Supplementary Materials for "Toward a Formalization of Human Intuitive Theories of Bodily Pain"

Marlene D. Berke[1], Katherine M. Collins[2,3,4], Joshua B. Tenenbaum[2], and Rebecca Saxe[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

Supplementary Materials for "Toward a Formalization of Human Intuitive Theories of Bodily Pain"

Contents

S1 Materials

Our vignettes were built from naturalistic sources—Reddit posts and "standardized-patient" medical training materials—and then recombined by permuting events, physical evidence, and pain reports within each of eight body-part categories. Section S1.1 describes how the stimuli were created and traces each element to its source; Section S1.2 then provides summary statistics about the distributions of the resulting stimulus features across body-part categories; Section S1.3 reports the debrief questionnaire on beliefs about pain that participants completed at the end of each experiment.

S1.1 Source provenance for each stimulus element

Each body-part category was built from 18 unique text elements: six inciting events, six descriptions of physical evidence, and six pain reports (144 elements across the eight categories). These elements were recombined to construct the vignettes used in Experiments 1–3. Table S1 lists every unique element and its source. Sources include Reddit posts[1] and seven standardized-patient (SP) teaching cases drawn from four publicly available clinical-simulation documents (California Simulation Alliance, 2018a, 2018b; Gilbert Program in Medical Simulation, Harvard Medical School, 2012; University of Wisconsin School of Medicine and Public Health, Department of Family Medicine, 2014). We occasionally remixed categories: for example, in an attempt to better balance the male and female genital pain descriptions, we reassigned the "carpet burn" pain description—originally from an r/vulvodynia thread (*female* genital & pelvic pain)—to the Male Genital & Pelvic Pain category. To produce more positive examples of evidence, we sometimes reversed the polarity of the reported evidence (e.g. "no signs of inflammation" → "signs of inflammation"). Within the Hands & Finger Burns category, the source posts often referred to different fingers; because we recombine events, physical evidence, and pain reports within a category, we rewrote every element to describe the same site—the thumb and forefinger—so that the permuted vignettes remained consistent. More broadly, we

[1] These posts are most reliably found by searching https://arctic-shift.photon-reddit.com by post ID.

often lightly edited (condensed, lightly reworded, and/or corrected for spelling) the text.

Table S1

Source provenance for the 144 unique text elements (eight body-part categories × 18 elements: six events, six physical evidence descriptions, and six pain reports) that were recombined to form the vignettes. A code of the form `r/Sub/postID` *denotes a Reddit post; the full URL is* `reddit.com/r/Sub/comments/postID/`*. Other sources are cited as standardized-patient teaching cases or, in one instance, given as the URL of an article; "Authored" marks text the authors wrote.*

Category	Type	Text	Source
Dental	Event	I had 2 permanent crowns placed and a few cavities filled.	`r/askdentists/1d7ktx4`
		I bit down too hard on a popcorn kernel.	`r/askdentists/1guoeak`
		I had food and accidentally chewed on the side with the root canal.	`r/askdentists/1fzac57`
		I am not sure if it's because of a solution I had to swish in my mouth for a Sailogram which was super super sour that's caused sudden sensitivity. Or if it is because I stopped using sensodyne.	`r/askadentist/uy1fn0`
		I had a root canal. It was a pretty bad cavity.	`r/Dentistry/emhtur`
		I had a rough dental cleaning.	Authored
Dental	Physical evidence	I looked and saw that there's a hole in my gum now. I don't see bone or anything in the hole.	`r/askdentists/1e2d3yc`
		The dentist said I was showing signs of infection and inflammation.	Authored inversion of `r/askdentists/1guoeak`
		There is a big cavity in my tooth.	Authored
		My teeth look normal.	Authored

(continued on next page)

(Table S1, continued)

Category	Type	Text	Source
		After reviewing X-rays, the dentist said I had an unusually long pulp horn and he is now highly suspicious of pulpitis causing the issue.	r/askdentists/178w4vd
		The dentist took an X-ray, inspected the tooth thoroughly, shined a light through it. She found no visible issues - no cavity, no cracks, nothing. The tooth looks as healthy as the rest of them.	r/askdentists/15d1u4a
Dental	Pain description	I have excruciating pain in my mouth. It's the worst pain I've ever felt in my life.	r/askdentists/1d7ktx4
		I have this mild toothache. Its definitely not painful, just rather annoying as it is this constant dull ache. I don't have any other issues such as a tooth being sensitive to hot or cold, or hurting when I chew with it.	r/askdentists/1d45iwh
		The tooth now feels painful. It's so sensitive it hurts when I close my mouth and the opposite tooth touches it.	r/Dentistry/emhtur
		It feels sensitive when I brush my teeth in a specific spot or when I eat crunchy or hard foods.	r/Dentists/1eakji7
		I am not feeling any pain in my mouth.	Authored (no-pain)
		I am having intense pain in that tooth. It's throbbing with sharp pain on its own without external stimulation. It's also extremely sensitive to touch and hurts even more if the slightest thing bumps it. The pain is so severe that I can't sleep.	r/askdentists/1gd6ib1

(continued on next page)

(Table S1, continued)

Category	Type	Text	Source
Diffuse / Generalized Everywhere	Event	I thought it was due to a workout injury or a series of them. Maybe I overdid it and weakened myself.	`r/AutisticAdults/171lufi`
		I live in Canada and left the country for a vacation in Scotland and Iceland for 2.5 weeks. It started within ONE DAY of returning to Canada.	`r/Autoimmune/17nbove`
		It was sudden for me. It just happened one day.	Authored
		I partied hard, consuming large amounts of hard liquor and snorting cocaine.	Standardized-patient case: Cocaine Intoxication case (Gilbert Program in Medical Simulation, Harvard Medical School, 2012)
		It started during my chemo treatments.	`r/breastcancer/11hxe1r`
		I came down with Covid.	Authored
Diffuse / Generalized Everywhere	Physical evidence	Nothing looks wrong except that my nails are flushed into an orange/brown colour.	`r/breastcancer/11hxe1r`
		They ultrasounded my ankles and MRI'd my hands and found excess joint fluid.	`r/ChronicIllness/1eniy0o`
		I had blood work, xrays, and an ultrasound, and in the end found inflammation and arthritis.	Authored inversion of `r/ChronicIllness/15fjns9`

(continued on next page)

(Table S1, continued)

Category	Type	Text	Source
		The doctors finally determined that there is literally nothing wrong with me.	r/AutisticAdults/171lufi
		I have had a thoracic and cervical mri as well as several chest X-rays. They found a herniated c3-c4 and bulging c5-c6.	r/AskDocs/djv3ir
		I have a fever up to 101.8 by mouth.	Standardized-patient case: pediatric sore-throat scenario (University of Wisconsin School of Medicine and Public Health, Department of Family Medicine, 2014)
Diffuse / Generalized Everywhere	Pain description	The pain is everywhere. Every joint, every muscle is stiff and achy.	r/Fibromyalgia/15ctkvv
		The pain is bearable, ignorable.	r/Fibromyalgia/1j626fw
		The pain is in all my joints. It's a constant dull ache, occasional numbness, TONS of cracking/popping, horrible morning stiffness, can't move as well in the cold, etc.	r/ChronicIllness/1eniy0o

(continued on next page)

(Table S1, continued)

Category	Type	Text	Source
		I have symmetrical joint pain in my hands, knees, elbows, and wrists. Never any stiffness, and the best way I can describe the pain is like overuse injuries but everywhere. I also have bizarre sensory/nerve issues where I can no longer do things like wear a cotton shirt or certain types of pants because I get pain from them. I get extremely aggravating scalp pain, nerve issues such as certain spots, on my body feeling "dead", and joint pain.	`r/ChronicIllness/15fjns9`
		I am feeling fine.	Authored (no-pain)
		Never in my life have I felt as if I'd been hit by a truck!! The fatigue is awful, all my joints and muscles aching. My hips hurt, my lower back hurts, my neck is stiff.	`r/Fibromyalgia/1gftuzz`
Abdominal & GI	Event	I'm pretty sure this all started from an injury directly to my stomach when I was sleeping - probably about 200 lbs of pressure directly on my unguarded stomach, plus a lot of lemon water and antibiotics around the same time.	`r/Gastritis/1b1yukr`
		It started when I got my period.	Authored
		I can kind of pinpoint that it started around the time I changed probiotics, and also when I was doing strenuous lifting in the garden.	`r/FODMAPS/1kh3nik`
		I had a colectomy operation — part of my colon removed.	Standardized-patient case: geriatric post-op colectomy case (California Simulation Alliance, 2018a)

(continued on next page)

(Table S1, continued)

Category	Type	Text	Source
		I had an abdominal hysterectomy - My uterus was removed via an incision in my abdomen.	Standardized-patient case: post-op abdominal hysterectomy case (California Simulation Alliance, 2018b)
		It started after I went out for breakfast and ate fried potatoes, an omelet, buttered toast, and lots of bacon.	Standardized-patient case: Acute Cholecystitis (Gilbert Program in Medical Simulation, Harvard Medical School, 2012)
Abdominal & GI	Physical evidence	I have rectal bleeding and loose bowel movements that are mainly red blood with a small amount of liquid stool.	SP case: Lower Gastrointestinal Bleeding (Gilbert Program in Medical Simulation, Harvard Medical School, 2012)
		I tested positive for SIBO (small intestinal bacterial overgrowth).	`r/IBSHelp/1hxm8vz`
		I can feel the contractions through my skin — if I put my hand there it's like a pulsing.	`r/TooAfraidToAsk/1dtctko` (comment)
		Everything looks fine.	Authored
		My colonoscopy revealed a large tumor causing a partial intestinal obstruction.	Standardized-patient case: geriatric post-op colectomy case (California Simulation Alliance, 2018a)

(continued on next page)

(Table S1, continued)

Category	Type	Text	Source
		My abdomen is slightly distended.	Authored inversion from Standardized-patient case: Acute Cholecystitis (Gilbert Program in Medical Simulation, Harvard Medical School, 2012)
Abdominal & GI	Pain description	I have dull and uncomfortable pain above my belly button in the middle of my abdomen.	`r/Gastritis/14qunw2`
		It was like someone twisted everything in my lower abdomen together like a rope, and now they are steadily pulling on it. Not a sharp, stabbing pain but unrelenting and unyielding, the kind that makes you a bit nauseous.	`r/TooAfraidToAsk/1dtctko` (comment)
		It feels like someone is squeezing my stomach hard and letting go and repeating.	`r/ibs/xuc93z`
		I feel pressure and pain in my lower abdomen. A dull pain/pressure accompanied by stabbing pains.	`r/WomensHealth/17r3a27`
		I feel a discomfort in my abdomen. It is a burning sensation above my belly button. I'd describe the 'pain' as a gnawing, burning, pulling feeling. It doesn't have me doubled over.	`r/Hernia/1khn4gz`
		My tummy doesn't hurt.	Authored (no-pain)
Head	Event	I hit my head.	Authored

(continued on next page)

(Table S1, continued)

Category	Type	Text	Source
		I was hit over the head and left temple by the edge of a kayak paddle. It was a very concentrated impact point.	r/Concussion/k9cz2v
		I had an incident where I fell backwards and hit my head resulting in a concussion.	r/Concussion/nbifwt
		A 45lb barbell fell on my head in the weight room.	r/Concussion/1arj7nr
		I drank a lot of alcohol and I don't remember what happened.	Standardized-patient case: Migraine case (Gilbert Program in Medical Simulation, Harvard Medical School, 2012)
		I can't think of a specific event that caused it.	Authored (no-event)
Head	Physical evidence	I did an MRI and they found nothing and I've also been tested for sleep apnea.	r/migraine/za0b0m
		The bump was literally a thin swollen line that went all the way across the top of my skull.	r/Concussion/k9cz2v
		The MRI results said "small foci with volume loss with gliosis in frontal gyrus." My doc basically told me in other words, brain scarring.	r/Concussion/nbifwt
		I had CT scans which showed a fractured skull, but I was told I had meningitis. I have a huge scar on the back of my head.	r/TBI/1ewolcs
		I've had MRIs on both my head and my neck but all seem normal.	r/Concussion/1ffvur1
		My head is bleeding.	Authored

(continued on next page)

(Table S1, continued)

Category	Type	Text	Source
Head	Pain description	It's very much a throbbing headache with pressure behind the eyes. It usually radiates around my temples and is very equal on both sides. At my highest opening pressures it's been unbearable pain where I'm thrashing.	`r/iih/14aprh6` (comment)
		I have head pain sort of radiating from my neck to the back of my head. It's super uncomfortable and feels like part of my skull muscles are cramping or getting into a spasm.	`r/ankylosingspondylitis/17rfmto`
		It's not painful, but my head feels really full and there's a lot of pressure and I feel faint.	`r/dysautonomia/1fgh1gs`
		When the base of skull/neck pain starts, it will eventually turn into a horrible migraine and basically feels like someone is stabbing an icepick from the base of my skull to the adjoining eye. To describe the type of pain more clearly, it can feel like a piercing, shocking, and/or throbbing pain.	`r/migraine/13sjusi`
		I'm doing ok today, just still a headache and tired.	`r/Concussion/1arj7nr`
		My head doesn't hurt.	Authored (no-pain)
Back	Event	It came out of the blue.	`r/backpain/g7dnyl`
		It started during my IB examinations where I would have to sit for 6 hours with little to no break.	`r/backpain/1ey5dtc`
		It started since having children and I figured it was from poor posture and lifting/carrying my kids for too long or due to my cycle.	`r/backpain/15kz046`

(continued on next page)

(Table S1, continued)

Category	Type	Text	Source
		I was run over and dragged over 100 feet under a hot rod classic car that I built.	`r/IAmA/yzifw` (comment)
		I was involved in a rollover accident. I rolled 4 times with no seatbelt.	`r/AskDocs/djv3ir`
		I helped a friend move and carried a very heavy sleeper sofa up 2 flights of stairs.	Standardized-patient case: Scenario D, carpenter with low-back pain (University of Wisconsin School of Medicine and Public Health, Department of Family Medicine, 2014)
Back	Physical evidence	I have had 3 X-rays (lumbar/MRI), two MRI's (lumbar/SI). My discs look bad, and there's wear on my SI joint.	Authored inversion of `r/backpain/g7dnyl`
		I tested positive for arthritis, inflammatory disease, and Ankylosing Spondylitis.	Authored inversion of `r/backpain/g7dnyl`
		There's a small curve in my lumbar that's causing nerve impingement.	`r/backpain/1gahcv0`
		The curve seems so slight to me with a little bit of unevenness in my hips.	
		I've managed to get a back x-ray and a CT scan done of the area.	`r/backpain/tt9p4k`
		Both showed nothing of concern, only 'mild scoliosis' on the x-ray.	
		The doctor sent me for xrays there's narrowing at one of the discs in my lower back.	`r/Autoimmune/17nbove`

(continued on next page)

(Table S1, continued)

Category	Type	Text	Source
		The doc told it was probably a herniated disc causing the problem. I got myself an MRI and it was official.	r/backpain/1kkc5l4
Back	Pain description	I have back pain, mainly in my sacrum. I literally have to leave restaurants early because it hurts so bad to sit. I only feel comfortable when laying flat.	r/backpain/1fyk0i8
		A particular spot on my lower spine/vertebrae that is painful to the touch and feels like a bruise. It's super painful to bring my knees to my chest, and child's pose is nearly impossible. It feels extremely tight, almost like there's a ball under it, but not a normal tightness from inflexibility if that makes sense. It hurts when I lift things or lean over.	r/backpain/15kz046
		When the pain is at its worst I get this feeling of pressure in my middle back, directly in the center of my back in fact, not really in either side but right there in the middle. It really feels like something might explode! Of course it doesn't, but it sure feels like it. crushing, clenching, gripping, and severe pain that feels like it's coming from much deeper. All I feel is the pain, and like I might go crazy from this pain.	r/backpain/tt9p4k
		I've struggled with back pain for years but nothing like this. My back is extremely painful. It hurts most when I bend or try to get up from bed. I feel like the pain itself will make me puke.	r/ankylosingspondylitis/1h38ygn

(continued on next page)

(Table S1, continued)

Category	Type	Text	Source
		My back doesn't hurt.	Authored
		I feel aching around the SI joint/low back area.	r/backpain/g7dnyl
Female Genital & Pelvic Pain	Event	It started when I switched to my low dose combo birth control pill.	r/vulvodynia/tf448h
		It started after I gave birth.	r/vaginismus/7931c9
		It started with cycling last summer. I believe it all started when I moved my saddle too high trying to adjust it perfectly.	r/PelvicFloor/7y8eeu
		It started when I finally got rid of a yeast infection.	r/vulvodynia/180pgtp
		Nothing in particular happened.	Authored
		It has always been like this.	Authored
Female Genital & Pelvic Pain	Physical evidence	My urine culture came back positive.	Authored inversion of r/PelvicFloor/1869d1g
		Just outside the entrance to my vagina on either side there's localised redness.	r/vulvodynia/180pgtp
		The doctor kept saying there was nothing wrong with my body.	r/vaginismus/1ksw7r5
		I visibly show signs of atrophy and low estrogen according to my doctor.	Authored inversion of r/vulvodynia/tf448h
		I have been cleared of infection and been to a urologist who and discharged me as he couldn't find anything.	r/PelvicFloor/1dt0k8o

(continued on next page)

(Table S1, continued)

Category	Type	Text	Source
		I got a handheld mirror to look at my vulva. It is bright red and inflamed. Tests from bloodwork show my testosterone levels are very low.	https://www.self.com/story/keena-batti-vulvodynia-diagnosis
Female Genital & Pelvic Pain	Pain description	My urethra feels so uncomfortable all the time, like I always need to pee.	r/PelvicFloor/1869d1g
		I have slight pain when trying to have intercourse.	r/vaginismus/x8nqgu (Edited from "a lot of" to "slight")
		I don't have strong pain. Just strange sensations like spasms and pinching feelings around my vagina, anus, and urethra.	r/PelvicFloor/1kcbpdi
		I have extreme pain in my vulva (mostly mons pubis but it radiates down). The pain is like someone putting out lit cigarettes on my genitals and causes me to breakdown in tears at work.	r/vulvodynia/152512t
		I try to push myself to just have sex but the farther he goes the more excruciating it gets and I just can't.	r/vaginismus/7931c9
		During my Pap smears, I start flinching with pain, followed by screaming and crying - it is so painful.	r/vaginismus/1ksw7r5
Male Genital & Pelvic Pain	Event	It started after my vasectomy.	r/medicine/15rzhq2
		It started about a mile into my run. I stopped for a second to run across the road or catch my breath that's when it started.	r/beginnerfitness/1j7o2o4

(continued on next page)

(Table S1, continued)

Category	Type	Text	Source
		It started after having sex.	Authored
		It came on seemingly out of nowhere.	Authored
		I got my kidney stone surgery. It started after catheter removal.	r/UrethralStrictureAid/1gzas4m
		I can't think of a specific event that caused it.	Authored (no-event)
Male Genital & Pelvic Pain	Physical evidence	Physical exam is normal except under incision scar there is a small "mass" akin to a granuloma or varicose vein. Right testicle feels slightly larger than left.	r/medicine/15rzhq2
		After that I did a ultrasound where they took a bunch of pictures. With my right testy they found epididymitis.	r/ChronicPain/1gkqr95
		The pelvis/scrotum MRI saw a very small inguinal hernia on the left side.	r/Hernia/1jmsi5p
		The medical provider said there is nothing to worry about and every thing is fine.	r/UrethralStrictureAid/1gzas4m
		Saw a doctor and a urologist and they both came to the conclusion that it's most likely a varicocele.	r/Fitness/slj9g
		I don't see any sign of injury.	Authored
Male Genital & Pelvic Pain	Pain description	When I go for a pee it seems to really cause some discomfort and then seems to calm down. I also have a really tight left groin that really hurts.	r/PelvicFloor/1dt0k8o

(continued on next page)

(Table S1, continued)

Category	Type	Text	Source
		I have burning pee hole, glans irritation on underwear, painful perineum, and uncomfortable pain in my erection.	r/PelvicFloor/d3ulgc
		It feels like someone is squeezing my testicle. It's a really unpleasant feeling.	r/Fitness/slj9g
		I have terrible pain in my balls.	r/ChronicPain/1gkqr95
		I experience awful burning after sex. It honestly feels like carpet burn.	r/vulvodynia/tf448h
		I have a dull achy pain in the perineum.	r/PelvicFloor/7y8eeu
Hands & Finger Burns	Event	I put my hand in lukewarm water.	Authored inversion of boiling water events.
		I spilled boiling water on my hand/fingers.	r/Advice/mrcysu
		I was doing an experiment with sodium hydroxide and aluminum to make some hydrogen gas. I did eventually get some NaOH on my hands when I was mixing it with the water. I washed my hands straightaway.	r/chemistry/16813uj
		I missed the bowl and cream of mushroom soup stuck to my thumb and forefinger and burned the hell out of me.	r/TalesFromYourServer/uaba97

(continued on next page)

(Table S1, continued)

Category	Type	Text	Source
		I began cooking boba in a large pot, which sits on top of an electric stove set to 3500 watts (240-260 degrees). When I lifted the pot to take out the boiling water and bubbles I pressed my right hand against the rim of the metal pot. The cloud of searing hot water vapor also hit my hand.	r/Advice/tf6o79
		I burnt my thumb and forefinger on a pan while cooking.	r/Cooking/lchg4p
Hands & Finger Burns	Physical evidence	My hand and fingers are turning white, swelling, the skin is separating and blisters are forming on my knuckles.	r/Advice/tf6o79
		You can't see any damage or anything visually wrong with my hand.	r/chemistry/16813uj
		My skin looks pretty damaged.	Authored inversion of r/chemistry/16813uj
		I have blisters on my thumb and forefinger just past the knuckle.	Authored
		My thumb and index finger just look a bit red.	r/Cooking/lchg4p
		My thumb and index finger look like a leathery wrinkle.	r/TalesFromYourServer/uaba97
Hands & Finger Burns	Pain description	Whenever I remove my ice pack, this very painful burning sensation returns to the affected area.	r/askscience/ujm9du
		It's like intense pins and needles in my hand, my fingers feel like they're on fire.	r/Advice/mrcysu
		I never felt any pain.	Authored
		It stings and feels as if it's being burnt all over again.	r/Advice/tf6o79
		It hurts a little bit, like a mild sunburn.	Authored

(continued on next page)

(Table S1, continued)

Category	Type	Text	Source
		IT FEELS LIKE IT'S BURNING. It hurts a lot.	r/Cooking/lchg4p

S1.2 Stimulus feature distributions by body-part category

Table S2 reports the mean and standard deviation of LLM ratings (see S2.1) of the severity of the events, physical evidence, and pain reports in our stimulus set by body-part category.[2]

Table S2

Mean (SD) of LLM-rated stimulus features by body-part category, from the Experiment 2 and 3 stimulus pool ($n = 1728$ vignettes, 216 per category).

Body part	Event M (SD)	Evidence M (SD)	Pain M (SD)
Back	48.9 (34.2)	53.8 (27.1)	64.4 (34.2)
Dental	16.2 (6.1)	47.3 (29.2)	48.2 (37.0)
Diffuse / Generalized Everywhere	43.5 (22.4)	35.9 (26.5)	52.2 (33.6)
Female Genital & Pelvic Pain	9.8 (10.7)	31.5 (18.8)	60.9 (29.3)
Male Genital & Pelvic Pain	33.8 (9.7)	28.0 (16.6)	67.6 (14.5)
Hands & Finger Burns	61.5 (33.0)	43.1 (30.7)	55.5 (33.1)
Head	68.8 (17.4)	55.8 (30.0)	51.8 (37.9)
Abdominal & GI	41.3 (31.6)	50.1 (29.5)	52.0 (29.5)

The uneven spread in pain reports drives the variation in per-category model–human fit reported in the main text: the more a category's pain reports vary, the more signal there is for any model to explain. The clearest case is Male Genital & Pelvic Pain, which has the narrowest pain report distributions (SD=14.5; see Table S2) and is the only category without a "no pain" report. Its weaker cross-experiment fit reported in the main text is therefore best read as a stimulus-range effect. Indeed, across the eight categories, the spread of pain-report ratings correlates strongly with the model's cross-experiment fit for that category ($r = 0.94$, $CI_{95\%} = (0.69, 0.99)$, $p < .001$, $n = 8$). The

[2] We make no claim that these statistics reflect the real-world distribution of these features in clinical populations: they reflect the distribution of stimuli we curated and permuted, and should be interpreted as a property of our stimulus set rather than as representative of real pain experiences.

Male Genital & Pelvic Pain category drives this effect; excluding it, the correlation among the remaining seven categories is weaker and unreliable ($r = 0.70$, $p = .08$).

S1.3 Questionnaire on beliefs about pain

At the end of each experiment, participants completed a debrief questionnaire that included ten statements about the nature and assessment of pain (Table S3), each rated on a 7-point agreement scale (1 = Strongly Disagree to 7 = Strongly Agree). The individual differences analysis reported in Section S3.2.2 uses participants' responses on these items to predict their prescribing behavior.

Table S3

The ten-item pain-beliefs questionnaire, administered at the end of each experiment. Participants rated their agreement with each statement.

1. The best way to find out how much pain someone is in is to ask them.
2. The best way to find out how much pain someone is in is to look at them (e.g., their face, their body, their skin).
3. The best way to find out how much pain someone is in is to have a doctor assess them.
4. People tend to exaggerate pain.
5. It is possible for two people to have the same exact injury and damage to the body but feel different amounts of pain.
6. An injury twice as severe (e.g., a burn twice as bad, a cut twice as wide or deep, two broken ribs rather than one) will always hurt more.
7. You can have physical damage to the body without experiencing pain.
8. You can have physical pain without physical damage to the body.
9. Once the body heals, pain will go away.
10. Two people with the same injury feel mostly the same thing.

S2 LLM pipeline, prompts, and validation

S2.1 LLM ratings of stimuli elements

We mapped each vignette into numeric inputs for the causal models using Llama-3.1-8B-Instruct-Turbo (Grattafiori et al., 2024), prompting the model to rate each component (event, physical evidence, pain report) 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 temperature = 0.1, and we averaged the resulting ratings.

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 models 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 models as μ_{event}, the expected magnitude of damage caused by the described event. 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 models as the observed evidence.

We validated the LLM's ratings against two human coders (one an author and the other not), rating each of 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 ($r_{LLM\ and\ coder\ 1} = 0.79$, $CI_{95\%} = (0.73, 0.84)$; $r_{LLM\ and\ coder\ 2} = 0.84$, $CI_{95\%} = (0.79, 0.88)$; $r_{coder\ 1\ and\ 2} = 0.83$, $CI_{95\%} = (0.77, 0.88)$).

S2.2 Direct-LLM baseline

As a benchmark for the structured causal models, we constructed a Direct-LLM baseline that rated pain (Experiments 1), prescribed painkiller (Experiment 2 and 3), and rated motive suspicion (Experiment 3). The baseline used the same Llama-3.1-8B-Instruct model as the LLM-rating pipeline (including averaging predictions over 10 rollouts at temperature = 0.1) described in Section S2.1. The baseline received the same task instructions as human participants, with the additional line: "You are a participant in a psychology study." For Experiments 1 and 2, the model was given the same instruction as participants: "You will be shown a brief description of a person who is experiencing pain. Your job is to rate how much pain you think they are experiencing, on a scale from 0 to 100, where 0 means no pain at all and 100 means the worst pain imaginable. Answer with only a number from 0 to 100." For Experiment 2, "pain" was replaced with "painkiller prescription in milligrams." For Experiment 3, the motive-rating prompt mirrored the instructions seen by human participants in that experiment: "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)?", rated on a 0–100 scale from "definitely no ulterior motive" (0), through "as likely as not to have an ulterior motive" (50), to "definitely has an ulterior motive" (100), preceded by the same framing given to participants (that some patients may have an ulterior motive for claiming pain and should not be prescribed painkillers). Outputs were parsed as integers; responses that could not be parsed were excluded (fewer than 0.5% of rollouts across all experiments).

All LLM ratings in this paper use Meta's Llama-3.1-8B-Instruct open weights (Grattafiori et al., 2024). Over the course of the project, the serving provider changed, as serverless endpoints were deprecated. The stimulus-feature ratings for Experiments 1 and 2 (and, by inheritance, the Experiment-3 model inputs) were generated using Together AI's "Turbo" build. For Experiment 3, the pre-registered Direct-LLM motive baseline was served by Cerebras through the Hugging Face inference router; after that endpoint became

unavailable, Experiment 3 pain and prescription ratings were regenerated with the same Llama-3.1-8B-Instruct weights served by Nscale.

S3 Additional analyses

This section reports analyses that supplement those in the main text, including pre-registered analyses not presented in main. These analyses are organized and presented by experiment.

S3.1 Experiment 1

S3.1.1 Model–human fit on painkiller prescription

In the main text, we presented model evaluations on pain ratings for Experiment 1 (Pain Prediction). We had preregistered these evaluations on painkiller prescriptions rather than pain ratings. In practice, because the pain ratings and prescribed painkiller dosages were nearly identical ($r = 0.98$), this choice did not matter, and we chose to present pain ratings for stylistic reasons. For completeness, we present the results using the painkiller prescription in Table S4. Prescription judgments showed similar reliability to pain ratings: split-half $r = 0.67$, $CI_{95\%} = (0.61, 0.72)$; ICC$(2, 1) = 0.40$, $CI_{95\%} = (0.35, 0.44)$.

Table S4

Experiment 1 model–human fit. Pearson r (with 95% CIs in brackets) between each model's predictions and humans' average response, across n = 269 vignettes, on both pain rating and prescription. Results are nearly identical regardless of the dependent variable.

Model	Pain rating	Prescription
Integrated	0.71 [0.64, 0.76]	0.70 [0.64, 0.76]
Physicalist	0.71 [0.64, 0.76]	0.70 [0.64, 0.76]
Experientialist	undefined (constant prediction)	
Direct-LLM	0.70 [0.64, 0.76]	0.70 [0.63, 0.75]

S3.1.2 Event-only and evidence-only lesions

We also pre-registered an analysis comparing the Integrated model against two lesions specific to Experiment 1: an *event-only* lesion (uses only the event rating to infer pain, ignoring physical evidence) and an *evidence-only* lesion (uses only physical evidence, ignoring the event). For parallelism with Experiments 2 and 3, the main text instead presents the model comparison using the set of lesion models common across experiments (Experientialist model, Physicalist model, Direct-LLM); we report the pre-registered event-only lesion and evidence-only lesion here for completeness. Both lesions performed reliably worse than the Integrated model (Table S5).

Table S5

Pre-registered Experiment 1 lesion analyses. Pearson r (with 95% CIs in brackets) between each model's predictions and humans' average response, computed across $n = 269$ vignettes for each dependent variable.

Model	Pain rating	Prescription
Integrated model	0.71 [0.64, 0.76]	0.70 [0.64, 0.76]
Event-only lesion	0.43 [0.33, 0.52]	0.43 [0.33, 0.52]
Evidence-only lesion	0.60 [0.52, 0.67]	0.60 [0.52, 0.67]

S3.2 Experiment 2

S3.2.1 Bernoulli vs. Gaussian event model

The model preregistered for Experiment 2 treated the inciting event as a Bernoulli variable occurring with probability p (based on the LLM event ratings) that increases the probability of a problem in the body (i.e., an event either "happens" or not, with no graded magnitude). Under this model, the extent of the damage is sampled from a uniform distribution when an event occurs, or from an exponential distribution (with mean=$n/10$) in the absence of an event (capturing the idea that, even without an external physical event to cause it, damage can be present, but it's unlikely to be severe).

The main text presents the Gaussian-event model variant, in which the LLM-rated event severity sets the mean of a Gaussian over the underlying damage variable. The two models make very similar predictions ($r = 0.954$ across the $n = 1709$ analyzed vignettes) and yield essentially the same fit to humans' average prescriptions (Bernoulli-event model: $r = 0.780$, $CI_{95\%} = (0.761, 0.798)$; Gaussian-event model: $r = 0.780$, $CI_{95\%} = (0.761, 0.798)$), so the conclusions of Experiment 2 are unchanged.

However, the models differ in how they distribute weight across the three input features. Regressing each model's predicted prescription on the LLM-rated event, evidence, and pain-report features (Table S6) reveals that the Bernoulli model assigns essentially no weight to the event predictor while loading more strongly on physical evidence, whereas the Gaussian model weighs event and evidence nearly the same. We adopted the Gaussian model in the main text because it has a more parsimonious causal structure (event is a continuous magnitude rather than a binary occurrence) and because event severity is then directly comparable to physical evidence severity on the same scale. As we discuss in the main text, returning to a *Bernoulli*-type event model—one that treats the inciting event as a probabilistic, insufficient cause of damage rather than a magnitude on the same scale as the physical evidence—would give a structural reason for people's higher relative weighting of physical evidence over events, rather than relying on parameter fitting to capture this weighting.

Table S6

Experiment 2 regression coefficients (raw β, with 95% CIs in brackets) from $\hat{y} \sim event + evidence + pain\ report$, fit separately to humans, the preregistered Bernoulli-event model, and the Gaussian-event model used in the main text. $n = 1709$ vignettes.

Response	β_{event}	$\beta_{evidence}$	β_{report}
Human Rx	0.09 [0.06, 0.11]	0.24 [0.22, 0.27]	0.51 [0.49, 0.53]
Bernoulli model Rx	0.02 [0.02, 0.02]	0.36 [0.35, 0.36]	0.67 [0.67, 0.68]
Gaussian (Integrated) model Rx	0.19 [0.19, 0.20]	0.20 [0.20, 0.21]	0.61 [0.60, 0.61]

S3.2.2 Individual differences in evidence weighting

People might not 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 Integrated model, while others might view physical events and evidence as less relevant to the mental experience of pain. To explore this possibility, as per our pre-registration, we correlated each participant's responses with the two leading models. We found that the Integrated model better captured more participants (219/276; $p_{binom} < .001$) and that individual participants' responses were, on average, better correlated with the Integrated model than the Experientialist 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. S1A).

We also compared all three models (not pre-registered): the Integrated model best described 189 participants, the Experientialist model 55, and the Physicalist model 32. The Experientialist group was nearly identical to the pairwise result: 55 of the 57 participants better fit by the Experientialist model than the Integrated model remained best described by the Experientialist model.

As a complementary (non-preregistered) analysis, we fit a linear regression to each

participant's prescriptions individually:

$$\text{prescription} \sim \beta_{\text{event}}\,\text{event} + \beta_{\text{evidence}}\,\text{evidence} + \beta_{\text{report}}\,\text{pain report}$$

This analysis asks a slightly different—and more permissive—question from the model comparison above: rather than requiring the Experientialist model to *beat* the Integrated model, it only requires that the event and evidence coefficients each fail to reach significance individually. Under this more liberal criterion, we found that 112/276 participants (40.6%) showed no reliable sensitivity to either event or evidence (both $p > .05$). This group overlapped substantially with the group best described by the Experientialist model: 44 of the 55 participants best described by the Experientialist model (80.0%) also had non-significant event and evidence coefficients in the regression. In contrast, nearly all participants positively weighed the pain report: 274/276 (99.3%) participants showed a positive pain-report coefficient, and 246/276 (89.1%) were significantly positive ($p < .05$), confirming that reliance on the pain report is universal. This provides converging evidence that all participants cared about the pain report, and a meaningful minority of participants cared about pain reports *exclusively*.

We suspected that participants' differences in weighting physical evidence when prescribing might map onto differences in explicit beliefs about pain assessment. In a pre-registered analysis, we used people's beliefs about pain, as measured on the 10-item questionnaire presented in Section S1.3 to predict the difference in correlation between the Integrated and Experientialist 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 Integrated model, which takes physical evidence into account ($r = 0.17$; $CI_{95\%} = (0.05, 0.29)$; Bonferroni-corrected $p = 0.048$; see Fig. S1B). This finding 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.

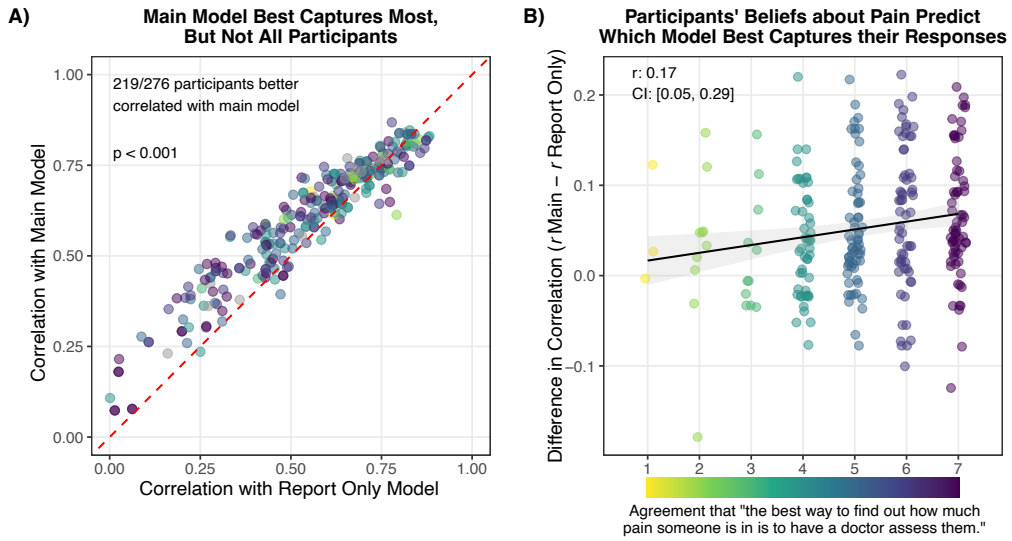

Figure S1

Individual differences. A) Participants' correlation with the Integrated model (y-axis) and Experientialist model (x-axis). The dashed red line denotes x=y. For 219 out of 276 participants, the Integrated 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) Each point displays a participant's agreement with the doctor item (x-axis) against how much better the Integrated model fits their prescriptions than the Experientialist model (y-axis).

S3.3 Experiment 3

S3.3.1 Individual differences in motive judgments

In Experiment 3, individual motive judgments were highly variable across both vignettes and participants. A variance-components analysis (random intercepts for vignette and participant) attributed 20.1% of the variance in motive ratings to vignette content ($CI_{95\%} = (18.7\%, 21.7\%)$), 20.7% to between-participant differences in rater style ($CI_{95\%} = (18.3\%, 23.1\%)$), and the remaining 59.2% to within-rater, within-vignette residual variance. Vignette content and stable individual differences in rater style therefore explained roughly equal, and comparatively small, shares of the variance in motive judgments. This pattern was specific to motive ratings: for pain ratings and painkiller

prescriptions, the vignette component dominated the participant component (pain ratings: vignette ICC ≈ .42, participant ICC ≈ .14; prescriptions: vignette ICC ≈ .39, participant ICC ≈ .18), indicating substantially greater consensus about pain and treatment than about motive.

S3.3.2 Model comparison on painkiller prescription

Table S7 shows the model comparisons on painkiller prescriptions in Experiment 3. The pattern parallels the model comparisons reported for pain ratings and motive judgments in the main text: the Integrated model best predicts human prescriptions, followed by Direct-LLM, Experientialist, and Physicalist models.

Table S7

Model comparison on Experiment 3 painkiller prescription. Pearson r (with 95% CI) between each model's predicted prescription and average human prescription, across $n = 1711$ *vignettes.*

Model	r [95% CI]
Integrated	0.79 [0.78, 0.81]
Experientialist	0.65 [0.63, 0.68]
Physicalist	0.42 [0.39, 0.46]
Direct-LLM	0.75 [0.73, 0.77]

S3.3.3 Explaining away across all pain reports

Our factorial stimulus design let us isolate the explaining-away effect from any confound with the content of the pain report itself. Each of the 48 unique pain reports is paired with all 36 combinations of event magnitude and physical evidence, yielding 48 report-matched vignette sets in which the reported pain is held constant while the surrounding event and evidence vary. Within each report-matched vignette set, we can therefore ask how suspicion of an ulterior motive relates to the inferred pain, with the pain report held fixed. The explaining-away account predicts a negative relationship: the more a

participant suspects an ulterior motive, the less pain they attribute. Figure S2 plots this relationship for every report-matched vignette set. In 47 of the 48 sets the slope is negative ($p_{binom} = 1.7 \times 10^{-13}$), and the Integrated model reproduces the same negative relationship for all of the sets where the pain report is non-zero (42 of the 48). This effect occurs for every non-zero pain report.

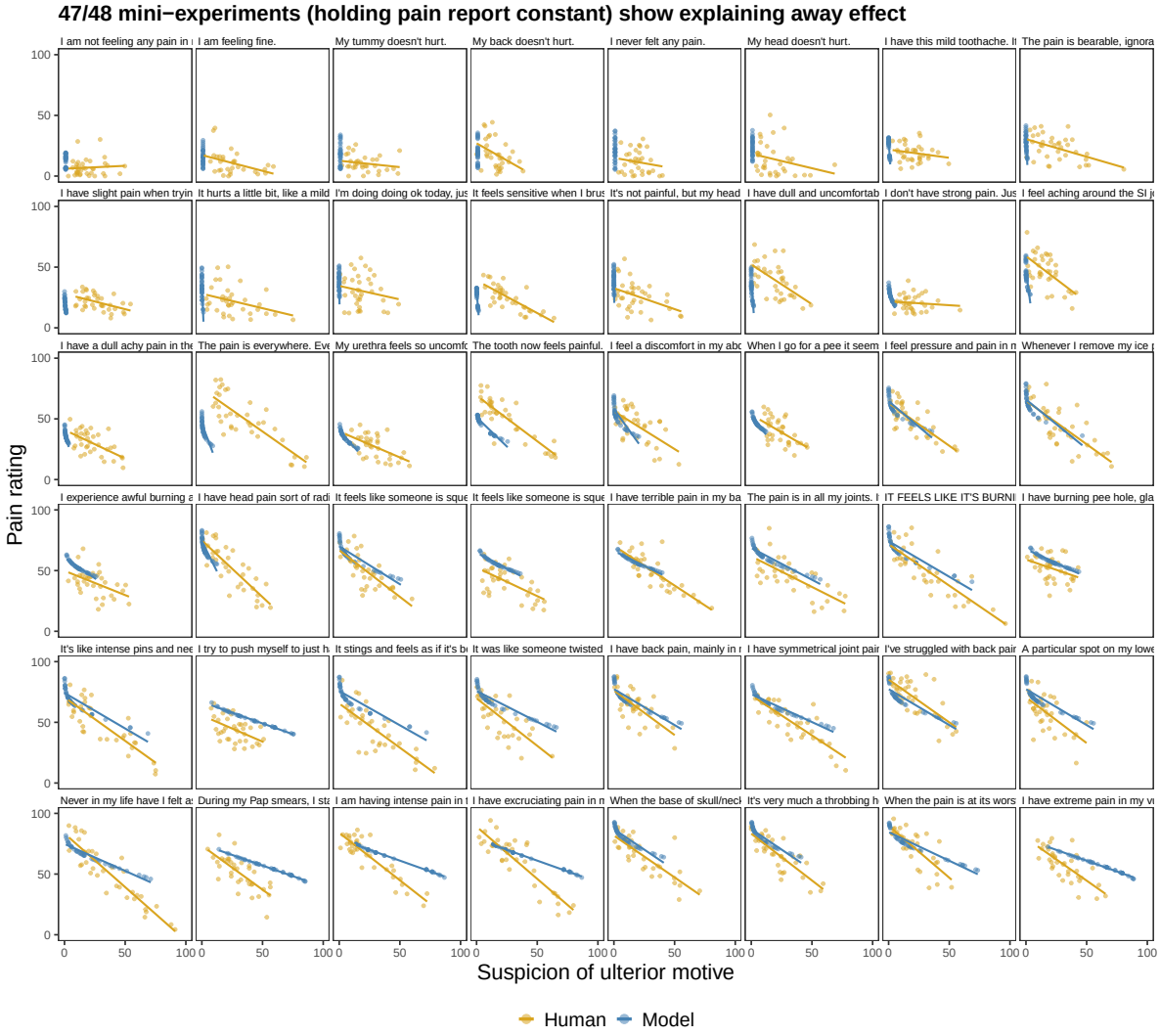

Figure S2

Explaining away effect in all 48 report-matched vignette sets. Each panel holds the pain report constant and varies the event magnitude and physical evidence across its 36 vignettes; the panel title gives the start of the pain report. Within each panel, the x-axis represents the average suspicion of an ulterior motive and the y-axis represents the average pain rating. Blue points and lines show human judgments; red points and lines show the Integrated model's predictions. Panels are ordered by LLM-rated pain severity of the report (lowest top-left to highest bottom-right). In 47 of 48 report-matched vignette sets, the human slope is negative, demonstrating the explaining-away effect.

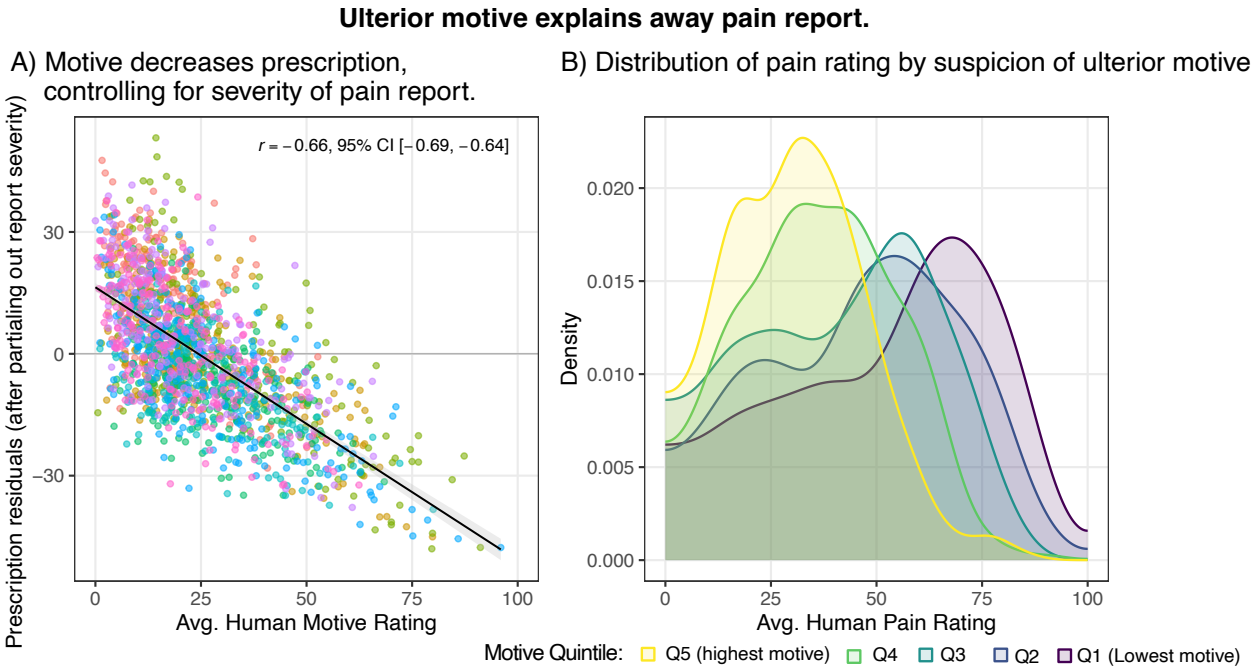

Figure S3

Ulterior motive explains away pain report (human judgments). (A) Residual prescription (after partialling out the LLM-rated severity of the pain report) plotted against the average human motive rating; the negative slope indicates that suspecting motive lowers prescribed dose beyond what the report itself implies. (B) Distribution of average human pain ratings across vignettes, binned into quintiles by humans' average motive rating. As suspected motive increases (lighter colors), the distribution of inferred pain shifts toward lower values, demonstrating the explaining-away effect.

S3.3.4 Explaining away within experiment

Fig. S3 further visualizes the explaining-away effect in Experiment 3. Panel A shows it on residualized prescriptions, and panel B on the distribution of inferred pain across motive quintiles.

S3.3.5 Coding self-reported strategies for judging ulterior motives

A large language model (Anthropic's Claude, Opus 4.8) coded all 426 free-response answers to the debrief question, "How did you decide whether someone had an ulterior

motive?" ($n = 426$ of 432 included participants) based on a pre-specified scheme (Table S8) and ten author-coded examples. An author validated the coding on a random 20% sample ($n = 85$) plus every response the model flagged on a rarer category—drug/alcohol, intimate-credible, or intimate-suspicious.

The model coded responses as reflecting comparisons between the physical evidence and the pain report more conservatively than the human coder ($\kappa = .57$; 33% vs. 49% of the validation sample) because the LLM required the comparison to be explicitly stated, whereas the human coder accepted implied comparisons. The main text reports the more conservative count. Counts for the rarer categories (drug/alcohol, $n = 17$; intimate-as-credible, $n = 6$; intimate-as-suspicious, $n = 3$) are the human-coded values, since the human coder verified every response the model flagged on those categories.

Table S8

Coding scheme for free responses about strategies for assessing motive and their prevalence across all 426 responses. Categories are not mutually exclusive, except that a response could not be coded as both intimate-credible and intimate-suspicious.

Category	**Definition**	**Prevalence (out of 426)**
Evidence–report comparison	Relates the pain report to the physical evidence, test/exam results, diagnosis, physiological cause, or the injury/event	142 (33%)[a]
Physical evidence, no explicit comparison	Assesses or names the physical evidence or the injury/event, but does not explicitly compare it with the pain report.	179 (42%)[a]
Drug/alcohol use	Cites drug use, substance use, alcohol, or addiction as a suspicion cue.	17 (4%)[b]
Intimate–credible	Treats intimate, embarrassing, or genitourinary complaints as *more* credible (people rarely fabricate vulnerable disclosures).	6 (1%)[b]
Intimate–suspicious	Treats intimate or sexual complaints as *less* credible (hard to prove, or a way to avoid intimacy).	3 (1%)[b]
Uncodable	Too vague, circular, or off-topic to convey a usable strategy.	≈0

[a]From the model's coding of all 426 responses. On a random 20% sample that a human coder also coded, agreement on the split between explicit comparison and no explicit comparison was moderate ($\kappa = .57$), with the model stricter, so the explicit-comparison count (142) is a conservative estimate. [b]Human-coded counts; the human coder verified every response the model flagged on these categories (drug/alcohol $\kappa = .94$, intimate-credible $\kappa = .84$).

S3.4 Cross-experiment analyses

S3.4.1 Fitted model recovers the event–evidence weighting asymmetry

Humans weighed physical evidence higher than the inciting event. This asymmetry can be captured through fitting the parameters of the Integrated model. Figure S4 extends Figure 9 from the main text by adding a third column depicting the fitted Integrated model with noise parameters fit separately for each experiment and dependent variable. Relative to the default model, the fitted model assigns higher weight to evidence and lower weight to the event, bringing both weights closer to the human pattern.

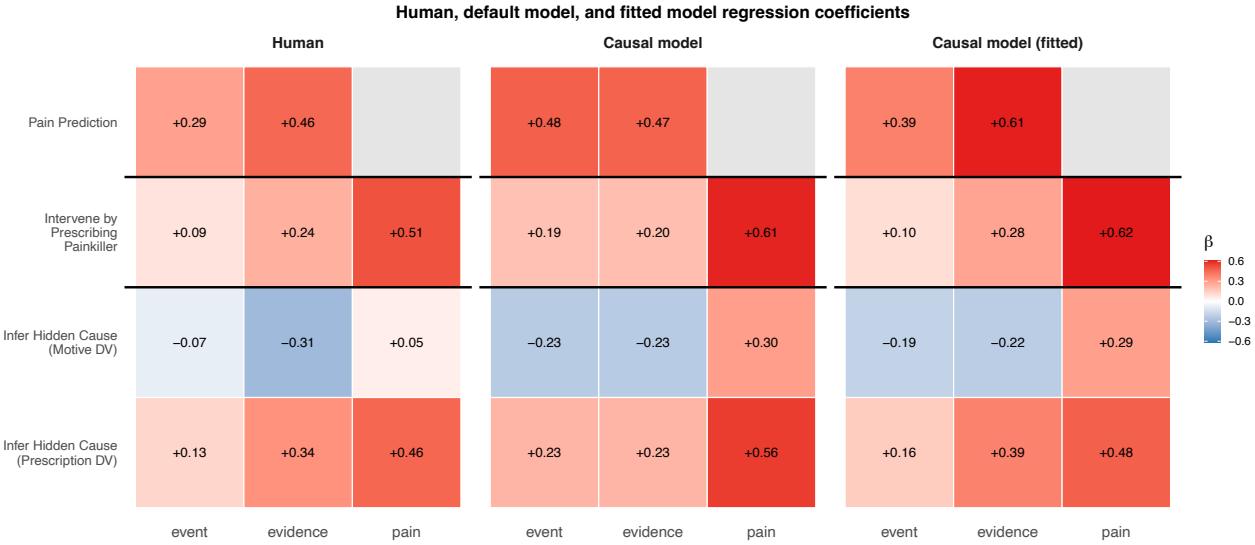

Figure S4

Regression coefficients for human judgments (left), the Integrated model with default parameters (center), and the Integrated model with per-DV fitted parameters (right). Each cell shows the raw regression coefficient from $\hat{y} \sim event + evidence + pain\ report$; red = positive, blue = negative. Fitting the model parameters substantially reduces the event–evidence asymmetry seen in the default model and brings the causal model's weighting pattern closer to the human pattern.

S3.4.2 Body-part residual analysis

As an exploratory analysis, we asked whether the Integrated model leaves any body-part-specific structure in human judgments unexplained, beyond what the LLM-rated stimulus features (the model's own inputs) already account for. For each experiment, we computed each vignette's mean human rating on that experiment's primary judgment—pain rating in Experiment 1, painkiller prescription in Experiment 2, and suspected-motive rating in Experiment 3—and regressed it on (i) the Integrated model's prediction for that vignette and (ii) the LLM-rated stimulus features that the model takes as input (event magnitude, physical evidence, and pain report where applicable; Experiment 1 has no pain report so this feature is dropped). The residual from this regression is the part of the mean judgment that neither the model prediction nor the stimulus features can account for on their own. We regressed out both the model's prediction and its input features—rather than simply comparing raw ratings across body parts—so that any remaining body-part structure reflects something the Integrated model cannot capture, and not body-part differences in the features the model already uses. These residuals are plotted in Fig. S5.

Residuals are not uniform across body parts. We ran a one-way ANOVA of these vignette-level residuals on body-part category for each experiment. Residuals are decisively non-uniform across body parts in every experiment: Exp 1 (Pain Prediction, $n = 269$ vignettes): $F(7, 261) = 6.07$, $p = 1.4 \times 10^{-6}$; Exp 2 (Pain Inference, $n = 1{,}709$): $F(7, 1701) = 39.10$, $p < 2 \times 10^{-16}$; Exp 3 (Ulterior Motive, $n = 1{,}711$): $F(7, 1703) = 18.85$, $p < 2 \times 10^{-16}$.

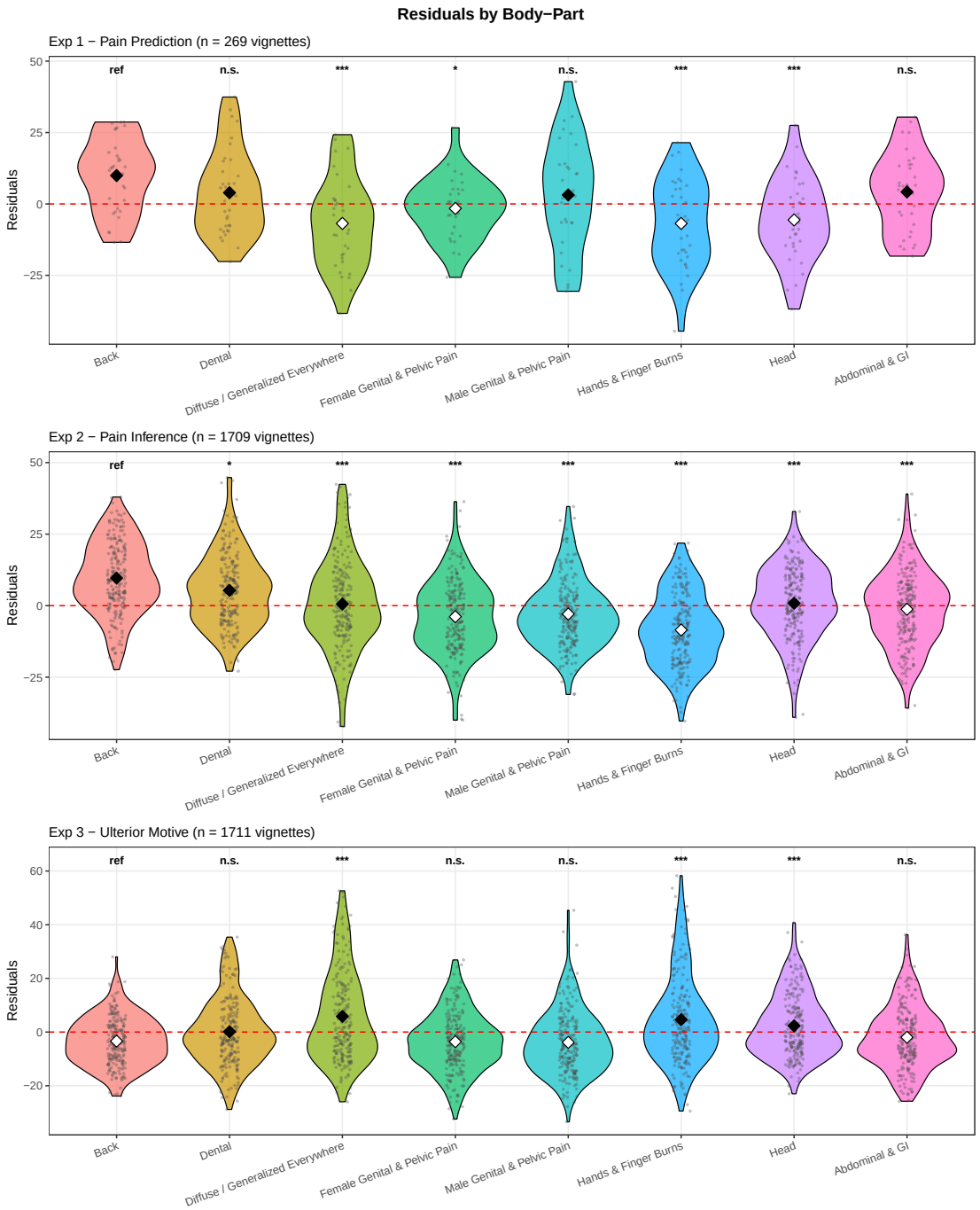

Figure S5

Vignette-level residuals by body-part category in each experiment. Each gray point is one vignette; the diamond marks the body-part mean (filled black when the mean is above zero, white when below). The dashed red line at zero indicates that the Integrated model plus LLM-rated stimulus features (event magnitude, physical evidence, pain report where applicable) predict the mean human judgment for that body part on average. Asterisks above each non-reference violin show Tukey-HSD significance of the contrast with Back (the reference category, marked `ref`*):* `***` $p < .001$, `**` $p < .01$, `*` $p < .05$, `n.s.` *otherwise.*

To localize the body parts driving these omnibus effects, we ran Tukey's HSD on the same residuals using Back as a common reference category across all three experiments (Table S9). Reported p-values are family-wise error corrected for the 28 within-experiment pairwise comparisons; only the 7 contrasts vs. Back are tabulated. In Exp 1 and Exp 2, Back's mean residual is among the highest (humans give higher pain ratings and prescriptions than the model predicts for back-pain vignettes). Hands & Finger Burns has the most strongly negative mean residual in both experiments, alongside Diffuse / Generalized Everywhere in Exp 1 and Female Genital & Pelvic Pain in Exp 2. In Exp 3 the pattern flips: Back has one of the lowest mean residuals, while Hands & Finger Burns, Diffuse / Generalized Everywhere, and Head are significantly more motive-suspicious than Back ($p < .001$ for each).

Table S9

Tukey HSD contrasts vs. Back per experiment. Each body part's value is its mean *vignette-level residual from the regression of mean rating on Integrated model prediction + stimulus features (Section S3.4.2); p is the Tukey-HSD-corrected p-value for the pairwise contrast between that body part's mean residual and Back's.*

	Exp 1 (Pain)		Exp 2 (Rx)		Exp 3 (Motive)	
Body part	resid.	p	resid.	p	resid.	p
Back (ref)	+9.95	—	+9.63	—	−3.38	—
Abdominal & GI	+4.20	.745	−1.27	$<10^{-12}$	−1.95	.947
Dental	+3.93	.682	+5.35	.017	+0.14	.092
Male Genital & Pelvic Pain	+3.15	.591	−2.92	$<10^{-12}$	−3.84	1.00
Female Genital & Pelvic Pain	−1.58	.038	−3.85	$<10^{-12}$	−3.59	1.00
Head	−5.60	5.1×10^{-4}	+0.85	1.4×10^{-10}	+2.31	1.5×10^{-4}
Hands & Finger Burns	−6.82	1.0×10^{-4}	−8.50	$<10^{-12}$	+4.62	6.6×10^{-9}
Diffuse / Generalized Everywhere	−6.83	9.9×10^{-5}	+0.57	3.3×10^{-11}	+5.84	1.1×10^{-11}

Gender effects on the residuals. The body-part pattern might differ between raters of different genders, since gender is associated with anatomy and therefore determines the kinds of pain a person can experience or has experienced, which might in turn affect pain ratings. We therefore also explored gender effects. For each vignette in each experiment, we computed the mean rating given by men and the mean rating given by women and subtracted the model-plus-features regression's prediction for that vignette (the same prediction used in Section S3.4.2). This gives one residual per (vignette, gender) pair. Comparing both genders against this single, rater-independent benchmark puts them on the same scale—each gender's deviation from what the model and features predict—so that a systematic difference between men and women registers in the residual itself. We then ran a 2-way ANOVA of these residuals on body part × gender. Free-text gender responses were harmonized to man/woman; non-binary or unclear responses were excluded.

No body-part × gender interaction reached significance in any experiment: the residual pattern across body parts is essentially the same for men and women (Exp 1: $F(7, 481) = 0.57$, $p = .78$; Exp 2: $F(7, 3402) = 1.24$, $p = .28$; Exp 3: $F(7, 3406) = 0.95$, $p = .47$). Gender main effects were also null in Exp 1 ($F(1, 481) = 0.01$, $p = .91$) and Exp 2 ($F(1, 3402) = 2.22$, $p = .14$). Exp 3 showed a small main effect of gender ($F(1, 3406) = 5.27$, $p = .022$): on average, men rated motive 1.3 points higher than women did, but this shift is uniform across body parts and small on the 0–100 slider.

Cocaine vignettes drive high motive residuals in the Diffuse / Generalized Everywhere category. The Diffuse / Generalized Everywhere category in Experiment 3 includes the 36 illicit-drug-use vignettes (whose Event text describes the patient as having "partied hard, consuming large amounts of hard liquor and snorting cocaine"). These vignettes ($n = 35$ retained after the standard $\geq 25\%$-of-participants-said-doesn't-make-sense exclusion) have a mean residual of +25.7 on the 0–100 motive scale and account for a substantial fraction of the largest positive residuals in Exp 3, where participants reported more motive suspicion than the Integrated

model + stimulus features could explain. Removing them lowers the body-part mean residual for Diffuse / Generalized Everywhere from +5.84 to +1.91, while the rest of the body-part residual pattern is unchanged.

References

California Simulation Alliance. (2018a). *Pain management: Geriatric post-op (75 yr)* (tech. rep.) (Standardized-patient simulation case (post-op colectomy). https://www.californiasimulationalliance.org/wp-content/uploads/2018/12/Pain_Mgt_Geri_Post-op_04_18.pdf (original URL now redirects to https://westernsim.org; subscriber access)). California Simulation Alliance (now Western States Simulation Alliance).

California Simulation Alliance. (2018b). *Post-operative assessment, case b: Pain management* (tech. rep.) (Standardized-patient simulation case (total abdominal hysterectomy). https://www.californiasimulationalliance.org/wp-content/uploads/2018/12/Post-op_Assess_Case-B_Pain_Mgt_04_18.pdf (original URL now redirects to https://westernsim.org; subscriber access)). California Simulation Alliance (now Western States Simulation Alliance).

Gilbert Program in Medical Simulation, Harvard Medical School. (2012). *Simulation casebook* (Z. Howard, J. Siegelman, E. Guterman, E. M. Hayden, & J. A. Gordon, Eds.; Draft of the 1st edition (2011), updated 2 March 2012, tech. rep.) (Contains Cocaine Intoxication, Migraine, and Acute Cholecystitis cases. https://learn.partners.org/docs/84/HMS-Simulation-Case-Book-2012-03-02.pdf). G. S. Beckwith Gilbert and Katharine S. Gilbert Medical Education Program in Medical Simulation, Harvard Medical School.

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*.

University of Wisconsin School of Medicine and Public Health, Department of Family Medicine. (2014). *Primary care clerkship: Practice OSCE scenarios* (tech. rep.) (Practice OSCE scenarios (incl. Scenario D, carpenter with low-back pain; pediatric sore-throat scenario). https://www.fammed.wisc.edu/files/webfm-

uploads/documents/med-student/pcc/practice-osce-scenarios.pdf). University of Wisconsin School of Medicine and Public Health.