UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

Title
Toward a Formalization of Human Intuitive Theories of Bodily Pain

Permalink
https://escholarship.org/uc/item/22s0g3j3

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors
Berke, Marlene
Collins, Katherine M
Tenenbaum, Joshua B.
et al.

Publication Date
2026

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Toward a Formalization of Human Intuitive Theories of Bodily Pain

Marlene D. Berke (berkem@mit.edu)[1], Katherine M. Collins (katiemc@mit.edu)[2, 3, 4], Joshua B. Tenenbaum (jbt@mit.edu)[2] & Rebecca Saxe (saxe@mit.edu)[1,2]
[1]McGovern Institute for Brain Research, Massachusetts Institute of Technology
[2]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology
[3]Princeton AI Lab, Princeton University
[4]Cambridge Centre for Human-Inspired AI, Cambridge University

Abstract

Pain is a subjective experience that people often communicate through words. How do observers, from family members to medical professionals, interpret these reports? We propose that people use causal inference to infer someone else's subjective pain experience, treating the pain experience as an unobservable latent variable and the report as one of several sources of observations. We formalized this as Bayesian inference over a causal model of bodily pain and tested it against human judgments on naturalistic stimuli drawn from Reddit posts. Critically, the model is stimulus-computable, taking as input exactly what participants see. The model quantitatively predicted human judgments about pain intensity and appropriate painkiller dosage and generalized across different kinds of pain (e.g., from headaches to finger burns). Furthermore, the model captured individual differences correlated with people's explicit beliefs about how best to assess pain. This work takes a first step toward a scientific understanding of folk intuitive theories of pain.

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

Introduction

Imagine that it's a Monday morning, and your child doesn't want to go to school. They are saying that their head hurts and refusing to get out of bed. Is your child belligerent, or are they actually in pain, or both? And is it serious? Suppose you take their temperature, and it's normal. They look fine. What do you do? Now, suppose you notice that their pupils are constricted, and they seem sensitive to the light. Does this change your intuitions? What if it's the day of the field trip to the zoo that they've been looking forward to all weekend? Or what if they had bonked their head the night before?

When reasoning about cases like this, people use an *intuitive theory*—a model of how causes generate effects that supports inferring latent causes from observed effects and correlates (Cheng, 1997; Gerstenberg & Tenenbaum, 2017; Gopnik et al., 2004; Lagnado & Sloman, 2004; Pearl, 2000). A long tradition in cognitive science has explored and developed formal models of people's intuitive theories in areas like intuitive physics (e.g., Battaglia et al., 2013; McCloskey, 1983), intuitive psychology (e.g., Baker et al., 2009; Gopnik & Wellman, 1992), and intuitive biology (e.g., Atran, 1990; Tenenbaum et al., 2006). However, how people think pain works and how these intuitions vary across people remain underexplored.

It is particularly important to understand people's intuitive theories of pain because these theories affect when and how people choose to seek medical care, workplace accommodations, or even support from family members. Furthermore, intuitive theories affect how people respond to these requests. As in other domains (e.g., ownership, Zhang et al., 2024), people may differ from one another in their intuitive theories of pain, posing a challenge for communicating about pain and pain treatment. A mature scientific understanding of people's intuitive theories of pain could inform interventions in medical education and clinician-patient communication.

What then might lay intuitive theories of pain look like? Because folk theories often resemble early scientific theories (Gopnik & Wellman, 1992; Kuhn, 1962), we looked to historical and current scientific theories of pain to generate hypotheses about the causal structure of intuitive theories of pain. The International Association for the Study of Pain (IASP) defines pain as a fundamentally subjective experience that need not be tied to a stimulus or physical cause (Raja et al., 2020). This view privileges subjective reports over physical findings. By contrast, the Cartesian biomedical tradition considered pain as the result of tissue damage and emphasized physical findings (for review, see Gatchel et al., 2007). In our work, we consider three competing hypotheses about lay people's theories of pain: 1) that pain is a wholly subjective experience and as such, "pain is what the person says it is" (McCaffery, 1968); 2) that pain arises from physical problems in the body and therefore physical evidence is the best way to assess it; and 3) that pain is a subjective experience that must be inferred through multiple sources of evidence, integrating pain reports with physical findings into a coherent causal model.

Our work presents a first step toward formalizing an intuitive theory of bodily pain. In particular, we implement competing hypotheses as computational models, allowing us to test their predictions against human behavior. In our experiment, observers are provided with subjective pain reports and physical evidence and asked to assess and treat 'true' pain. We find that people behave most like the model that integrates both sources, suggesting that lay theories treat pain as arising from physical problems in the body, such that physical events and findings serve as evidence of 'real' pain. We find that our intuitive causal model generalizes well across a wide variety of types of pain (e.g., from headaches to finger burns). Finally, we provide evidence that intuitive theories of pain differ across people, and people have reportable insight into their own intuitive theories of pain.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society* ©2026 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

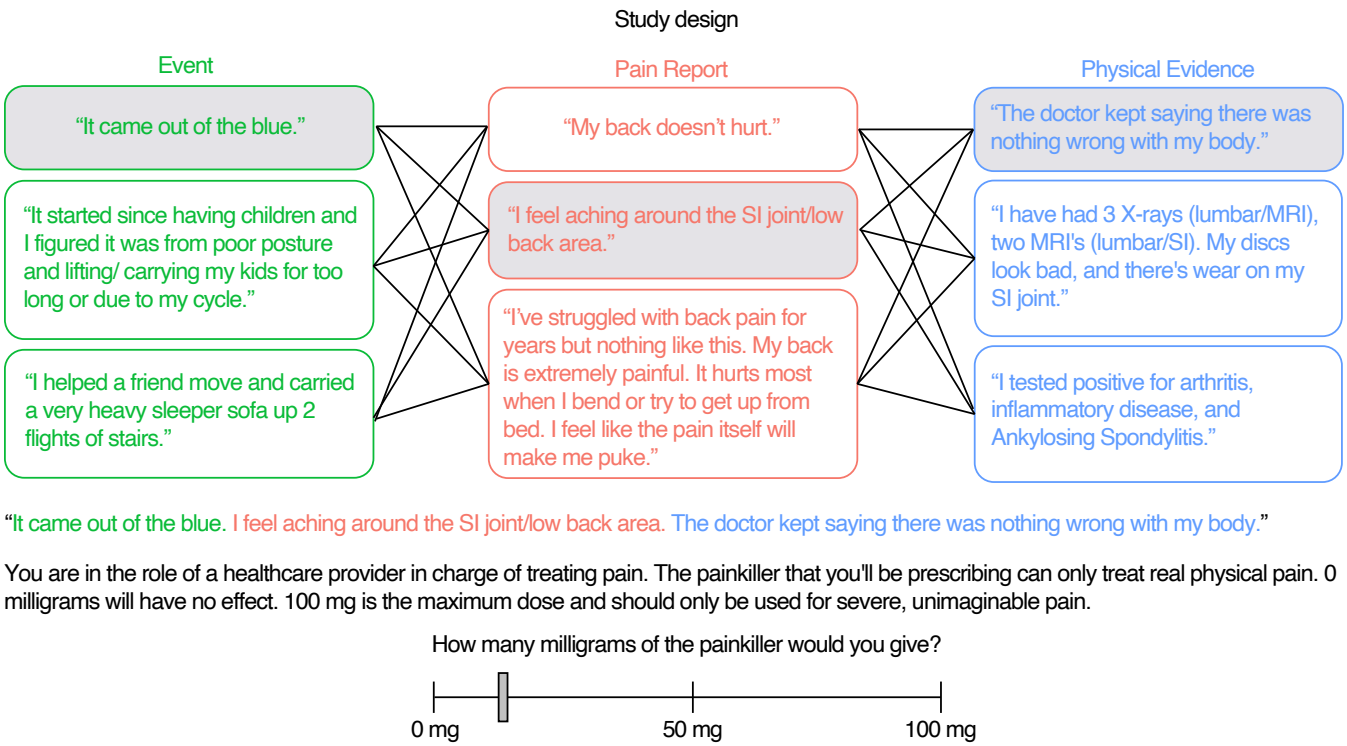

Figure 1: Our stimuli are generated by permuting descriptions of events, pain, and physical evidence. Participants read a vignette and use a slider to prescribe the dosage of a painkiller.

Study Design and Stimuli Curation

To study people's intuitions about pain, its causes, and evidence, we developed a novel behavioral task in which participants play the role of a healthcare provider prescribing doses of an imaginary perfect painkiller. Participants are instructed to prescribe the dosage as precisely as possible to treat the pain that the patient is experiencing. The prescribed dosage therefore serves as a proxy for the amount of pain that participants attribute to the patient in the story.

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 (or a statement about the lack of an event, e.g., "It just happened one day."), 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.").

To create a set of stories that varied independently along these dimensions, we permuted these event, pain, and evidence descriptions, creating a large set of vignettes systematically varying in the severity of the event, the pain, and the evidence described. Pain stories were drawn from eight categories (Back, Dental, Diffuse / Generalized Everywhere, Female Genital & Pelvic Pain, Male Genital & Pelvic Pain, Hands & Finger Burns, Head, Abdominal & GI). Combinations of event, damage, and pain reports were permuted within, but not across, categories. This process generated naturalistic vignettes spanning a variety of types and categories of pain and independently varying the variables of interest (severity of the inciting event, severity of reported pain, and severity of evidence). We measured the influence of each dimension on people's inferences about "real" pain.

Computational Framework

Next we implemented three competing hypotheses as computational models: 1) people consider pain to be a subjective mental experience independent from the physical body, and do not take physical events and physical evidence into account (report-only model; inspired by the 'it-is-what-they-say-it-is' view of pain); 2) people treat solely based on the event and evidence that they can see (physical-only model; inspired by the Cartesian biomedical model); or 3) people integrate the pain report, event, and physical findings into their inference about how much pain someone is 'really' in (main model);

We formalized competing accounts of pain attribution within the Bayesian causal reasoning framework (e.g., Pearl, 2000) and adopted the standard approach of inferring latent causal variables from observations. Typically, these types of models require numerical input, and mapping from vignette stimuli to model inputs requires gathering human ratings (or priors) about features of the vignettes (as in, e.g., Radkani et

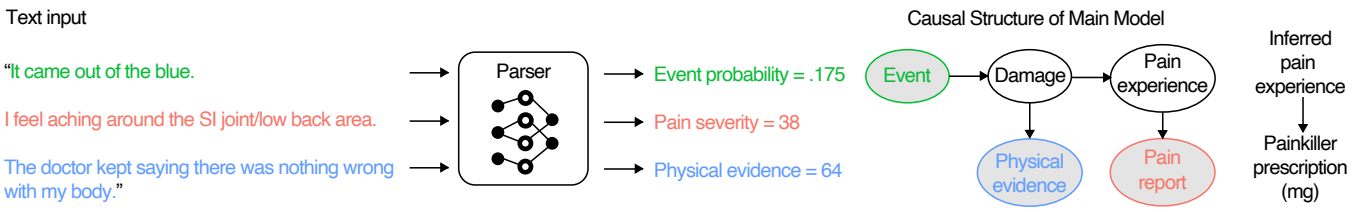

Figure 2: Schematic of our stimulus-computable model. Information flows from left to right. Gray shading denotes observable variables.

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 the described event, pain report, and physical evidence, and we used the resulting numeric ratings as inputs to our causal models. Thus, our models are "stimulus-computable" (Raz et al., 2025).

Main Model

Core causal structure Our main model describes a possible intuitive theory of pain as an experience resulting from a physical problem in the body. This model intuitively captures a physicalist view of pain: it is naive about the true complexities of how pain works, and instead aims to capture core elements of how people might assume or expect pain to work. The model makes four key commitments:

1) *Physical events that happen to the body produce problems in the body.* This principle captures the intuition that physical events often initiate the start of pain, and that more severe events tend to cause more severe problems (e.g., falling down the stairs is more likely to cause major damage than waking up in the morning). In our model, the severity of the problem in the body (e.g., damage) is sampled from a normal distribution with mean μ_{event} (the expected magnitude of damage caused by the described event) and variance σ^2_{event}. The parameter σ_{event} captures the intuition that even benign events can sometimes cause problems, and even severe events can sometimes leave the body unharmed.

2) *Physical problems in the body leave physical evidence.* This principle captures the intuition that evidence of bodily damage can be observed by others (e.g., through visual inspection, imaging, diagnostic tests, etc.). Furthermore, the more severe the problem, the more evidence there should be for it (e.g., a more severe stroke should leave more evidence in imaging and behavior, or a more severe fever should lead to a higher temperature reading on a thermometer). To capture the intuition that the amount of evidence is related to, but not fully determined by, the severity of the underlying problem, we modeled the severity of the evidence as sampled from a normal distribution with mean $\mu = severity_{problem}$ and variance $\sigma^2_{evidence}$. This parameter $\sigma_{evidence}$ controls how tightly or loosely physical evidence reflects the severity of underlying problems in the body.

3) *More severe or extensive physical problems in the body tend to produce a higher magnitude of experienced pain.* This principle aims to capture the intuition that the severity of physical pain that someone experiences tracks with the severity of the problem in their body (e.g., a more severe burn hurts more than a less severe burn). We modeled pain experience as a noisy signal reporting the extent of the physical problem in the body, such that the pain signal *pain* follows a normal distribution with mean $\mu = severity_{problem}$ and variance σ^2_{pain}. As before, σ_{pain} controls the strength of this relationship.

4) *People's verbal reports of pain track their experience of pain.* This principle aims to capture the intuition that people try to report the severity of their pain accurately, but that translating the experience into natural language introduces noise. We modeled the magnitude of reported pain as following a normal distribution with mean $\mu = severity_{pain}$ and variance σ^2_{report}. σ_{report} controls the accuracy with which people are expected to verbalize the severity of their pain.

So far, our model describes an intuitive causal theory of how events in the body produce the experience of pain and verbal reports of pain. By conditioning on the observable evidence (the event, the physical evidence, and the pain report), an observer can estimate the magnitude of the patient's pain experience using Bayesian inference (see Fig. 2). However, mapping from the amount of attributed pain to the painkiller prescription requires an additional model component.

Main Model Pseudocode

$\text{severity}_{\text{problem}}$	$\sim N(\mu = \mu_{event}, \sigma^2_{event})$
$\text{evidence}_{\text{physical}}$	$\sim N(\mu = \text{severity}_{\text{problem}}, \sigma^2_{evidence})$
$\text{experience}_{\text{pain}}$	$\sim N(\mu = \text{severity}_{\text{problem}}, \sigma^2_{pain})$
$\text{report}_{\text{pain}}$	$\sim N(\mu = \text{experience}_{\text{pain}}, \sigma^2_{report})$
$\text{dose}_{\text{painkiller}}$	$\sim \sigma(\tau, f(\text{experience}_{\text{pain}} \mid \text{dose}_{\text{painkiller}}\mid))$

Utility over treatment Moving from the observer's inference about the mental state of the patient to their decisions about pain treatment requires positing a utility function over pain treatment. For simplicity in our first version, we assumed that the cost of under- and over-prescribing painkillers is symmetric and scales linearly with the error, resulting in the cost function $f(dose_{painkiller} \mid experience_{pain}) = -1 * |\text{experience}_{\text{pain}} - \text{dose}_{\text{painkiller}}|$. Future extensions of the model might incorporate considerations like the risk of addiction into this utility function. Finally, we model people's treatment decisions as following a softmax (denoted by σ in Main Model Pseudocode) over the utilities, parametrized by the temperature parameter τ (Luce, 1959). Thus, the model

predicts a distribution over dosages of painkiller that the observer might prescribe.

Parameter settings Our model was 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 $\sigma_{event} = \sigma_{evidence} = \sigma_{pain} = \sigma_{report} = \frac{n}{10}$ and the temperature parameter to $\tau = 0.00001$.

Although these parameters affect the exact predictions that the model makes, 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.[1]

Model inputs and stimulus-computability To map natural-language narratives about pain into inputs to our model, we used a large language model. Specifically, we prompted Llama-3.1-8B-Instruct-Turbo (Grattafiori et al., 2024) to rate each component of the story on a scale from 0-100. The LLM was instructed to answer like "a participant in a human study without any specialized medical knowledge." For each component, the LLM was queried 10 times, with a temperature = 0.1; we took the average score as the coding.

For each component, the LLM was instructed to first provide an explanation and then a numeric rating. The instructions for rating pain were to "rate the severity of the described pain on a scale from 0-100, with 0 = no pain and 100 = the most excruciating, unbearable pain imaginable. Only consider real pain (not pretend)." This pain rating was input to the causal model as the observed pain report. To rate the event, the LLM was prompted to "rate on a scale from 0-100 the magnitude of physical injury, bodily damage, disease, or irritation that this event would cause, with 0 = no injury or damage and 100 = most severe, catastrophic injury or damage." This event rating was input into the causal model as μ_{event}, the expected magnitude of damage caused by the described event (see Main Model Pseudocode). Finally, the instructions for rating evidence were to "rate the strength of the physical evidence for physical injury, bodily damage, disease, or irritation on a scale from 0-100, with 0 = no evidence of any physical damage and 100 = strong, incontrovertible evidence of severe physical damage." This rating was input into the causal model as the observed evidence. Using the LLM to transform the stories into a format compatible with our causal model enabled the cognitive model to compute predictions directly from the stimuli.

We validated the LLM's ratings against two human coders (one an author and the other not), rating each the 48 events, 48 pain reports, and 48 descriptions of physical evidence. We found that the LLM ratings correlated as strongly with each coder as the coders did with each other ($r_{LLM\ and\ coder\ 1} = 0.79$, $CI_{95\%} = (0.73, 0.84)$; $r_{LLM\ and\ coder\ 2} = 0.84$, $CI_{95\%} = (0.79, 0.88)$; $r_{coder\ 1\ and\ 2} = 0.83$, $CI_{95\%} = (0.77, 0.88)$).

[1]We confirmed this intuition by fitting parameters to maximize Pearson correlation with average participant responses. Fitting the parameters of the main model increased the correlation only minimally (from $r = 0.78$ to $r = 0.79$), and fitting the parameters of the alternative models had negligible effects (report-only: unchanged at $r = 0.72$; physical: unchanged at $r = 0.30$).

Alternative Models

In addition to our main model, we implemented alternative models meant to capture competing theories of how people might reason about pain and how to treat it.

Report-only model People might view pain as an unobservable subjective experience for which the only relevant evidence is the person's pain report. This possibility is instantiated in the report-only model, which is a structured causal model like the main model, but ablates the event, damage, and physical evidence components, leaving only the pain experience and pain report.

Physical-only model Conversely, people might view pain as something physical in the body that cannot be accurately reported. In this case, they might consider the event and physical evidence as indicators of a pain experience, but discount the pain report. The physical-only model instantiates this hypothesis.

Direct-LLM The Direct-LLM model serves as a test of whether language associations are enough to produce human-like behavior on a painkiller prescription task without an explicit causal model. The LLM received instructions identical to those the human participants received and responded in the same way: by prescribing between 0-100mg of painkiller. As when parsing the story, we averaged over 10 rollouts of Meta-Llama-3.1-8B-Instruct-Turbo with temperature = 0.1.

Behavioral Experiment

All materials (including the stimuli, predictions, analyses, and sample size) were preregistered and can be found on our OSF repo: https://tinyurl.com/4bb7hj8e. An interactive visualization for browsing through our results is also available: https://marleneberke.shinyapps.io/exp1_vignette_browser/. We deviated from the preregistration by replacing the preregistered model (which treated events as Bernoulli random variables that increased 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 main model presented here both yield $r = 0.78$).

Participants

288 U.S.-based participants were recruited via Prolific and completed our experiment. As preregistered, we excluded

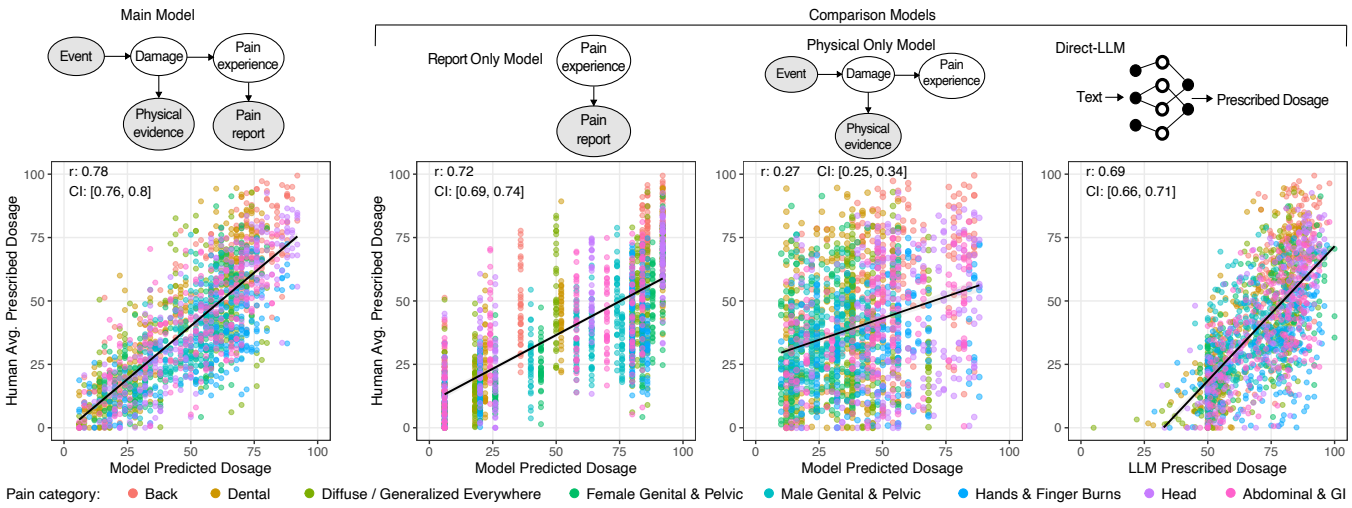

Figure 3: Scatterplots showing the relationship between average participant responses and each model's predictions. Participants' prescribed painkiller dosages are on the y-axis, and model predictions are on the x-axis. Each vignette results in one point, and they are color-coded by pain category. Pearson correlations and 95% bootstrapped confidence intervals are displayed.

participants who reported paying less than 90% attention, resulting in the exclusion of 12 participants.

Stimuli

Stimuli consisted of 1728 vignettes, generated from permuting six events, six pain reports, and six evidence descriptions for each of eight pain categories (e.g., back pain). These 1728 vignettes were split into 36 sets of 48 vignettes (six from each body category) so that no participant saw any vignette element repeated.

Procedure

Participants were introduced to the painkiller prescription task. They were told to play the role of a healthcare provider treating patients' pain. Their job was solely to manage the patients' current pain; diagnosing and treating injuries or ailments was the job of another doctor. To treat pain, they would prescribe a painkiller that only works on "real physical pain, not pretend or imaginary pain." Prescribing too little painkiller would result in the patient suffering from untreated pain, and prescribing too much would risk harming the patient's liver and causing more pain later down the line. Participants were told that 0mg would have no effect, and 100mg was the maximum dose that they could prescribe and should only be used for "severe, unimaginable pain." Participants used a slider from 0mg to 100mg to indicate the dosage that they would prescribe in each scenario.

Each participant was assigned to one of 36 sets of 48 vignettes, which they viewed in random order. We gathered responses from eight participants per vignette. For each vignette, they were provided with a checkbox to indicate if a scenario did not make sense. They were also asked how long the pain would last if untreated and how long the body would take to heal; these data are not analyzed here.

Following the main task, participants completed a survey. They rated their agreement with 10 different statements about pain and how to assess it on a seven-point Likert scale and answered open-ended questions about their firsthand experiences of pain. They also self-reported their attention using a slider from 0 to 100.

Results

Population-level analyses and results As preregistered, we excluded trials for which > 25% of participants indicated that the scenario did not make sense, resulting in the exclusion of 19/1728 trials.[2] To assess people's intuitive theories of pain at the population level, for each vignette, we averaged participants' responses and compared the results against the models' predictions. The main model predicted participants' prescribed painkiller dosages with high quantitative accuracy ($r = 0.78$; $CI_{95\%} = (0.76, 0.80)$; see Fig. 3), indicating that the causal model incorporating events, physical evidence, and pain reports into inferences about pain experience predicts people's treatment decisions well. This model generalized well across types of pain and body-part categories (see Fig. 4).

To assess alternative hypotheses about how people infer the severity of another person's pain, we compared the main model to alternative models. The report-only model considers the pain report as the only valid evidence of the pain experience; as pain is a subjective mental experience, this model disregards physical events and physical evidence. The report-only model predicted participants' painkiller prescriptions with high accuracy ($r = 0.72$; $CI_{95\%} = (0.70, 0.74)$; see Fig. 3). However, it failed to capture the variability in

[2] 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."

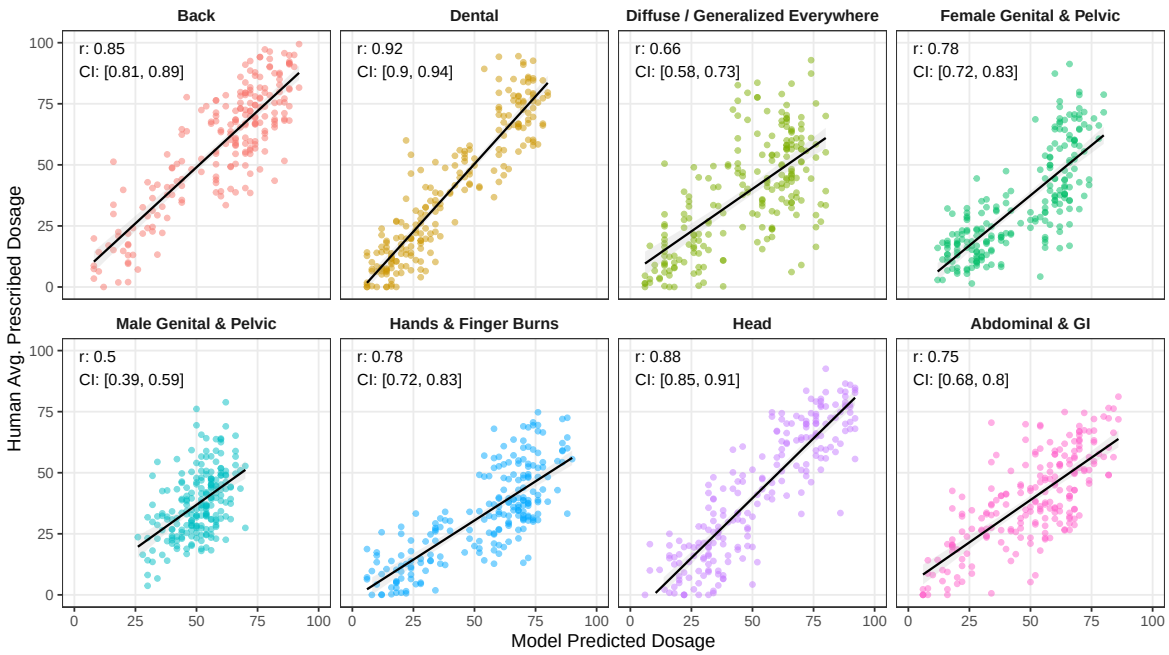

Figure 4: Scatterplots showing the relationship between average participant responses and each model's predictions, broken down by body parts/category of pain. The stimuli in the male genital & pelvic pain category had a smaller range of pain and evidence ratings, resulting in compressed model predictions and reduced performance.

people's prescriptions for vignettes containing the same pain reports, resulting in the columns of datapoints shown in Fig. 3. This report-only model performed reliably worse than the main 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, the physical-only model only considers physical events and evidence and ignoring the patient's pain report. If people do not trust subjective pain reports (perhaps because they expect that pain might be too difficult to introspect on and accurately report, or because they suspect that ulterior motives corrupt pain reports), then the physical-only model should better capture the patterns of people's prescribing behavior. The physical-only model predicted participant responses weakly ($r = 0.30$; $CI_{95\%} = (0.25, 0.34)$; see Fig. 3) and reliably worse than the main 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.

Additionally, we found that Llama 3.1 was able to mimic much of how participants' responses varied across vignettes ($r = 0.69$; $CI_{95\%} = (0.66, 0.71)$), but it consistently over-prescribed, often prescribing 50mg or more of painkiller in cases where participants prescribed far less (see Fig. 3). Furthermore, the main model reliably outperformed Llama 3.1 ($\delta_r = 0.09; CI_{95\%} = (0.07, 0.12)$), providing evidence that people's estimations of pain follow a structured reasoning process more similar to that of our structured causal model than the large language model. Future work could compare a wider range of language and reasoning models' judgments against people's.

Participant-level analyses, results, and individual differences People might not all share the same intuitive theory of how pain works and attribute it in the same way. Perhaps some people place more emphasis on physical events and evidence than others, and might be better described by the main model, while others might view physical events and evidence as less relevant to the mental experience of pain. To explore this possibility, we correlated each participant's responses with the two leading models. We found that the main model better captured more of the participants (219/276; $p_{binom} < .001$) and that individual participants' responses were, on average, better correlated with the main model than the report-only model ($r_{main} = 0.60$; $r_{report\ only} = 0.54$; $\delta_r = 0.054$; $CI_{95\%} = (0.046, 0.062)$, $t_{paired}(275) = 12.8; p < .001$; see Fig. 5A).

Still, some participants were better described by the report-only model. This difference in prescription behavior might map onto differences in explicit beliefs about pain assessment. To test this possibility, we used people's beliefs about pain, as measured on the 10-item questionnaire, to predict the difference in correlation between the main and report-only models. We found that participants' agreement with the statement "The best way to find out how much pain someone is in is to have a doctor assess them" predicted that their painkiller prescription behavior better matched the main model, which takes physical evidence into account ($r = 0.17$; $CI_{95\%} = (0.05, 0.29)$; Bonferroni-corrected $p = 0.048$; see Fig. 5B). This finding

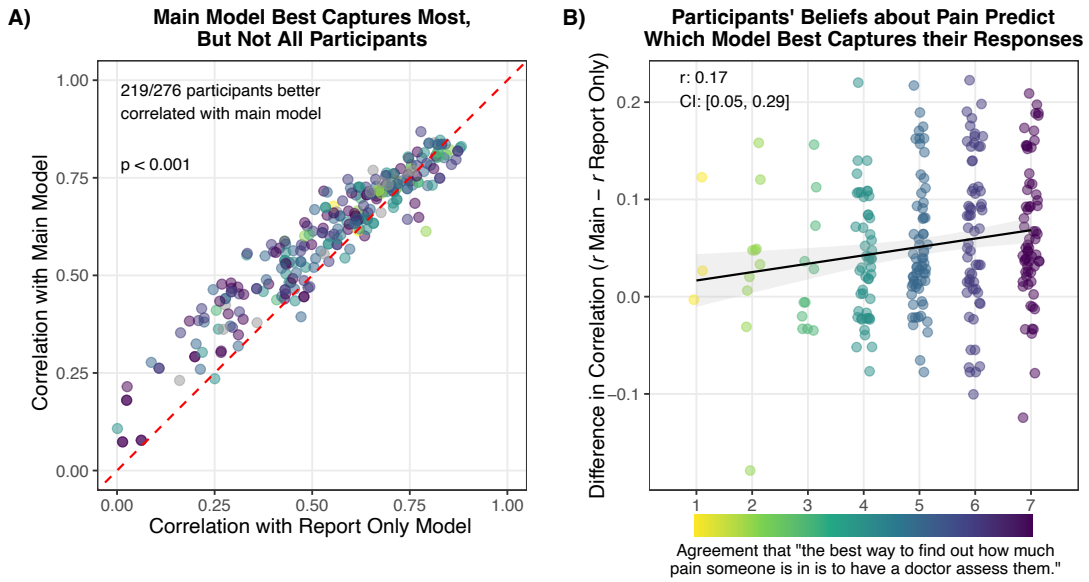

Figure 5: Individual differences. A) Participants' correlation with the main model (y-axis) and report-only model (x-axis). The dashed red line denotes x=y. For 219 out of 276 participants, the main model better described their responses (so 219/276 points are above the red dashed line). Point color indicates the participant's agreement with the doctor item on the pain beliefs questionnaire. B) Scatterplot showing the predictive ability of agreement with the doctor item in the pain beliefs questionnaire.

suggests that individuals have at least partial insight into their own causal model of pain, so individual differences in explicit responses correlate with individual differences in task performance patterns.

Discussion

This work presents a first step toward characterizing folk intuitive theories of pain. We formalized possible intuitive theories of bodily pain as simple causal models, allowing us to describe the problem of attributing pain to someone else as Bayesian inference about a latent pain experience. We tested our models' abilities to capture lay people's intuitions about pain magnitude and appropriate treatment, using a naturalistic stimulus set of real people describing pain in their own words. Across a wide variety of stories about pain, we found that the dominant intuitive theory held by a majority of participants incorporated physical events and physical evidence into pain estimates. At the same time, our task also revealed substantial individual differences in people's patterns of painkiller prescription. Individuals' self-reported beliefs about pain reflected these differences, perhaps pointing toward true heterogeneity in people's intuitive theories of pain.

A strength of this work is the use of semi-naturalistic stimuli, which combine the richness of real first-person pain narratives with experimental control. In fully naturalistic datasets (such as Reddit posts or medical training materials), the distribution of narratives over key dimensions—event severity, pain reports, and physical evidence—is likely 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 substantial confounds. By independently permuting events, reports, and evidence, we constructed vignettes in which these three dimensions vary orthogonally, enabling us to test stimuli that discriminate among competing models. This process frequently yielded "unnatural" pairings (e.g., a severe car crash with no reported pain, or pain with no identifiable cause), and it is precisely these unintuitive edge cases that pull apart our cognitive theories and probe the limits of people's intuitive understanding of pain.

In future work, we will use the core computational causal model presented here as a building block in larger models and theoretical expansions, including the temporal aspects of bodily healing and pain (e.g., Melzack et al., 1982) and the psychological factors that contribute to pain experience (e.g., Bolles & Fanselow, 1980). Furthermore, we will embed this core into communication models, allowing us to explore what happens when people with different intuitive theories of pain try to talk about it, when doctors try to explain to a patient the cause of their pain (Collins et al., 2025), or when we suspect someone may have ulterior motives for claiming to be in pain (e.g., not wanting to go to school). A natural next step is to build or even synthesize a variety of models to capture people's diverse and perhaps idiosyncratic theories of pain (for approaches to model synthesis, see Rmus et al., 2025; Wong et al., 2025). Additionally, a full account of pain attribution and communication will incorporate interpersonal and epistemic trust, relationships, and social categories (e.g., child vs. adult), building toward a full scientific theory of how people reason about, attribute, and communicate subjective experiences, especially across social and epistemic divides.

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

Atran, S. (1990). *Cognitive foundations of natural history: Towards an anthropology of science*. Cambridge University Press.

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

Bolles, R. C., & Fanselow, M. S. (1980). A perceptual-defensive-recuperative model of fear and pain. *Behavioral and Brain Sciences*, *3*(2), 291–301.

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). *psyarxiv preprint*. https://doi.org/10.31234/osf.io/pt863

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

Collins, K. M., Chandra, K., Ragan-Kelley, J., Weller, A., & Tenenbaum, J. B. (2025). Empathy in explanation. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*.

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*.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.

Lagnado, D. A., & Sloman, S. A. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(4), 856–876. https://doi.org/10.1037/0278-7393.30.4.856

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

Melzack, R., Wall, P. D., & Ty, T. C. (1982). Acute pain in an emergency clinic: Latency of onset and descriptor patterns related to different injuries. *Pain*, *14*(1), 33–43.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.

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

Rmus, M., Jagadish, A. K., Mathony, M., Ludwig, T., & Schulz, E. (2025). Generating computational cognitive models using large language models. *Advances in Neural Information Processing Systems*. https://arxiv.org/abs/2502.00879

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

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*.

Zhang, X., Bloom, P., & Jara-Ettinger, J. (2024). People have systematically different ownership intuitions in seemingly simple cases. *Psychological Science*, *35*(8), 858–871.