Cartesian specificity theory of pain posited that pain is a direct readout of tissue damage (Descartes, 1664). The biomedical descendants of this theory considered pain a symptom secondary and proportional to tissue pathology, placing physical findings at the center of clinical assessment (see van der Linden et al., 2025 for review). Similarly, empirical work from experimental philosophy has shown that people sometimes seem to consider pain a property of an appendage rather than a

Body-centric theories**Mind-centric theories****Figure 1**

The body–mind axis, along which theories of pain vary. The leftmost pole depicts body-centric theories of pain; the rightmost depicts mind-centric theories. Our Physicalist, Integrated, and Experientialist models are positioned along this axis. Leftmost cartoon element adapted from Descartes (1664); rightmost generated using ChatGPT (OpenAI, 2026a).

mind (Sytsma, 2010). Additionally, the neural correlates of pain attribution offer another piece of converging evidence that people might hold a body-centric view of pain: attributing pain to another person engages a brain region adjacent to but distinct from the right temporoparietal junction (the canonical seat of mental-state attribution; Saxe and Kanwisher 2003). Crucially, this same region also responds when attributing body states like hunger to other people (Saxe & Powell, 2006). The recruitment of body-state rather than mental-state machinery when reasoning about another's pain suggests that people might view pain as a body state, and so the body-centric pole might best describe people's

intuitions.

Consider the rightmost pole in Figure 1, the purely mind-centric view. This view finds support from the same three threads of evidence. According to some philosophical views, pain is an inherently subjective experience: if pain is not felt, it does not exist (Kripke, 1980; Nagel, 1974). On this view, pain is fundamentally personal and private, and first-person report is the authoritative evidence. Some scientific views support this position as well. Melzack’s neuromatrix theory (Melzack, 1999) proposes that pain is generated by a widely distributed brain network whose activity can but need not be triggered by peripheral inputs—as illustrated most dramatically by phantom-limb pain, in which people experience centrally-driven, vivid pain in limbs that no longer exist. The mind-centric view has clinical analogues in the view that pain is “whatever the experiencing person says it is, existing whenever and wherever the person says it does” (McCaffery, 1968) and the modern IASP definition of pain as a subjective experience that need not be tied to actual or potential tissue damage (Raja et al., 2020), which is now the dominant view in the modern clinical pain literature. Lay people sometimes seem to endorse versions of this view, largely agreeing that pain reports and behaviors can indicate pain even if there is no corresponding tissue damage to explain it (e.g., see the results on the “Unexplained stomach pain” vignette in Salomons et al., 2022).

Between these two poles sits a family of integrative theories that treat pain as arising from a more complex interplay of bodily and mental factors. The landmark gate-control theory (Melzack & Casey, 1968; Melzack & Wall, 1965) established that the relationship between tissue damage and pain experience is not fixed (Wall, 1979): psychological, attentional, and contextual factors modulate pain signals at the spinal cord, such that damage and experience can dissociate. The biopsychosocial framework (Engel, 1977; Gatchel et al., 2007; Waddell, 1987) generalizes this idea into a clinical paradigm in which sensory and psychological components combine to produce and modulate pain experience. Under these kinds of integrated accounts, multiple sources of evidence

(physical findings, behavioral evidence, and first-person report) are valid.

The theories of pain surveyed above are verbal, qualitative, and descriptive: they suggest that this body-mind axis is an important dimension in scientific, clinical, and lay views of pain, but these theories do not specify how their commitments would cash out as quantitative predictions of pain or generalize to related tasks like, e.g., inferring motives behind pain reports. Translating verbal theories into computational causal models increases quantitative precision, spells out assumptions, and makes clear the consequences for generalizing to related tasks.

We implement instances of these three families of theories (body-centric, mind-centric, and integrated) as causal models that take as input vignettes describing the sources of evidence on which the theories disagree—physical evidence vs. first-person pain reports. This allows us to test the models against human judgments. We first formalize the models (§ 2), then describe the experimental tasks and stimuli on which they are tested (§ 3), and then turn to the behavioral results (§ 4).

2 Computational Model

2.1 The intuitive theory: abstract causal principles

We take as a starting point the idea that intuitive theories can be captured through causal models (Gerstenberg & Tenenbaum, 2017; Gopnik et al., 2004; Pearl, 2000), and that people use their background causal knowledge about the pain domain to synthesize a causal graph for any particular reasoning problem involving pain (Tenenbaum et al., 2007). Motivated by the theories of pain surveyed in Section 1, we propose four abstract causal principles that background knowledge about pain might contain. Each of these principles constrains the space of possible causal graphs that might be synthesized for this task. We first overview these principles, and later, in § 2.2, we articulate how committing to subsets of these principles motivates different causal graphs instantiating points along the mind-body axis.

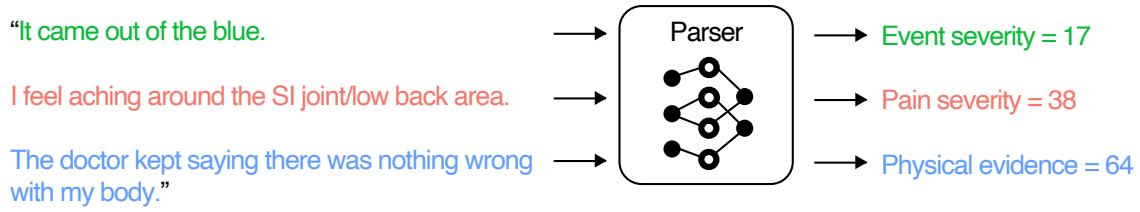
These four principles are as follows:

1. *Physical events that happen to the body can cause problems in the body.* This principle captures the intuition that physical events can cause injury (e.g., kicking a rock can cause a broken toe) or illness (e.g., eating a berry can cause sickness) and thus initiate the onset of pain.
2. *Physical problems in the body leave physical evidence.* This principle captures the intuition that bodily damage leaves behind evidence that others can observe (e.g., through visual inspection, imaging, diagnostic tests, etc).
3. *Physical problems in the body cause pain.* This principle captures the intuition that problems in the body (e.g., a broken toe, a sickness) cause pain. In other words, pain signals a problem in the body.
4. *People verbally report their experience of pain.* This principle captures the broad expectation that people can, and do, introspect and report on their pain experiences. Experiencing pain causes pain reports.

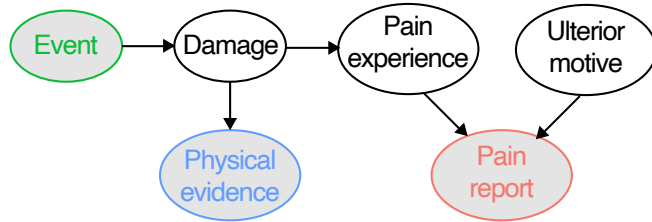
It is possible that people’s intuitive theories of pain include all, some, or none of these principles. For example, people might consider pain to be a state of the body, perhaps inaccessible to introspection and self-report, and hold principles 1-3 without holding 4. Our modeling approach is to cash out theories’ variation along the body–mind axis as commitment to different subsets of the four principles above. The theories surveyed in Section 1 inspire the body–mind axis, but our work does not seek to instantiate faithful renditions of any particular theory—actual philosophical theories are verbal, nuanced, and in some ways underspecified in terms of their predictions on our particular task (see Corns, 2017, for review). Instead, we instantiate computational models particular to our task that capture the spirit of the body–mind axis, along which the theories of pain in § 1 disagree.

First, we explain how we implemented these principles in causal graphs. Then, we introduce the Integrated model, which occupies a middle region between purely body-centric views of pain, and purely mind-centric ones (see Fig. 1); the Physicalist

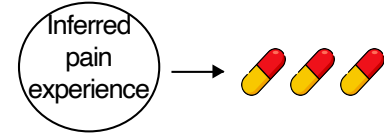
a) Stimulus processing into model input



b) Causal graph for Integrated model (including motive extension)



c) Observer prescribes painkiller based on inferred severity of speaker's pain experience

**Figure 2**

a) Information flows from left to right. The vignette stimulus consists of an inciting physical event, a pain report, and a description of any physical findings. An LLM parses the vignette text into numerical ratings of the severity of the event, pain report, and physical findings, which serve as model input. b) Conditioned on these observations (gray shading), the model performs inference over the latent variable in the causal graph. c) From the inferred pain experience, the model predicts how much painkiller participants will prescribe.

model, implementing the extreme of the body–mind axis; and the Experientialist model, which instantiates the mind-centric extreme of the body–mind axis by considering the pain report to be the only valid evidence of pain experience.

Implementation of the causal principles

These four principles define a causal graph (see Fig. 2). To implement these principles as causal relations, we formulate all of the variables (e.g., events, damage, pain experience, evidence, pain report) on the same, commensurate scale: severity. There is good reason to do this. Severity is one of the primary dimensions of pain experience (Melzack, 1975); pain is frequently assessed in terms of its severity (e.g., rating pain intensity along a 1-10 scale is common clinical practice); and severity can reasonably apply

to events, damage, evidence, pain, and pain reports alike.

To implement principle 1, the idea that events can cause physical problems, we relied on the intuition that more severe events tend to cause more severe problems (e.g., falling down the stairs is more likely to cause more damage than waking up in the morning). We sample the severity of the resulting problem in the body from a normal distribution centered on the severity of the event: $\text{severity}_{\text{problem}} \sim \mathcal{N}(\text{event}, \sigma_{\text{event}}^2)$. The variance parameter σ_{event} captures the intuition that even benign events can sometimes cause problems, and even severe events can sometimes leave the body unharmed, so the relationship is probabilistic rather than deterministic.

To implement principle 2, the idea that physical problems leave physical evidence, we relied on the intuition that more severe problems tend to leave more evidence (e.g., a more severe broken bone leaves more apparent evidence, e.g., on X-rays). We sample the severity of the physical evidence from a normal distribution centered on the severity of the problem, $\text{evidence}_{\text{physical}} \sim \mathcal{N}(\text{severity}_{\text{problem}}, \sigma_{\text{evidence}}^2)$. The parameter σ_{evidence} controls how tightly or loosely physical evidence reflects the severity of the underlying problem.

To implement principle 3, the idea that more severe problems produce more severe pain, we similarly sample the pain experience from a normal distribution centered on the severity of the problem, $\text{experience} \sim \mathcal{N}(\text{severity}_{\text{problem}}, \sigma_{\text{pain}}^2)$. The parameter σ_{pain} controls the strength of the relationship between magnitude of experienced pain and the severity of the underlying problem.

Finally, to implement principle 4, the idea that verbal reports track the pain experience, we sample the reported pain from a normal distribution centered on the experienced pain, $\text{report} \sim \mathcal{N}(\text{experience}, \sigma_{\text{report}}^2)$. The parameter σ_{report} controls the accuracy with which people are expected to verbalize the severity of their pain.

2.2 The Integrated model

The Integrated model commits to all four causal principles and therefore considers events, physical evidence, and pain reports as valid sources of evidence. We test the model

on three modes of reasoning: forward prediction from physical evidence, inverse inference of pain experience from multiple evidence sources, and detection of ulterior motives behind inflated pain reports.

Prediction

The Integrated model supports pain prediction from the inciting event and physical evidence. To predict pain, the model conditions on the observations and infers the subsequent pain experience:

$$\begin{aligned}
 P(\text{experience} \mid \text{event}, \text{evidence}_{\text{physical}}) &\propto \int P(\text{experience} \mid \text{severity}_{\text{problem}}) \\
 &\quad \cdot \underbrace{P(\text{evidence}_{\text{physical}} \mid \text{severity}_{\text{problem}})}_{\text{evidence (effect)}} \\
 &\quad \cdot \underbrace{P(\text{severity}_{\text{problem}} \mid \text{event})}_{\text{event (cause)}} d\text{severity}_{\text{problem}}.
 \end{aligned}$$

by marginalizing over $\text{severity}_{\text{problem}}$. This propagates information forward from event and evidence to pain experience. We test the model’s forward prediction in Experiment 1.

Inference and intervention

When pain reports are available, the Integrated model inverts the causal graph to infer the latent pain experience. Mathematically, inference is the same as prediction—with one additional observation, the pain report, entering as a likelihood term:

$$\begin{aligned}
 P(\text{experience} \mid \text{event}, \text{evidence}_{\text{physical}}, \text{report}) &\propto P(\text{report} \mid \text{experience}) \\
 &\quad P(\text{experience} \mid \text{event}, \text{evidence}_{\text{physical}}).
 \end{aligned}$$

Now, the Integrated model integrates over the pain report, event, and physical evidence to infer a posterior over true underlying pain experience. From this posterior, the model chooses a dose of painkiller expected to maximize utility by appropriately treating the underlying pain, no more and no less (see § 3.2). We test inverse inference and the resulting painkiller prescription in Experiment 2.

Motive inference

A third task this model might support is detecting ulterior motives behind inflated pain reports (e.g., wanting attention, time off work or school, etc). We ask whether the same model of bodily pain and pain reports that supports predicting and inferring pain also supports inferring this alternative hidden cause of pain reports. If it does, then a portion of people’s motive suspicions should fall out as a direct consequence of the intuitive theory of pain.

To allow the model to infer motives, we add a minimal extension: a binary latent variable $M \in \{0, 1\}$ representing whether the patient has an ulterior motive for inflating their report. The report distribution becomes

$$\text{report} \sim \mathcal{N}(\text{experience} + \delta \cdot M, \sigma_{\text{report}}^2),$$

such that when $M=1$, the patient inflates their report by a fixed amount δ ; when $M=0$ the report follows the original distribution. The body–mind causal structure is untouched. We chose this minimal implementation to test how much traction the intuitive theory of pain provides on its own, without a rich model of the many cues that influence suspicion of motive or the sophisticated ways in which someone with an ulterior motive might be expected to alter their pain report. It is possible to build richer models of how people expect ulterior motives to operate, but doing so would not help us test the contribution of the intuitive theory of bodily pain on its own.

Given the full evidence set $\mathcal{O} = \{\text{event}, \text{evidence}_{\text{physical}}, \text{report}\}$, the observer jointly infers M and the pain experience:

$$P(M, \text{experience} \mid \mathcal{O}) \propto P(\text{report} \mid M, \text{experience}) \cdot P(\text{experience} \mid \text{event}, \text{evidence}) \cdot P(M).$$

Crucially, M introduces an alternative cause for elevated reports: a high report could be produced either by genuinely high pain or by an ulterior motive. As suspicion of motive rises (as $P(M=1 \mid \mathcal{O})$ rises), the attributed pain decreases (the marginal posterior

$P(\text{experience} \mid \mathcal{O})$ shifts downward). This model therefore predicts that suspicion of motive *explains away* pain reports. We test this prediction in Experiment 3.

2.3 The Physicalist model

Unlike the Integrated model, which considers pain reports valid evidence of a pain experience, the Physicalist model only considers physical evidence. A purely body-centric view of pain holds that pain is a readout of bodily damage, and so physical evidence is the most—or, in the most extreme version of this view, the *only*—relevant type of evidence. This view can also be motivated through distrust of self-report: people might expect others to exaggerate pain reports in practice (e.g., Boring et al. 2022), and so discount them in favor of bodily evidence. Either motivation arrives at the same model: the Physicalist model drops principle 4 from the causal graph.

The Physicalist model conditions only on what can be observed of the body $\mathcal{O}_{Physicalist} = \{\text{event}, \text{evidence}_{\text{physical}}\}$. When no report is available, the Physicalist model makes the same predictions as the Integrated model. When a report is available, the Physicalist model disregards it, predicting that pain attribution and prescriptions will be sensitive to physical events and evidence but not to reported pain. The motive extension applies to the Physicalist model, but because M is independent of the event and evidence, the Physicalist model’s motive posterior collapses to the prior on M .

2.4 The Experientialist model

In contrast to the Physicalist model, which only considers physical evidence, the Experientialist model considers pain reports to be the only valid evidence of a pain experience. This model represents a purely mind-centric view of pain—“pain is what the person says it is” (McCaffery, 1968)—in which the first-person report is the only authoritative evidence and physical observations are irrelevant. Therefore, this model cannot predict pain without a pain report. When a report is available, the model’s inferences about pain experience depend solely on the report:

$$P(M, \text{experience} \mid \text{report}) \propto P(\text{report} \mid M, \text{experience}) \cdot P(\text{experience}) \cdot P(M).$$

To assess ulterior motives, this model relies on the pain report itself. Because an ulterior motive inflates reports by δ , the motive hypothesis shifts the expected report upward relative to the no-motive hypothesis, making a high report more probable under $M=1$ than under $M=0$. The motive posterior then follows from Bayes' rule,

$$P(M=1 \mid \text{report}) = \frac{P(\text{report} \mid M=1) P(M=1)}{\sum_{m \in \{0,1\}} P(\text{report} \mid M=m) P(M=m)},$$

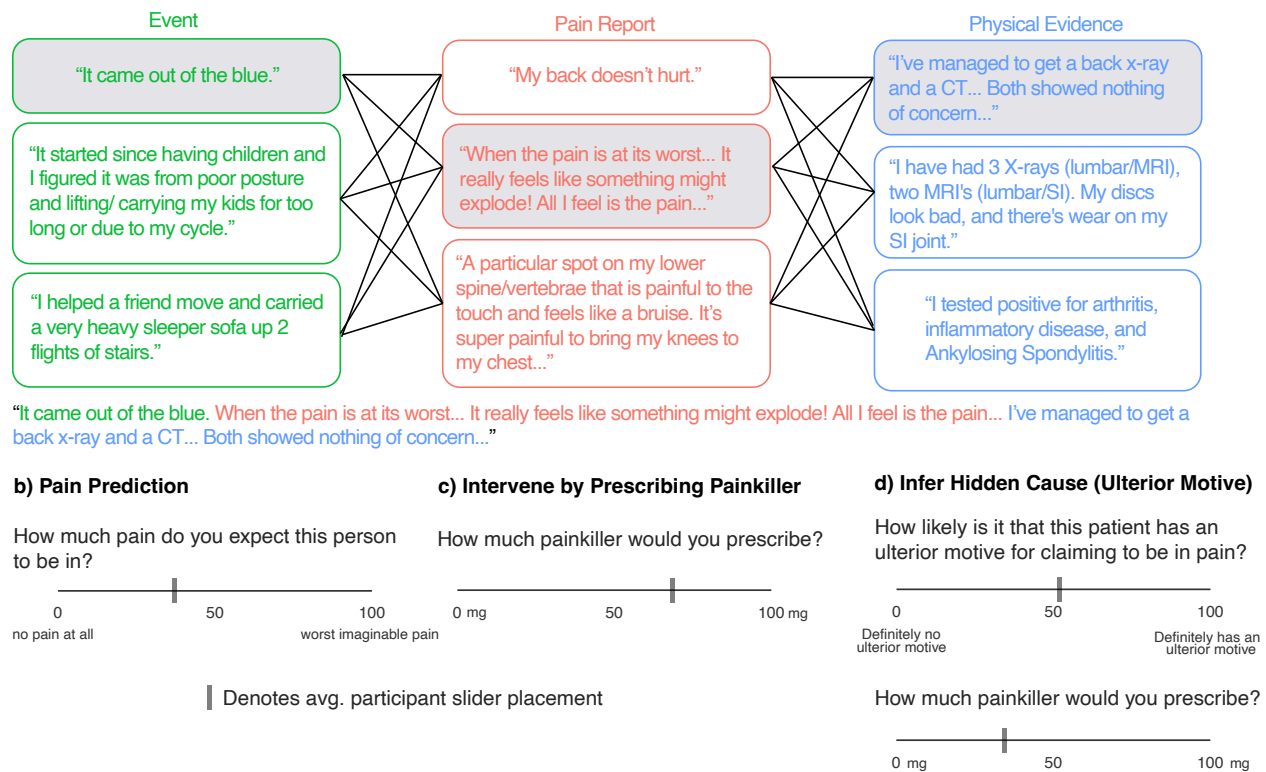
and since the likelihood ratio $P(\text{report} \mid M=1)/P(\text{report} \mid M=0)$ grows with the report, so does the suspicion of motive. With no event or physical evidence against which to gauge the report, the Experientialist model can only judge a report as inflated relative to its fixed prior over pain experience, so the larger the pain report, the more the model suspects an ulterior motive.

Holding a non-zero prior over $P(M=1)$ may appear contrary to the spirit of the Experientialist view and the “pain is what the person says it is” adage. A model with $P(M=1) = 0$ will always make the trivial prediction of no ulterior motive. In the interest of generating non-trivial predictions against which to compare the Integrated model, we gave the Experientialist model the same prior over motives as the Integrated model.

Having specified three models spanning the body-centric to mind-centric spectrum of pain theories, we next describe the stimuli on which these models are tested, followed by the study-specific implementation details.

3 Study Design

To study people's intuitions about pain, its causes, and evidence, we developed three behavioral tasks. In each, participants prescribe doses of an imaginary perfect painkiller, and the prescribed dosage serves as a proxy for the amount of pain that participants attribute to the patient (we validate this proxy directly in Experiment 1). The experiments

a) Study Design**Figure 3**

a) Our stimuli are generated by permuting descriptions of events, pain, and physical evidence. b) In Experiment 1, vignettes contain only events and physical evidence (no pain report), and participants are asked to predict how much pain they expect the patient to be in. c) In Experiment 2, vignettes consist of events, physical evidence, and a pain report, and participants are asked how much painkiller they would prescribe. d) In Experiment 3, participants are asked to infer whether the patient has an ulterior motive for claiming to be in pain. Comparing painkiller prescription in c) to d) shows how suspicion of motive explains away pain, and thus, reduces painkiller prescriptions.

differ in the evidence observers receive and the additional judgments they provide.

Experiment 1 provides only the inciting event and physical evidence (no first-person pain report), and asks for both an explicit pain estimate and a prescription. Experiment 2 additionally provides the patient's own first-person pain report. Experiment 3 introduces

the possibility of ulterior motives and asks for explicit judgments of motive likelihood.

3.1 Stimuli design & curation

To construct ecologically-valid stimuli on which to test people’s intuitions, we drew from Reddit posts and medical “simulation” training materials. From each post or simulation, we manually extracted language containing the description of the inciting event (e.g., “A 45lb barbell fell on my head in the weight room.”) or non-event (e.g., “It came out of the blue.”), a description of the pain quality and intensity (e.g., “It basically feels like someone is stabbing an icepick from the base of my skull to the adjoining eye.”), and any physical evidence or findings indicating a problem in the body (e.g., “They ultrasounded my ankles and MRI’d my hands and found excess joint fluid.”), or lack thereof. See SM Section S1 for details.

Stories depicted pain in eight regions of the body (Back, Dental, Diffuse / Generalized Everywhere, Female Genital & Pelvic, Male Genital & Pelvic, Hands & Finger Burns, Head, Abdominal & GI). To create a set of stories that varied independently along the dimensions of interest, we permuted the event, pain, and evidence descriptions within, but not across, body-region categories, creating a large set of vignettes systematically varying in the severity of the event, the pain, and the evidence described. This process generated naturalistic vignettes spanning a variety of types and categories of pain while independently varying the variables of interest.

3.2 Inputting stimuli to the model

To generate model predictions on each vignette, we needed a way to map vignette text into numeric values for the observable variables event, evidence_{physical}, and report. Mapping from vignette stimuli to model inputs typically requires gathering human ratings (or priors) about features of the vignettes (as in Radkani et al., 2025). Instead, we computed model predictions directly from the text (without human ratings as an intermediary) by using a Large Language Model (LLM) to rate the severity of each component on a scale from 0–100, and we used the resulting numeric ratings as inputs to

our causal models. Specifically, we used Llama-3.1-8B-Instruct (Grattafiori et al., 2024). The full prompts and the validation of the LLM ratings against two human coders are reported in SM Section S2.1. Thus, our models are “stimulus-computable” (Raz et al., 2025).

3.3 Direct-LLM baseline

In addition to the Integrated, Physicalist, and Experientialist models, we also compared human judgments against a *Direct-LLM* baseline. We used the same Llama-3.1-8B-Instruct model for this baseline as for stimulus parsing (see SM Section S2.2 for details). Matching the LLM used for parsing and for the Direct-LLM baseline means that any advantage of a causal model over the Direct-LLM baseline can be attributed to the structure of the causal model. It also lets us assess how much of human judgment can be explained by the features of the full vignette text other than the three severity dimensions alone, since the Direct-LLM baseline can pick up on semantic content and specific linguistic cues beyond what is available to the causal models. Critically, this comparison is insufficient for making claims about LLMs’ capacities to capture human judgments in general; we simply ask what this specific model does when responding directly to the same prompts that humans receive.

3.4 Model implementation details

Parameters

Our models were implemented in Memo, a probabilistic programming language for expressing computational cognitive models (Chandra et al., 2025). This language requires describing a discrete domain over which the model operates. We used the domain of integers in $[0, n]$, with $n = 50$. We set the standard deviations of the normal distributions to $\sigma_{event} = \sigma_{evidence} = \sigma_{pain} = \sigma_{report} = \frac{n}{10}$. The constant capturing pain report inflation based on an ulterior motive was $\delta = n/5$. The prior on motive depended on the experiment: for Experiments 1 and 2, we set $P(M=1) = 0$ to render M dormant; for Experiment 3, the task framing introduces the possibility of motivated reporting, and so we

set $P(M=1) = 0.01$ to reflect that ulterior motives, while now possible, remain rare. All parameters were pre-registered and fixed prior to data collection.

Although these parameters influence the exact predictions that the models make, their effect is minimal: the qualitative patterns of the model predictions follow from the causal relationships and structure of the model, rather than from the exact values of the parameters themselves. For each experiment, we also present the results of fitting each model’s parameters.

From inferred pain to prescription

Each of the three models produces a posterior over experience. Mapping from inferred pain to a painkiller prescription requires considering the costs and benefits of different possible interventions (operationalized as the dosage of a painkiller). We implemented this mapping using a linear cost function

$f(\text{dose}_{\text{painkiller}} \mid \text{experience}) = -|\text{experience} - \text{dose}_{\text{painkiller}}|$, which assumes that the cost of under- and over-prescribing painkillers is symmetric. Future extensions might incorporate considerations like addiction risk via an asymmetric cost function. We model the observer’s decision about how much painkiller to prescribe as a softmax over the resulting utilities, parametrized by a temperature parameter $\tau = 0.00001$ (Luce, 1959). Because the utility function is shared across the Integrated, Physicalist, and Experientialist models, differences in predicted prescriptions across the three models come entirely from differences in their inferred posteriors over pain experience.

4 Behavioral Experiments

Having formalized three intuitive theories of pain as causal models, we turn to testing their predictions against human judgments. Across three experiments (Fig. 3b-d), we look for classical signatures of causal reasoning: forward prediction (Experiment 1), inverse inference from multiple sources of evidence (Experiment 2), and inferring a hidden cause to explain away evidence (Experiment 3). These experiments vary the evidence that the participants observe and the judgments that they are asked to make.

All materials, predictions, analyses, and sample sizes were preregistered. Preregistration packages, stimuli, model code, and human data for all three experiments are available on our OSF repository:

https://osf.io/3n4zf/?view_only=8464679768ac419d9403cf8d8e397da2. An interactive visualization for browsing through the vignettes and our results from each experiment is also available at https://marleneberke.shinyapps.io/intuitive_theory_of_pain_browser/.

The experiments were approved by MIT’s Institutional Review Board (protocol #0602001639). Claude Code was used to assist in writing data analysis code (Anthropic, 2026).

4.1 Experiment 1: Pain Prediction

We first tested people’s ability to predict how much pain someone would experience given an inciting event and physical evidence of a problem in the body. If people hold an intuitive theory of pain as arising from physical problems, then they should be able to use this theory to predict pain, even in the absence of a pain report. By contrast, if people consider pain to be a subjective experience that “is what the person says it is,” then people might not be able to predict pain from physical events and physical evidence alone, or if they do so, they might rely on heuristics rather than the same causal model that they use when a pain report is available. Thus, Experiment 1 tests a critical prediction of the causal model: it tests whether people can predict pain from physical causes, in the absence of any first-person report.

Participants

We recruited 48 U.S.-based participants via Prolific, who completed the experiment. The study was approved by MIT’s Institutional Review Board. As pre-registered, we excluded participants who reported paying less than 90% attention on the task ($n = 11$).

Stimuli

Stimuli for Experiment 1 consisted of vignettes describing only the inciting event and physical evidence, with no first-person pain report. Because a first-person narrator

describing an injury might be expected to mention pain, we rewrote events and evidence in the third person to avoid the lack of pain report implying a lack of pain. We crossed six events with six evidence descriptions in each of the eight pain categories, yielding 288 unique vignettes split into six sets of 48, with no event or evidence description repeated within a set.

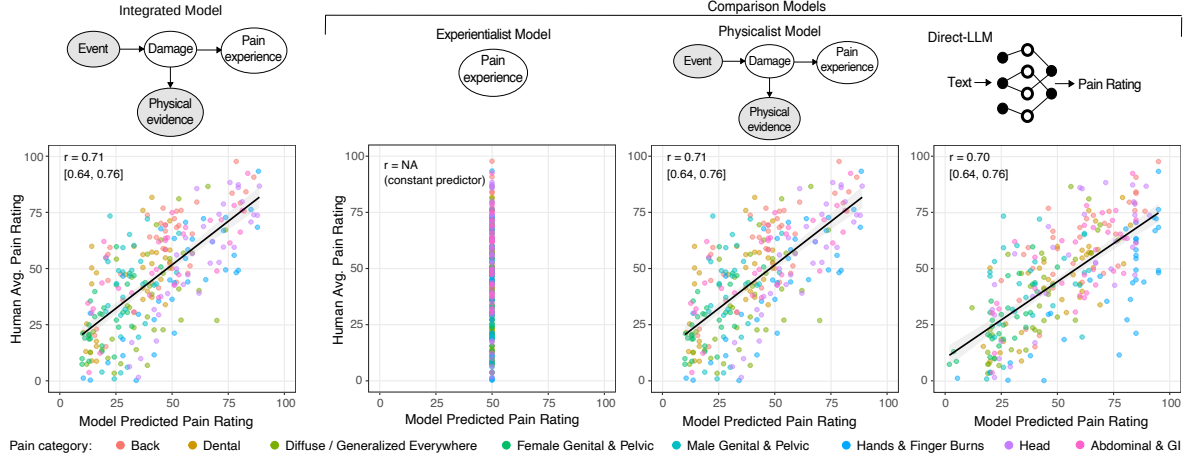
Procedure

Participants were told that they would read stories describing a medical issue happening to someone. For each story, they were asked, “How much pain do you expect this person to be in?” and they responded using a slider from 0 “no pain at all” to 100 “worst imaginable pain.” Additionally, participants were asked, “How much painkiller would you give them?” and they responded using a slider from 0 to 100mg. Participants were told that “the painkiller can only treat real pain, not pretend or imaginary pain” and that “if you don’t give enough painkiller, the person will still suffer from untreated pain, and if you give too much painkiller, you risk hurting their liver and causing more pain later down the line.” Participants were also provided with a checkbox for indicating, “this scenario does not make any sense.” Each participant was randomly assigned to one of six sets of 48 vignettes, presented in random order. We collected eight responses per vignette. After the main task, participants completed a survey about their attention during the task, beliefs about pain, painful experiences, and their demographics.

Results

As preregistered, we excluded vignettes that >25% of participants indicated did not make sense ($n = 19$), leaving 269 vignettes for analysis. Participants’ pain estimates and prescribed dosages were tightly correlated ($r = 0.98$), validating the use of prescription as a proxy for attributed pain in Experiments 2 and 3. Our analyses from here on focus on pain ratings; the same analyses on prescriptions are reported in SM Section S3.1.

The full causal model predicted participants’ average pain ratings with correlation $r = 0.71$; $CI_{95\%} = (0.64, 0.76)$, comparable to the split-half reliability of the human ratings

**Figure 4**

Results on Experiment 1: Pain Prediction. Scatterplots show the relationship between average participant pain ratings (y-axis) and each model’s predictions (x-axis). Each point represents one vignette, color-coded by pain category. Pearson correlations and 95% bootstrapped confidence intervals are displayed. Because no first-person pain reports are available on this task, the Physicalist model is identical to the Integrated model, and the Experientialist model has nothing to observe (resulting in a constant prediction). The Direct-LLM baseline predicts pain directly from the vignette without an explicit causal model.

themselves ($r = 0.68$; $CI_{95\%} = (0.63, 0.72)$; single-rater inter-observer agreement $ICC(2,1) = 0.41$).¹ Because there are no first-person pain reports in this task, the Physicalist model makes identical predictions to the Integrated model, and the Experientialist model predicts a constant (the mean of the prior over pain). See Fig. 4.

The Direct-LLM baseline predicted human pain ratings with $r = 0.70$; $CI_{95\%} = (0.64, 0.76)$, comparable to the full causal model (see Fig. 4, rightmost panel). The two models are strongly correlated with each other ($r = 0.78$), but diverge

¹ Fitting the parameters of the Integrated model (σ_{event} , $\sigma_{evidence}$, σ_{pain}) to maximize Pearson correlation with humans’ pain ratings only minimally improved fit: $r = 0.71 \rightarrow 0.72$ at $\sigma_{event} = 2.76$, $\sigma_{evidence} = 2.20$, $\sigma_{pain} = 0.34$.

systematically on vignettes where the inciting event and the physical evidence are at odds. As the discrepancy between event and evidence severity grows, the Direct-LLM baseline’s divergence (absolute error) from human ratings increases ($r = 0.17$; $CI_{95\%} = (0.05, 0.29)$, reliably positive), whereas the causal model’s does not ($r = -0.05$; $CI_{95\%} = (-0.17, 0.07)$, not distinguishable from zero). The divergence is sharpest on the most conflicting vignettes—those describing a severe inciting event (event severity $> 75/100$) that nonetheless left little or no physical evidence (physical evidence $< 25/100$; $n = 21$).² On these vignettes, the Direct-LLM baseline overestimates pain (mean 72.7/100) relative to both human participants (mean 51.1/100) and the causal model (mean 52.0/100), and it is correspondingly less accurate (mean absolute error 23.0 vs. 13.3 for the causal model; paired Wilcoxon $p = 0.035$).

As a complementary descriptive analysis, we fit a linear regression of average human pain rating on the LLM-rated event and evidence features. We found that, unlike the Integrated model (under fixed parameters), participants weighed physical evidence more than the event: $\beta_{evidence} = 0.46$ ($[0.39, 0.53]$) and $\beta_{event} = 0.29$ ($[0.23, 0.36]$). Additional pre-registered analyses testing the relative contributions of events and physical evidence are available in SM Section S3.1.

Together, these findings establish that people make predictions about how much pain a physical event and subsequent physical damage might cause, even without any information about first-person subjective experience. This finding is consistent with both the physicalist view of pain and with the integrated causal model of pain. However, it is inconsistent with the view that pain is a purely subjective state detached from the physical world (in which case, it should be impossible to infer pain experience without a pain

² An example of such a vignette is: “They burnt their thumb and forefinger on a pan while cooking. You can’t see any damage or anything visually wrong with their hand.” The Direct-LLM baseline gives a pain rating of 77.5/100, while the causal model gives 42.0/100 and participants averaged 11.7/100; $CI_{95\%} = (2.5, 20.8)$.

report).

The next two experiments test how people integrate first-person reports of pain with physical evidence, and whether they use the same causal machinery to do so.

4.2 Experiment 2: Pain Inference

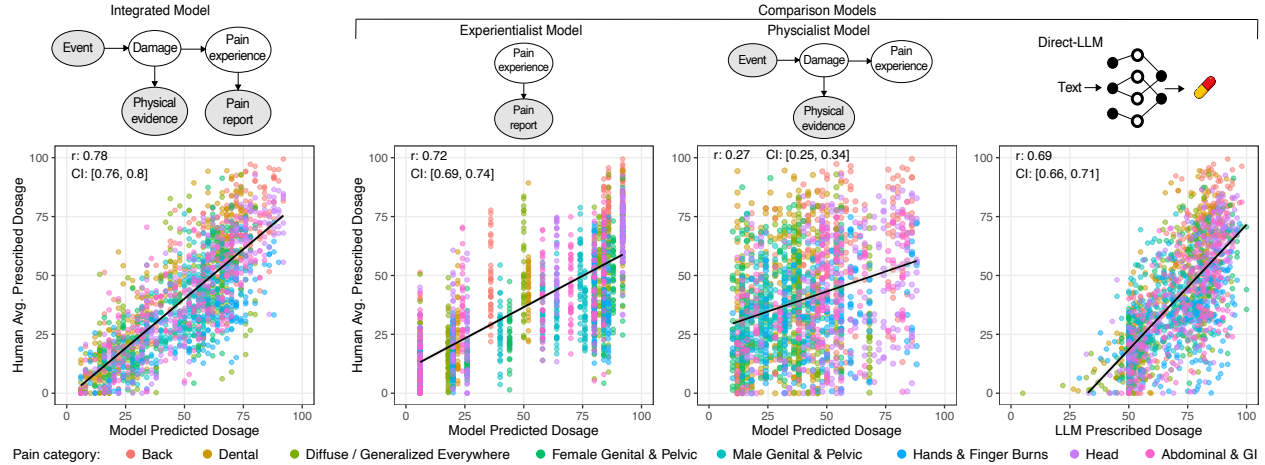
Experiment 1 tested people’s ability to predict pain from physical causes. But if people truly have a causal model of pain, they should be able to use it for other tasks, like intervention. Experiment 2 tests how people integrate multiple sources of evidence to infer how much pain someone is in and how much painkiller to prescribe in order to treat that pain. Experiment 2 tests the critical prediction of the Integrated model: People use Bayesian inference to integrate both physical evidence and verbal pain reports so as to infer the latent pain experience. In contrast, the Experientialist model predicts that only verbal pain reports matter, while the Physicalist model predicts that only physical evidence matters. On this experiment, the Physicalist, Experientialist, and Integrated causal accounts each make distinct, non-trivial predictions, and any of them could, in principle, win.

Participants

We recruited 288 U.S.-based participants via Prolific, who completed our experiment. As preregistered, we excluded participants who reported paying less than 90% attention, resulting in the exclusion of 12 participants.

Stimuli

Stimuli consisted of 1,728 vignettes, generated from permuting six events, six pain reports, and six evidence descriptions for each of eight pain categories. These vignettes were split into 36 sets of 48 (six from each body category), with no vignette element repeated within a set.

**Figure 5**

Experiment 2 results. Scatterplots show the relationship between the average painkiller dosage that participants prescribe (y-axis) and each model's predictions (x-axis). Each point represents one vignette, color-coded by pain category. Pearson correlations and 95% bootstrapped confidence intervals are displayed.

Procedure

The procedure was identical to Experiment 1 except for three changes. First, the task was reframed such that participants were asked to play the role of a healthcare provider whose sole job was to manage the patient's current pain ("diagnosing and treating injuries or ailments was the job of another doctor"). Participants responded to each vignette by prescribing the painkiller dosage to best treat the patient's pain (using a slider from 0–100mg). Unlike in Experiment 1, they did not separately rate pain. Second, because Experiment 2's stimulus pool was larger, each participant was randomly assigned to one of 36 sets of 48 vignettes (vs. 6 in Experiment 1), presented in random order; we again collected eight responses per vignette. Third, participants were asked additional questions about how long the pain would last if untreated and how long the body would take to heal; these responses were collected for use in future work and are not analyzed here.

Results

We deviated from the preregistration by replacing the preregistered model (which treated events as Bernoulli random variables that increase the probability of a problem in the body) with the model presented above, which treats events as having magnitudes. The change provides a more parsimonious causal structure and a more direct interpretation of event rating without affecting the results. (The preregistered model and the Integrated model presented here both yield $r = 0.78$; see SM Section S3.2.1).

As in Experiment 1, we excluded trials for which $> 25\%$ of participants indicated that the scenario did not make sense, resulting in the exclusion of 19/1,728 trials.³ For each vignette, we averaged participants' responses and compared the results against the models' predictions. The Integrated model predicted participants' prescribed painkiller dosages with $r_{Rx} = 0.78$; $CI_{95\%} = (0.76, 0.80)$,⁴ exceeding the split-half reliability of the human ratings themselves ($r = 0.72$; $CI_{95\%} = (0.70, 0.73)$; single-rater inter-observer agreement $ICC(2,1) = 0.40$), and indicating that people integrate both physical evidence and verbal pain reports into their treatment decisions (see Fig. 5).

To assess alternative hypotheses about how people infer the severity of another person's pain, we compared the Integrated model to alternatives. The Experientialist model predicted participants' painkiller prescriptions with high accuracy ($r_{Rx} = 0.72$; $CI_{95\%} = (0.70, 0.74)$; see Fig. 5), indicating that the pain report drove participants' prescriptions to a large extent. However, it failed to capture the variability in people's prescriptions for vignettes containing the same pain reports, resulting in the columns of points shown in Fig. 5. The Integrated model reliably outperformed the Experientialist

³ A representative excluded trial is the following: "I put my hand in lukewarm water. It hurts a little bit, like a mild sunburn. My skin looks pretty damaged."

⁴ Fitting the σ values and τ to maximize Pearson correlation with humans only minimally changed any model's fit: main $r_{Rx} = 0.78 \rightarrow 0.79$; Experientialist model unchanged at $r_{Rx} = 0.72$; Physicalist model unchanged at $r_{Rx} = 0.30$.

model ($\Delta r = 0.065$; $CI_{95\%} = (0.052, 0.079)$), indicating that, on average, participants took physical events and evidence into account when prescribing painkillers.

Conversely, if people do not trust subjective pain reports (perhaps because they expect that pain might be too difficult to accurately report or too prone to exaggeration; Boring et al., 2022), then the Physicalist model should better capture the patterns of people’s prescribing behavior. The Physicalist model predicted participant responses weakly ($r = 0.30$; $CI_{95\%} = (0.25, 0.34)$; see Fig. 5) and reliably worse than the Integrated model ($\Delta r = 0.48$; $CI_{95\%} = (0.44, 0.53)$), indicating that pain reports contribute substantially to people’s assessments of another’s pain.

The Direct-LLM was able to capture variance in participants’ responses across vignettes ($r_{Rx} = 0.69$; $CI_{95\%} = (0.66, 0.71)$), but it consistently over-prescribed, often prescribing 50mg or more of painkiller in cases where participants prescribed far less (see Fig. 5). The Integrated model reliably outperformed it ($\Delta r = 0.09$; $CI_{95\%} = (0.07, 0.12)$).

As a complementary descriptive analysis, we fit a linear regression of average human prescription on the LLM-rated event, evidence, and pain-report features (raw coefficients, 95% CIs in brackets). All three contributed, with pain report dominating: $\beta_{event} = 0.09$ $([0.06, 0.11])$, $\beta_{evidence} = 0.24$ $([0.22, 0.27])$, and $\beta_{report} = 0.51$ $([0.49, 0.54])$, $R^2 = 0.62$.

We also conducted additional pre-registered analyses at the participant level. We found that the Integrated model better describes the behavior of the majority of participants in a pairwise comparison than the leading alternative model, the Experientialist model ($n = 219/276$ participants; $p_{binom} < .001$). A three-way comparison that also includes the Physicalist model leaves this conclusion unchanged (see SM Section S3.2.2). Furthermore, participants’ self-reported explicit beliefs about the best ways to assess pain predicted which model best matched their behavior (see SM Section S3.2.2 for details).

4.3 Experiment 3: Ulterior Motive Inference

Experiments 1 and 2 suggested that people use an intuitive causal theory of pain to infer pain from observable evidence. Experiment 3 asks whether this same theory gains traction on a qualitatively different task: inferring hidden causes. Specifically, we test how well this causal model of pain can capture people’s suspicions of ulterior motives (e.g., attention-seeking) for pain reports. If people truly believe that pain arises from physical events and problems in the body, then when a pain report seems too high given the physical evidence, people may infer that there must be some other cause inflating the report. Thus, some portion of people’s suspicions of ulterior motives may fall out as a direct consequence of their intuitive theory of pain. Furthermore, if people are reasoning about motives, true pain, and pain reports through Bayesian causal reasoning, then suspicion of ulterior motives will explain away pain reports, leading to lower inferred pain and lower prescribed dosages on suspicious cases. By contrast, if people do not have a causal theory of pain, or if they have one but do not use it to reason about motives, then the model will fail to capture people’s suspicions of ulterior motives and the relationship between motive suspicion and pain inference.

Crucially, we do not assume the causal theory of pain is the only thing driving suspicions of ulterior motives. People certainly rely on other cues in the real world (e.g., patient demographics and personal history, the delivery of the pain report, etc.), and perhaps in our vignettes. Our aim is not to fully explain attribution of ulterior motives in pain reports. Instead, we aim to quantify how much of people’s motive suspicions falls out as a direct consequence of the intuitive theory of pain developed in this work and tested in Experiments 1 and 2.

Participants

We recruited 432 U.S.-based participants via Prolific. Per preregistered exclusion criteria, we excluded participants who self-reported $< 90\%$ attention ($n = 26$), reported any AI tool use ($n = 1$), or completed the task in < 20 minutes ($n = 13$), yielding a final

sample of 392 participants.

Stimuli

The 1,728 vignettes were identical to those in Experiment 2, with one design change: events and physical evidence were rewritten in the third person (as in Experiment 1) while pain reports remained in the first person. This reflects the common situation in which an observer (e.g., a clinician or juror) may have access to third-person confirmation of an inciting event or imaging findings, but only the patient’s own first-person account of the pain.

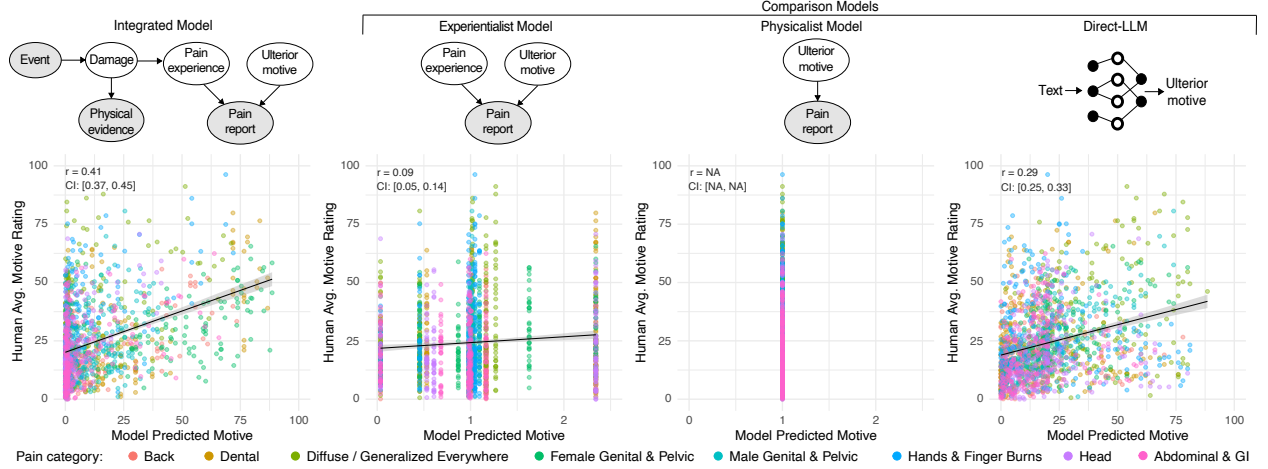
Procedure

The procedure was identical to Experiment 2 except for three changes. First, the instructions introduced the possibility of ulterior motives: “Not all patients coming to you are really in pain. Some patients have an ulterior motive for claiming to be in pain—for instance, they might want time off work or school, or want attention.” Second, instead of a single response per vignette, participants gave three: (1) “How likely is it that this patient has an ulterior motive for claiming to be in pain (e.g., wanting time off work or school, or wanting attention)?” (0–100; “Definitely no ulterior motive” / “Definitely has an ulterior motive”), (2) “How much pain is this person actually in?” (0–100; “No pain at all” / “Worst imaginable pain”), and (3) “How much painkiller would you prescribe?” (0–100mg). Third, each set of 48 vignettes was assigned to 12 participants (vs. 8 in Experiments 1–2).

Results

Motive judgments. The primary, pre-registered analysis tested how much of the variance in people’s motive judgments falls out as a consequence of their intuitive theory of bodily pain. If that intuitive theory of pain provides traction on motive inference, then the Integrated model’s posterior probability of ulterior motive should capture some of the variation in people’s motive ratings.

Participants’ motive judgments were less reliable than the pain and prescription judgments of the earlier experiments, even compared to prescription judgments (Rx)

**Figure 6**

Experiment 3 primary results. Scatterplots show the relationship between average participant motive ratings (y-axis) each model’s inferred motive (x-axis). Each point represents one vignette, color-coded by pain category. Pearson correlations and 95% bootstrapped confidence intervals are displayed.

collected within this experiment (split-half $r_{\text{motive}} = 0.57$; $CI_{95\%} = (0.55, 0.60)$; $r_{\text{Rx}} = 0.79$; $CI_{95\%} = (0.77, 0.80)$). The same pattern held for single-rater inter-observer agreement: $ICC(2,1) = 0.19$ for motive versus 0.38 for prescription. This lower reliability indicates that judging motives is a harder task with more inter-observer variability and a weaker signal for any model to capture.

Across the 1,711 vignettes that survived exclusions, the Integrated model predicted people’s judgments about the likelihood of the patient having ulterior motives moderately well ($r = 0.41$; $CI_{95\%} = (0.37, 0.45)$; see Fig. 6). For this experiment, fitting the parameters of the model further improved its fit to $r = 0.51$; $CI_{95\%} = (0.47, 0.54)$.⁵

As in Experiments 1–2, the Experientialist and Physicalist models failed to capture human judgments as well. The Experientialist model’s motive predictions (based solely on the pain report) weakly correlated with human motive ratings ($r = 0.10$;

⁵ This fit was achieved with parameter values $\sigma_{\text{event}} = 28.3$, $\sigma_{\text{evidence}} = 6.1$, $\sigma_{\text{pain}} = 10.7$, $\sigma_{\text{report}} = 3.3$, $\delta = 6.7$, $p_{\text{motive}} = 0.00045$ (LLM scale, 0–100).

$CI_{95\%} = (0.05, 0.14)$).⁶ The Integrated model performed reliably better than the Experientialist model ($\Delta r = 0.31$; $CI_{95\%} = (0.27, 0.36)$). As described in SM Section S2.2, the Physicalist model’s motive predictions collapsed to the prior and were uncorrelated with human ratings. Overall, the Integrated model captured people’s motive judgments in a way that neither alternative model could.

As a complementary descriptive analysis, we fit a linear regression of average human motive rating on the LLM-rated event, evidence, and pain-report severity. The dominant predictor was physical evidence; increasing evidence decreased suspicion of motive:

$\beta_{event} = -0.07$ ($[-0.09, -0.05]$), $\beta_{evidence} = -0.31$ ($[-0.33, -0.28]$), and $\beta_{report} = +0.05$ ($[0.03, 0.07]$), $R^2 = 0.32$. This pattern matches the intuition behind the Integrated model: that motive is suspected when the pain report cannot be supported by the available physical events and evidence.

Finally, the Direct-LLM baseline rated motive directly from the vignette, given the same instructions as participants. It could, in principle, make use of features of the vignette beyond what the structured models could (e.g., semantic content, tone). The Direct-LLM baseline reached a moderate correlation with human motive ratings ($r = 0.29$; $CI_{95\%} = (0.25, 0.33)$), but the Integrated model reliably outperformed it ($\Delta r = 0.12$; $CI_{95\%} = (0.05, 0.18)$).

The Integrated model and the Direct-LLM baseline capture complementary variance. After completing our pre-registered analyses, we then explored whether the LLM picked up on the same variance structure in people’s judgment of ulterior motives as the Integrated model; did the LLM, like the Integrated model, base its predictions on the mismatch between pain report and physical evidence, or other cues in

⁶ Fitting the parameters of the Experientialist model improved its fit only marginally, from $r = 0.10$ at default parameters to $r = 0.12$; $CI_{95\%} = (0.07, 0.16)$ at $\sigma_{event} = 2.5$, $\sigma_{pain} = \sigma_{report} = 10$, $\delta = 10$, $p_{motive} = 0.3$, confirming that the Experientialist model’s structure (not particular parameter settings) limits its ability to explain human motive judgments.

the vignettes? The two models’ predictions only weakly correlated with each other across vignettes ($r_{\text{main,LLM}} = 0.05$; $CI_{95\%} = (0.00, 0.10)$). A combined regression `human~main+LLM` achieved $R^2 = 0.24$, with the Integrated model uniquely explaining 16% of the variance, the LLM uniquely explaining 8% of the variance, and 1% shared. Thus the structured causal model and the LLM each contribute independently to motive judgments: the causal model captures suspicion that arises from a structured event–evidence–report mismatch, while the LLM captures additional suspicion grounded in other cues. This indicates that a substantial portion of people’s judgments of ulterior motives arise as a direct consequence of their intuitive theory of pain, and some from other features of the vignette text.

Our preregistration hypothesized that the Integrated model would perform poorly on 36 vignettes that reference illicit drug-use (e.g., “They partied hard, consuming large amounts of hard liquor and snorting cocaine.”), since in these vignettes participants might suspect drug-seeking as an ulterior motive for reasons that could not be captured by a mismatch in physical evidence and pain report. As preregistered, the Integrated model failed to identify these vignettes as suspicious (mean motive rating of 3 out of 100 on the 35 vignettes surviving exclusion), while the Direct-LLM baseline rated motive much higher on these (mean = 52), as did participants (mean = 49), suggesting that the LLM and participants picked up on semantic features in the vignette (e.g., “partied hard”, “snorting cocaine”) implying drug-seeking as a motive for elevated pain reports.

To further characterize how participants judged ulterior motives, we examined free-response answers to the question “How did you decide whether someone had an ulterior motive?” collected at the end of the experiment ($n = 426$ of 432 total participants responded; see SM Section S3.3.5 for coding details). About a third of participants explicitly described comparing pain report against physical event and/or evidence⁷ ($n = 142$ of 426). This strategy mirrors the logic of the Integrated model and suggests that

⁷ “I looked for inconsistencies or a lack of objective findings. If tests were normal and pain sounded exaggerated, I was suspicious.”

many people have introspective access into the source of these judgments. Others described relying on the physical evidence without drawing an explicit comparison to the pain report.⁸ Some participants also named illicit drug or alcohol use as suspicious ($n = 17$), sometimes flagging such pain as self-inflicted.⁹ Interestingly, participant intuitions diverged regarding intimate complaints (i.e., genital pain and/or sex-related pain): A few participants ($n = 6$) reported viewing these complaints as more credible, while others ($n = 3$) found such complaints especially suspicious.¹⁰ These free-responses demonstrate the rich and sometimes contradictory motive intuitions that our naturalistic stimuli elicit.

Motive explains away pain. The models make specific, quantitative predictions about the strength of the explaining away effect. We compared these predictions to the empirical effect in humans in two ways (see Table 1). First, within Experiment 3, we partialled out the variance in prescription attributable to the LLM-rated severity of the pain report—a confound which inflates both motive ratings and prescriptions—and correlated the residual prescription with motive. Human judgments showed strong discounting ($r(\text{motive, prescription} \mid \text{report}) = -0.66$; $CI_{95\%} = (-0.68, -0.63)$). The Integrated model predicts a closely matching correlation ($r = -0.66$; $CI_{95\%} = (-0.68, -0.63)$); while the Experientialist model predicts a weaker effect size ($r = -0.32$; $CI_{95\%} = (-0.33, -0.30)$). The Direct-LLM predicts the weakest effect size (-0.19 ; Table 1), and the Physicalist model cannot predict discounting at all. Thus, the Integrated model captures this effect better than the alternative models.

⁸ “If there was a clear physiological cause for the pain, such as a tumor, intestinal obstruction, or nerve impingement, I concluded there was no ulterior motive.”

⁹ “If they did it themselves such as partied too hard and get headaches and pain all over their body, that is ulterior motive if they come in for medicine.”

¹⁰ Examples of participants’ diverging intuitions about the credibility of more intimate pain reports in free-response: “When people have ulterior motives, they don’t usually come up with weird stuff like burning in the vulva...” vs. “Some, like burning during sex, sound fabricated and hard to prove that they actually exist...”

Table 1

Each model predicts a different strength of the explaining-away effect. Within-experiment analysis: r between motive and prescription after partialling out the LLM-rated severity of the pain report. Between-experiment analysis: r between motive and the Experiment 2→3 change in prescription (ΔRx). The Integrated model predicts the strongest discounting, the Experientialist model a weaker one, the Physicalist model none (its motive posterior collapses to its prior, so its correlations are undefined), and the Direct-LLM baseline the weakest. The empirical human effect (bold) matches the Integrated model’s prediction (bold) for both analyses.

	Integrated	Experientialist	Physicalist	Direct-LLM baseline	Human (observed)
$r(\text{motive, prescription} \mid \text{report})$	-0.62 [-0.64, -0.59]	-0.32 [-0.33, -0.30]	—	-0.19 [-0.24, -0.13]	-0.66 [-0.68, -0.63]
$r(\text{motive, } \Delta Rx_{\text{Exp 3-Exp 2}})$	-0.29 [-0.34, -0.23]	-0.19 [-0.24, -0.15]	—	0.02 [-0.02, 0.07]	-0.35 [-0.39, -0.31]

Second, between experiments, we compared prescriptions for the same vignettes under Experiment 3’s motive framing and Experiment 2’s no-motive framing. For each of the 1,693 vignettes surviving exclusions in both experiments, we computed $\Delta Rx = \overline{Rx}_{\text{Exp3}} - \overline{Rx}_{\text{Exp2}}$ and correlated it with the average motive rating. Human judgments showed discounting ($r(\text{motive, } \Delta Rx) = -0.35$; $CI_{95\%} = (-0.39, -0.31)$), and as before, the Integrated model predicted a similar effect size ($r = -0.29$; $CI_{95\%} = (-0.34, -0.23)$). The Experientialist model predicted a weaker effect size ($r = -0.19$; $CI_{95\%} = (-0.24, -0.15)$). The Direct-LLM made its judgments about motive and painkiller prescriptions independently, and its judgments about each showed no relationship ($r = 0.02$; $CI_{95\%} = (-0.02, 0.07)$). As before, the Physicalist model could not predict discounting. We preregistered quantifying the empirical explaining-away effect with the correlations $r(\text{motive, prescription} \mid \text{report})$ and $r(\text{motive, } \Delta Rx)$, but we did not preregister using the models’ predicted effect size to distinguish between them. Together, these analyses show that only the Integrated model predicts the overall strength of the human explaining-away effect.

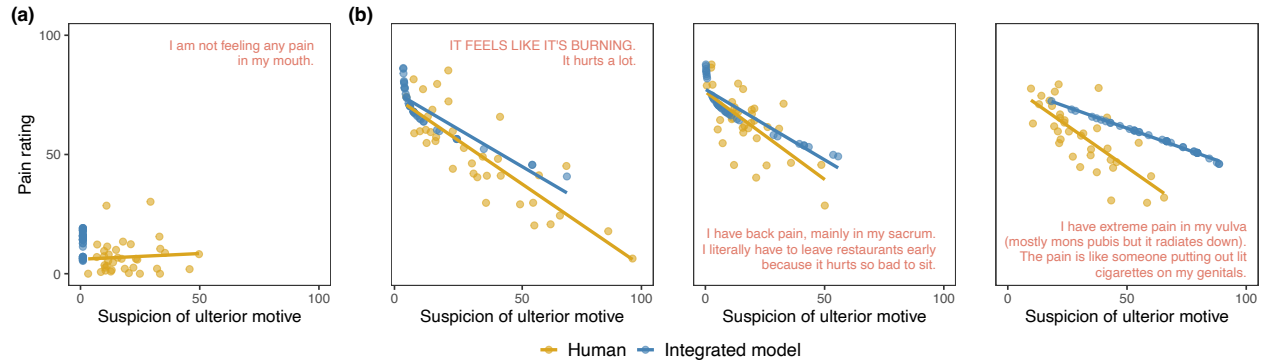


Figure 7

Motive explains away pain reports within report-matched vignette sets. Holding the pain report fixed isolates the effect of suspected motive on pain attribution. Each panel shows one report-matched set, in which the pain report is fixed while the event and physical evidence vary across the 36 vignettes in the set. Each point is a vignette; the Integrated model's predictions are in blue and average participant judgments in gold. A negative slope indicates the explaining-away effect. (a) The single report immune to explaining away, in which no pain is reported ("I am not feeling any pain in my mouth."). Because no pain is reported, the Integrated model infers no ulterior motive, resulting in the column of blue points; human judgments likewise show no discounting. (b) The explaining away effect on three example sets with non-zero pain reports, ordered left to right by increasing severity of the reported pain. All show the explaining-away effect for both humans and the Integrated model.

We next asked: Are any pain reports so compelling (e.g., through their vividness or other signals of honesty) that they cannot be explained away? To examine this question, we leveraged our stimulus design, in which each of the 48 unique pain reports was paired with 36 combinations of event and physical evidence, resulting in 48 sets of vignettes in which the event and physical evidence vary, but the pain report does not. Within each of the 48 sets of report-matched vignettes, we correlated human motive judgment and prescriptions. All but one set showed the explaining away effect (negative correlation; see Fig. 7). The only pain report immune to explaining away was one in which no pain was

reported: “I am not feeling any pain in my mouth.” The Integrated model largely reproduced this pattern (see SM Section S3.3.3). This indicates that reported pain—of any kind—can become suspect: the explaining-away effect generalizes across pain reports.

The fact that the Integrated model’s predicted explaining-away effect held across nearly every pain report indicates that the Integrated model’s account is not tied to the particulars of any one stimulus. More broadly, across Experiments 1–3 the model reproduced a series of hallmarks of causal reasoning: *forward prediction* of pain from physical causes (Experiment 1), *inference* of a latent experience from multiple sources of evidence (Experiment 2), *intervention* by prescribing painkillers (Experiments 2 and 3), and *inference of a hidden cause* together with *explaining away* (Experiment 3). For the most part, each form of causal reasoning was examined in isolation. A stronger test of the theory is whether one model with a fixed structure and parameters can account for all of these patterns, and whether it does so regardless of the kind of pain involved. We next examine the model’s generalizability across body parts and across experiments.

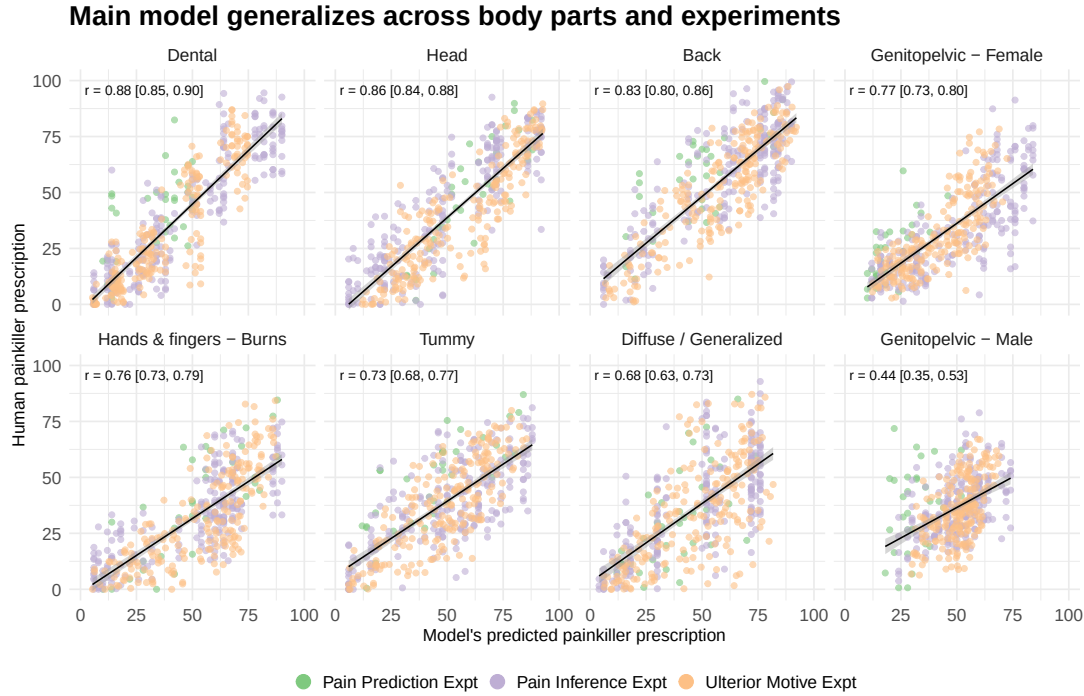
4.4 Cross-experiment generalization

The Integrated model generalizes across body parts.

The Integrated model might capture some types of pain (e.g., dental pain) better than others (e.g., headaches), especially if some of these types of pain are more commonly caused by physical events that leave evidence. We therefore examined whether the model generalizes across all eight body parts included in our stimuli.

Since all three experiments share prescribed painkiller dosage as a dependent variable, we used it as a common scale on which to compare the model’s predictions across body parts and across experiments. Fig. 8 shows the Integrated model’s prescription (pre-registered and fixed prior to data collection) vs. average human prescriptions.

The Integrated model predicted human prescriptions well for most body-part categories, but the exact fits varied: performance was strong for seven of the eight categories (average correlation $r = 0.79$) but weaker for Male Genital & Pelvic Pain

**Figure 8**

The Integrated model generalizes across body parts and experiments. Each panel shows one body-part category; humans' average painkiller prescription (y-axis) is plotted against the Integrated model's predicted prescription (x-axis), pooled across Experiments 1—3. Point color indicates the experiment. Within-panel correlations and 95% bootstrapped CIs are shown in the upper left. Panels are ordered by model–human correlation, from highest (top left) to lowest (bottom right).

($r = 0.44$; $CI_{95\%} = (0.35, 0.53)$). This variation was well explained by each category's spread (standard deviation) in pain-report severity, which predicted per-category model–human fit ($r = 0.94$; $CI_{95\%} = (0.69, 0.99)$; see SM Section S1.2 for per-category statistics). The larger the spread, the more signal available for the model to capture, and the better the fit between model and human judgments. Correspondingly, Male Genital & Pelvic Pain's pain-report range was unusually narrow.¹¹ The Integrated model's ability to

¹¹ Our vignettes were curated from Reddit posts and clinical-simulation materials, and the resulting feature distributions may reflect curation procedures. Whether the statistics of the stimuli reflect the

generalize across body parts suggests that the same intuitive causal theory of pain can capture reasoning about pain anywhere in the body.

Finally, the model also generalized across experiments. The fact that the same model structure was able to capture human judgments across three tasks with different framings, stimulus structures (i.e., vignette point of view), and participant samples indicates that this is a general account of how people reason about pain in a variety of contexts.

Generalization across body parts and experiments is one source of evidence for the causal account; another comes from how people flexibly reweight and recombine different sources of evidence across the experimental tasks. We now turn to this second source of evidence.

The Integrated model predicts the coefficients of the best-fitting regression.

Across experiments, the Integrated model takes three LLM-rated features as input: the severity of the inciting event, the strength of physical evidence for a problem in the body, and (where available) the severity of the patient’s pain report. For each experiment, a simple linear regression based on these three features achieves competitive accuracy on the key dependent variable in human judgments (Experiment 1 pain rating: $r = 0.72$; $CI_{95\%} = (0.66, 0.78)$; Experiment 2 prescription: $r = 0.79$; $CI_{95\%} = (0.77, 0.80)$; Experiment 3 motive: $r = 0.56$; $CI_{95\%} = (0.54, 0.59)$). One might therefore ask: if a regression does well, what does the structured causal model add?

Our answer is that the causal model provides a deeper level of explanation than the regression: it explains why the regression coefficients take the values they do. The causal model describes the relationship between variables in terms of a generative process—events cause damage; damage produces both visible evidence and pain experience; pain experience produces a report; ulterior motives produce an inflated report—and that generative structure determines both the direction and the relative magnitude of every feature’s

real-world distribution of features for these body-part categories is an open question left to future work.

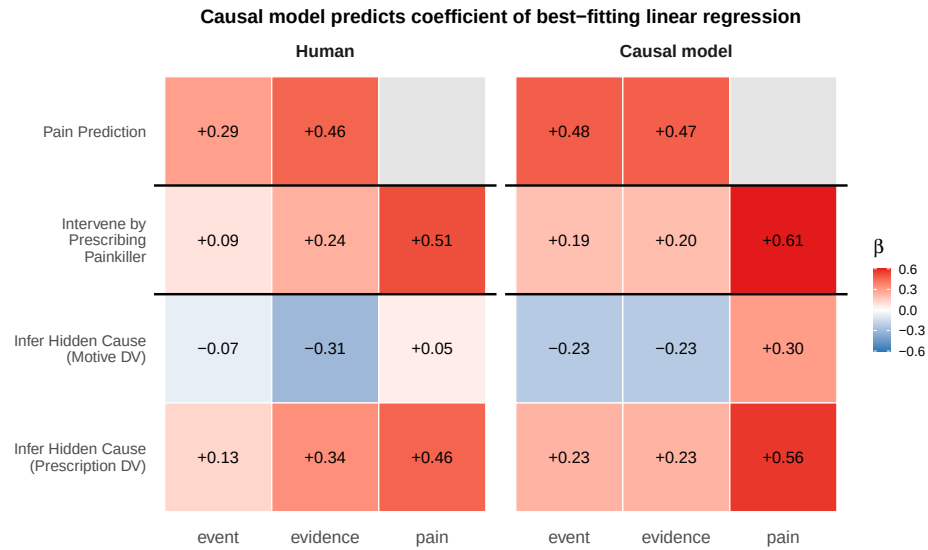


Figure 9

Regression coefficients from humans (left) and from the Integrated model’s predictions (right), for the primary dependent variable of each experiment (rows). Crucially, the Integrated model was not fit to data, while the regressions were. Each cell shows the raw regression coefficient and is colored by direction (red—positive; blue—negative). Across all 11 primary cells, the model predicts the sign of the human coefficient and largely tracks the rank order. The model fails to predict people’s relative weighting of physical evidence over the inciting event across all experiments, a discrepancy we address in the Discussion.

contribution to every inference. The causal model derives those weights from the generative graph. Fitting the same linear regression to the Integrated model’s *predictions* rather than to human responses, the regression coefficients largely match the coefficients fit to human judgments (Fig. 9), both in terms of their signs and their relative magnitudes. The causal model thus predicts the structure of the best-fitting regressions and provides a deeper account of why the regression coefficients look the way they do.

Finally, we consider how the models’ commitments affect the persuasiveness of their fits (Roberts & Pashler, 2000). The regressions can be interpreted as instantiating a cue-based account, in which people learn correlations among events, physical evidence,

pain, and pain reports without representing the causal process that relates them. Because this account can reproduce almost any pattern of results, it is difficult to falsify. The causal account, by contrast, makes far more restrictive commitments, as illustrated by the role of physical evidence. When observers judge how much pain a person is in, a more severe report and stronger physical evidence both indicate more pain (Experiments 1 and 2). But when observers judge ulterior motives, the two diverge: a severe report raises suspicion, while strong physical evidence corroborates the report and thereby reduces suspicion (Experiment 3). The Integrated model predicts this reversal in the role of physical evidence from a single causal structure and a single set of fixed parameters; it has no freedom to make the opposite prediction. Had human judgments not shown this pattern, the model would have been falsified. The cue-based account, by contrast, can only capture the reversal after the fact, by assigning physical evidence a positive weight for pain and a negative weight for motive. Because the causal model commits to predictions that could have turned out wrong, whereas the regressions commit to nothing, equal fit provides stronger evidence for the causal model.

5 Discussion

Building on a rich literature on conceptions of pain, this work takes a step toward formally characterizing folk intuitive theories of pain. We formalized intuitive theories of pain as Bayesian inference over causal models, and instantiated three competing models (Physicalist, Experientialist, and Integrated) representing different positions along the body-mind dimension of theories of pain. Formalizing these theories as models enabled us to test their quantitative predictions against human judgments across three experiments probing for signatures of causal reasoning: forward prediction, inverse inference, inferring hidden causes, and explaining away. Experiment 1 showed that people make quantitative predictions about the pain that will be experienced as a result of physical events. Experiment 2 showed that people invert this same causal structure to infer pain from subjective pain reports and physical evidence, rather than treat either source as fully

authoritative on its own. Experiment 3 then asked about the consequences of this simple intuitive theory, serving as a strong test of the theory’s generative reach by asking how far it could explain behavior in a qualitatively different task: inferring which pain reports are likely exaggerated or faked. We found that a substantial portion of people’s suspicions fall out as a direct consequence of their intuitive theory of pain. As predicted by the same model, suspicion of ulterior motives leads human participants to explain away pain reports. Taken together, these models and experiments show that people reason about pain through an intuitive theory that spans the physical and mental worlds, and they use this causal theory to make quantitative predictions about even the most subjective of human experiences.

We chose to study intuitive theories of pain in lay people rather than in clinicians for four reasons. First, pain is an important part of our everyday, commonsense understanding of the world, our bodies, and our place within it, but, unlike intuitive physics or intuitive psychology, it has rarely been formally characterized. Second, lay people’s attributions of pain have direct and widespread societal consequences: from parents deciding whether their kids can stay home from school to juries determining the monetary worth of a plaintiff’s pain and suffering, these decisions matter. Third, lay theories are the foundation on which clinical expertise is built. Clinicians are trained atop the folk concepts they already hold, and it is an open empirical question how medical education shifts their causal theories of pain—toward greater deference to self-report, toward greater reliance on physical findings, or along dimensions yet to be described. Understanding folk theories gives us the baseline from which to characterize how theories change with expertise. Last, characterization of lay theories of pain could inform medical education by identifying which aspects of modern pain science and management are most counterintuitive and require targeted training.

5.1 Design

A strength of this work is the use of semi-naturalistic stimuli created through factorial design, combining the richness of real first-person pain narratives with the benefits

of experimental control (Auspurg & Hinz, 2015). In fully naturalistic datasets (such as Reddit posts or medical training materials), the distribution of narratives is shaped by the reasons people produce such narratives in the first place (e.g., seeking help for medically unexplained pain, or illustrating instructive clinical cases), introducing confounding correlations between the events, physical findings, and pain reports. Our design circumvents this: by independently permuting events, physical findings, and reports—factors that are normally confounded in the wild—we can attribute variance in judgments to each source. This permutation occasionally yields “unnatural” pairings (e.g., a severe car crash with no reported pain, or pain with no identifiable physical cause), which we treat as informative edge cases; it is precisely these unintuitive cases that discriminate between competing models, pull apart our cognitive theories, and probe the limits of people’s intuitive understanding of pain.

The dependent variable common across all three experiments is painkiller prescription. Prescription is a less direct measure of pain attribution than direct pain ratings. Despite our careful instructions to “prescribe as accurately as possible” and to avoid under- or over-treating, measuring pain attribution through prescription risks inadvertently introducing considerations like addiction and asymmetric costs of over- versus under-treatment. However, our validation of prescription against explicit pain ratings in Experiment 1 (finding a near-perfect correlation of $r = 0.98$ between the two measures) should partly alleviate this concern. There are also benefits to using painkiller prescription as the dependent variable: it is an ecological measure that is consequential in the real world—the decisions that actually matter in clinical and everyday contexts—and we suspected that it might be more sensitive than explicit pain ratings for detecting subtle dismissals of pain as not quite “real” enough to treat with an intervention on the body. On the whole, we believe the benefits outweigh the weaknesses as a measure of pain attribution.

Our work addresses how people combine evidence to form intuitive attributions of pain. One might wonder whether this combination of evidence produces accurate

inferences—whether they approximate the experience of the person in pain. If there were a “mental dipstick” for measuring pain (Prkachin et al., 2007), this question might be answerable. But for now, we can only assess accuracy in two more limited senses. The first is the agreement between an observer’s inference and the sufferer’s own first-person report—the criterion used in studies of empathic accuracy (Ickes, 1993), including pain attribution (Prkachin et al., 2007); by this measure, observers’ inferences tracked pain reports closely ($r = 0.72$ in Experiment 2). The second is agreement among observers (Kenny, 1991), which we quantify with intraclass correlations (reported for each experiment above). Neither directly measures correspondence with the ground-truth experience itself, but together they bound how accurately these inferences can be characterized.

5.2 Interpreting model success

The Integrated model’s successes over alternatives only matter insofar as those alternatives are meaningful. The Physicalist and Experientalist models are not arbitrary: they instantiate the two poles of a long-standing debate about the nature of pain—the biomedical view that pain tracks tissue damage, and the experientalist view that pain is constituted by, and known through, the sufferer’s report. The experientalist pole is the prevailing normative stance in modern pain medicine, where self-report is the gold standard (Raja et al., 2020). Testing whether people actually reason this way is therefore a test of the establishment view, not a caricature. We grant that few hold either pole in its pure form—a committed physicalist might still admit self-report when physical findings are incomplete, and even an experientalist would defer to physical evidence when no report is available (as concerning a nonverbal patient, a non-human animal, or in Experiment 1). But in Experiments 2 and 3, where a pain report was always present, it was entirely plausible that people would take it at face value. They did not. People’s judgments of the same pain report varied based on the inciting event and physical evidence, resulting in the vertical columns of points visible in Figure 5—variance the Experientalist model could not explain.

We also compared our model directly against Llama 3.1, the Direct-LLM baseline. We chose Llama 3.1 for several reasons. As a non-reasoning language model, it serves as a stand-in for the general associations that can be learned from language, without layering an additional stage of explicit reasoning—planning, code synthesis, or extended deliberation—on top of those associations, isolating the contributions of such associations on their own. It is openly available, supporting reproducibility, and commonly used in cognitive science (e.g., Wong et al., 2025). Unlike the other alternative models, the direct LLM baseline received the full vignette text. Using the same LLM that parsed the input to the Integrated and alternative models as a baseline enabled us to directly assess what the structured causal model gains beyond its inputs. The Integrated model consistently outperformed the Llama 3.1, but we do not take this as a general claim about LLMs’ capacity to capture human judgment: what an LLM appears able to do depends heavily on the specific model and on how it is probed (McCoy et al., 2024), and the capability landscape shifts quickly enough that blanket claims tend not to age well. We simply observe that, for these tasks and this particular model, structured causal reasoning captured human judgments better than direct prediction based on text.

The cross-experiment regression analysis revealed one systematic discrepancy: across every experiment and dependent variable, human participants weighted physical evidence more heavily than the inciting event, whereas the model’s default parameters treat the two symmetrically (Fig. 9). A natural interpretation is that people treat the inciting event as an indirect, uncertain cause of bodily damage that may or may not produce lasting harm, while treating physical evidence as a more direct indicator of the damage itself, consistent with evidence that people reason differently from causes to effects than from effects to causes (Fernbach et al., 2011). Alternatively, this discrepancy may be a property of our materials: the events we sampled may simply have carried less diagnostic signal than the physical findings, in which case more intense event descriptions, or more ambiguous evidence, might reverse the weighting. In either case, the Integrated model can

accommodate this discrepancy by increasing the σ_{event} parameter (see SM Section S3.4.1), as it did during parameter fitting in Experiment 3. But if the asymmetry instead reflects the structure of people’s causal theory of pain, a model that represents the event as an insufficient, probabilistic cause (Cheng, 1997), rather than as a magnitude anchor symmetric with the evidence, might better allow the weighting to emerge from causal structure itself.

Despite the model’s overall success, Experiment 3 leaves substantial variance unexplained. The debrief responses help characterize its sources. The strategies participants agreed on—evidence–report mismatch and drug-use cues—are largely captured by the Integrated model and the Direct-LLM baseline, respectively. The remaining variance appears to reflect differences among raters rather than properties of the vignettes. Participants split, for example, over whether intimate disclosures are especially credible or especially suspect, offering a fruitful starting point for future work on motive judgments in their own right.

5.3 Conclusion

This work proposed that people attribute pain the same way they attribute other hidden physical or mental states: through an intuitive theory. We characterize this theory in terms of its position on the body-mind dimension along which theories of pain tend to disagree, and we find that a model that integrates over both physical (e.g., physical information) and social information (e.g., verbal reports) best matches human inferences about another person’s underlying pain experience. We formalized the causal models that lay intuitive theories of pain might synthesize for our experimental tasks, leading to a new, quantitative tool for predicting and probing people’s judgments in a variety of contexts. This formalization of intuitive theories of pain represents not only a theoretical advance, but has practical implications for the study of human behavior in clinical contexts. This work builds toward a science of how we understand, suspect, and trust each other as we reason and communicate about subjective experience.

6 Acknowledgments

We thank Laura Schulz and Sam Cheyette for many helpful conversations and Kartik Chandra for support using Memo. Special thanks to research assistants Zadriana Smith and Xinyue (Jessie) Yu. Finally, thank you to our CogSci reviewers for their helpful comments!

References

- Anthropic. (2026). Claude [Large language model].
- Atran, S. (1990). *Cognitive foundations of natural history: Towards an anthropology of science*. Cambridge University Press.
- Auspurg, K., & Hinz, T. (2015). *Factorial survey experiments*. SAGE Publications.
- Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, 113(3), 329–349.
<https://doi.org/10.1016/j.cognition.2009.07.005>
- Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences of the United States of America*, 110(45), 18327–18332.
<https://doi.org/10.1073/pnas.1306572110>
- Borg, E., Fisher, S. A., Hansen, N., Harrison, R., Ravindran, D., Salomons, T. V., & Wilkinson, H. (2021). Pain priors, polyeidism, and predictive power: A preliminary investigation into individual differences in ordinary thought about pain. *Theoretical Medicine and Bioethics*, 42(3–4), 113–135.
<https://doi.org/10.1007/s11017-021-09552-1>
- Borg, E., Harrison, R., Stazicker, J., & Salomons, T. (2020). Is the folk concept of pain polyeidic? *Mind & Language*, 35(1), 29–47. <https://doi.org/10.1111/mila.12227>
- Boring, B. L., Ng, B. W., Nanavaty, N., & Mathur, V. A. (2022). Over-rating pain is overrated: A fundamental self-other bias in pain reporting behavior. *The Journal of Pain*, 23(10), 1779–1789. <https://doi.org/10.1016/j.jpain.2022.06.002>
- Chandra, K., Chen, T., Tenenbaum, J. B., & Ragan-Kelley, J. (2025). A domain-specific probabilistic programming language for reasoning about reasoning (or: A memo on memo). *arXiv preprint*.
- Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, 104(2), 367–405. <https://doi.org/10.1037/0033-295X.104.2.367>

- Corns, J. (Ed.). (2017). *The Routledge handbook of philosophy of pain*. Routledge, Taylor & Francis Group.
- Descartes, R. (1664). *L'homme et un traitté de la formation du foetus du mesme auteur* (C. Clerselier, L. de La Forge, & F. Schuyt, Eds.). Charles Angot.
- Engel, G. L. (1977). The need for a new medical model: A challenge for biomedicine. *Science*, 196(4286), 129–136. <https://doi.org/10.1126/science.847460>
- Fernbach, P. M., Darlow, A., & Sloman, S. A. (2011). Asymmetries in predictive and diagnostic reasoning. *Journal of Experimental Psychology: General*, 140(2), 168–185.
- Gatchel, R. J., Peng, Y. B., Peters, M. L., Fuchs, P. N., & Turk, D. C. (2007). The biopsychosocial approach to chronic pain: Scientific advances and future directions. *Psychological Bulletin*, 133(4), 581–624. <https://doi.org/10.1037/0033-2909.133.4.581>
- Gerstenberg, T., & Tenenbaum, J. B. (2017). Intuitive theories. In M. R. Waldmann (Ed.), *The oxford handbook of causal reasoning*. Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199399550.013.28>
- Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and bayes nets. *Psychological Review*, 111(1), 3–32. <https://doi.org/10.1037/0033-295X.111.1.3>
- Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind & Language*, 7(1-2), 145–171. <https://doi.org/10.1111/j.1468-0017.1992.tb00202.x>
- Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, 315(5812), 619. <https://doi.org/10.1126/science.1134475>
- Ickes, W. (1993). Empathic accuracy. *Journal of Personality*, 61(4), 587–610.

- Kenny, D. A. (1991). A general model of consensus and accuracy in interpersonal perception. *Psychological Review*, 98(2), 155–163.
- Kripke, S. A. (1980). *Naming and necessity*. Harvard University Press.
- Luce, R. D. (1959). On the possible psychophysical laws. *Psychological Review*, 66(2), 81–95.
- McCaffery, M. (1968). *Nursing practice theories related to cognition, bodily pain and man-environment interactions*. University of California.
- McCloskey, M. (1983). Intuitive physics. *Scientific American*, 248(4), 122–131.
<https://doi.org/10.1038/scientificamerican0483-122>
- McCoy, R. T., Yao, S., Friedman, D., Hardy, M. D., & Griffiths, T. L. (2024). Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41), e2322420121.
- Melzack, R. (1975). The McGill pain questionnaire: Major properties and scoring methods. *Pain*, 1(3), 277–299. [https://doi.org/10.1016/0304-3959\(75\)90044-5](https://doi.org/10.1016/0304-3959(75)90044-5)
- Melzack, R. (1999). From the gate to the neuromatrix. *Pain, Suppl 6*, S121–S126.
[https://doi.org/10.1016/S0304-3959\(99\)00145-1](https://doi.org/10.1016/S0304-3959(99)00145-1)
- Melzack, R., & Casey, K. L. (1968). Sensory, motivational, and central control determinants of pain: A new conceptual model. In D. R. Kenshalo (Ed.), *The skin senses* (pp. 423–443). Charles C. Thomas.
- Melzack, R., & Wall, P. D. (1965). Pain mechanisms: A new theory. *Science*, 150(3699), 971–979. <https://doi.org/10.1126/science.150.3699.971>
- Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4), 435–450.
<https://doi.org/10.2307/2183914>
- OpenAI. (2026a). Chatgpt [Large language model].
- OpenAI. (2026b). Codex [Large language model].

- Outa, J., & Liu, S. (2023). Adults hold two parallel causal frameworks for reasoning about people's minds, actions and bodies. *Proceedings of the Annual Meeting of the Cognitive Science Society*.
- Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.
- Prkachin, K. M., Solomon, P. E., & Ross, J. (2007). Underestimation of pain by health-care providers: Towards a model of the process of inferring pain in others. *Canadian Journal of Nursing Research*, 39(2), 88–106.
- Radkani, S., Tenenbaum, J. B., & Saxe, R. (2025). What people learn from punishment: A cognitive model. *Proceedings of the National Academy of Sciences*, 122(32), e2500730122. <https://doi.org/10.1073/pnas.2500730122>
- Raja, S. N., Carr, D. B., Cohen, M., Finnerup, N. B., Flor, H., Gibson, S., Keefe, F. J., Mogil, J. S., Ringkamp, M., Sluka, K. A., Song, X.-J., Stevens, B., Sullivan, M. D., Tutelman, P. R., Ushida, T., & Vader, K. (2020). The revised International Association for the Study of Pain definition of pain: Concepts, challenges, and compromises. *Pain*, 161(9), 1976–1982. <https://doi.org/10.1097/j.pain.0000000000001939>
- Raz, G., Cao, A., Saxe, R., & Frank, M. C. (2025). A stimulus-computable rational model of visual habituation in infants and adults. *eLife*. <https://doi.org/10.7554/eLife.102713>
- Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, 107(2), 358–367. <https://doi.org/10.1037/0033-295X.107.2.358>
- Salomons, T. V., Harrison, R., Hansen, N., Stazicker, J., Sorensen, A. G., Thomas, P., & Borg, E. (2022). Is pain “all in your mind”? examining the general public's views of pain. *Review of Philosophy and Psychology*, 13(3), 683–698. <https://doi.org/10.1007/s13164-021-00553-6>

- Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in “theory of mind”. *NeuroImage*, 19(4), 1835–1842.
[https://doi.org/10.1016/S1053-8119\(03\)00230-1](https://doi.org/10.1016/S1053-8119(03)00230-1)
- Saxe, R., & Powell, L. J. (2006). It’s the thought that counts: Specific brain regions for one component of theory of mind. *Psychological Science*, 17(8), 692–699.
<https://doi.org/10.1111/j.1467-9280.2006.01768.x>
- Schulz, L. E., Bonawitz, E. B., & Griffiths, T. L. (2007). Can being scared cause tummy aches? naive theories, ambiguous evidence, and preschoolers’ causal inferences. *Developmental Psychology*, 43(5), 1124–1139.
<https://doi.org/10.1037/0012-1649.43.5.1124>
- Sytsma, J. (2010). Dennett’s theory of the folk theory of consciousness. *Journal of Consciousness Studies*, 17(3-4), 107–130.
- Tait, R. C., Chibnall, J. T., & Kalauokalani, D. (2009). Provider judgments of patients in pain: Seeking symptom certainty. *Pain Medicine*, 10(1), 11–34.
<https://doi.org/10.1111/j.1526-4637.2008.00527.x>
- Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, 10(7), 309–318.
<https://doi.org/10.1016/j.tics.2006.05.009>
- Tenenbaum, J. B., Griffiths, T. L., & Niyogi, S. (2007). Intuitive theories as grammars for causal inference. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 301–322). Oxford University Press.
<https://doi.org/10.1093/acprof:oso/9780195176803.003.0020>
- van der Linden, R. R., Schermer, M. H. N., Bolt, T. C., & Binney, N. R. (2025). From symptom to disease: A fleckian comparative epistemology of the conceptualization of chronic pain. *Synthese*, 206(100). <https://doi.org/10.1007/s11229-025-05180-7>

- Waddell, G. (1987). 1987 Volvo award in clinical sciences: A new clinical model for the treatment of low-back pain. *Spine*, 12(7), 632–644.
<https://doi.org/10.1097/00007632-198709000-00002>
- Wall, P. D. (1979). On the relation of injury to pain. *Pain*, 6(3), 253–264.
[https://doi.org/10.1016/0304-3959\(79\)90047-2](https://doi.org/10.1016/0304-3959(79)90047-2)
- Weisman, K., Dweck, C. S., & Markman, E. M. (2017). Rethinking people’s conceptions of mental life. *Proceedings of the National Academy of Sciences*, 114(43), 11374–11379.
<https://doi.org/10.1073/pnas.1704347114>
- Wong, L., Collins, K. M., Ying, L., Zhang, C. E., Weller, A., Gerstenberg, T., O’Donnell, T., Lew, A., Andreas, J., Brooke-Wilson, T., et al. (2025). Modeling open-world cognition as on-demand synthesis of probabilistic models. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 47.