Week 9 — Inverse RL, Social Cognition & POMDPs

Friday, June 26, 2026

Prof. Joseph Austerweil

Run the camera backwards

Admin — deadlines

- ★ **Final-project proposal due Sunday June 28, 8:00 PM** (pass/fail) — ~1 page: background, your question, method, ≥3 references.
- **Assignment 4 (Reinforcement Learning)** is out — due **Fri Jul 10, 8 PM** (Q-learning on the GardenPath grid).
- **Monte Carlo assignment** also due **Fri Jul 10, 8 PM** — pace your work around the proposal.
- **Reflections:** you need **5 of 11** (Weeks 2–12). After today, **3 chances remain** — Weeks 10, 11, 12. Check you're on track.

Next week is async — AI Safety workshop

- **Week 10 (Fri Jul 3) runs asynchronously** — no in-person session.
- There's an **AI Safety workshop** that week — you're warmly invited if it interests you. *Heads up: it runs **fully in English**.*
- I'll **record the lecture** (Bayesian nonparametrics) and post a **link to the recording** — watch it on your own time.
- The **weekly discussion thread still runs** — post and reply as usual (Week 10 is reflection-eligible).

From acting to reading minds

Last week we ran reinforcement learning **forward**: a goal becomes a policy becomes actions.

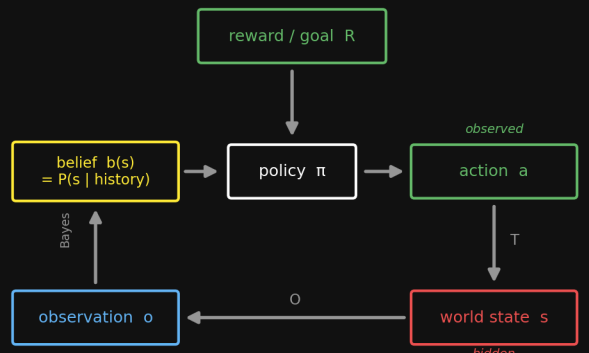
This week we run the *same machinery* **backward** — watch the actions, infer the **goal**, **belief**, or **reward** behind them. That inversion is **inverse RL**, and when the agent is a person, it *is* the computational story of **social cognition**.

The one equation for the whole week

$$P(\text{goal} \mid \text{actions}) \;\propto\; \underbrace{P(\text{actions} \mid \text{goal})}_{\text{a policy, in reverse}} \cdot P(\text{goal})$$

The likelihood `P(actions | goal)` is just a **policy** — how a goal-seeker would act. Last week we worked out *how good each action is* for a goal; this week we assume the agent **mostly takes the good actions, but slips up now and then** — and we read that backward. Everything today is this one inversion, on a richer hidden variable each time: **goal** → **belief** → **reward**.

The agent model — a POMDP



Week 8 ran this FORWARD (all pieces known → behavior). Week 9: watch the behavior, infer a hidden piece.

Every agent today fits **one picture**: a goal/**reward** R drives a **policy** π that picks an **action** a; the world has a hidden **state** s the agent sees only through a noisy **observation** o, so it tracks a **belief** $b(s)$. That is a **POMDP** — a *partially observable* MDP. Week 8 ran it **forward** (every piece known → behavior); all week we **invert** it — watch behavior, infer one hidden piece.

A map of the unknowns

The rest of the day is **this one model with a different piece hidden** each time:

- **reward R** hidden, world seen → **goal inference / IRL** (next)
- **belief b** hidden, possibly *false* → **theory of mind**, POMDPs
- **reward R at scale** → MaxEnt IRL, RLHF
- the **observer's belief** is what *you* steer → **teaching / showing**
- two agents, **one shared hidden R** → **alignment**

Same inversion, a richer unknown each time — we *point at the lit box*, we don't re-derive it.

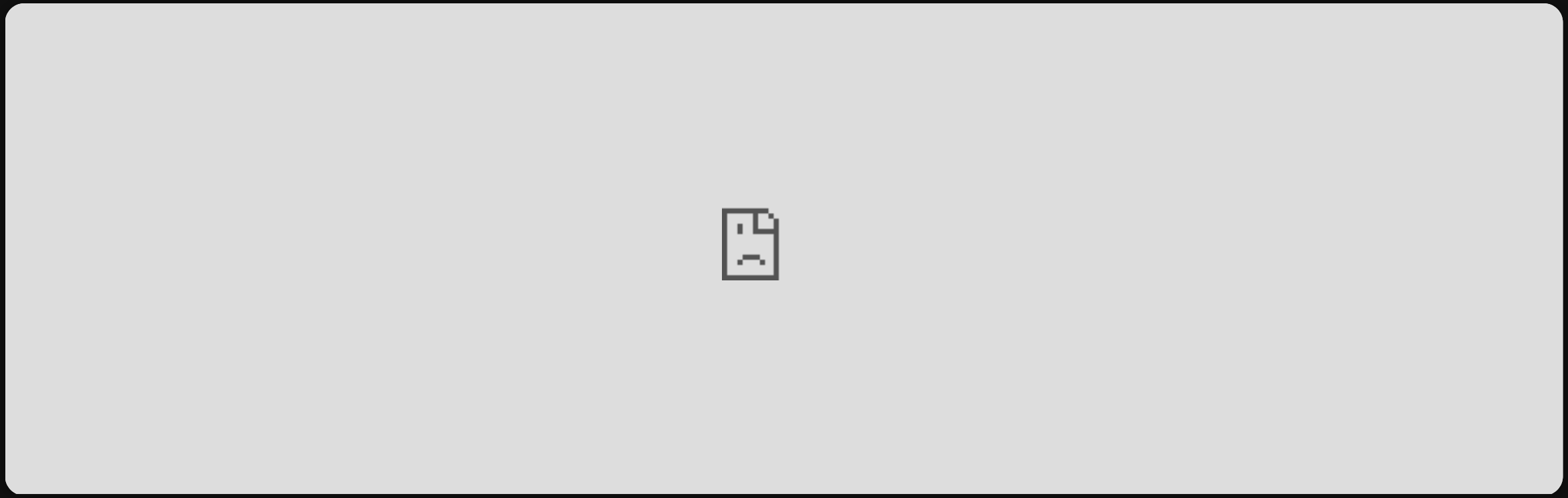
Agenda — 2 hours

Run the camera backwards	0:00
Goal inference as inverse planning	0:07
What is theory of mind?	0:37
POMDPs — inferring beliefs	0:45
Break	1:07
Inverse RL — recovering rewards	1:12
Teaching, flipped	1:26
Alignment & LLM Theory of Mind	1:38
Close & Week 10	1:52

Goal inference as inverse planning

Where we are

Recall: the forward model



Week 8’s forward solve: value iteration turns a reward into a value, then a policy. *This* is the machine we now run backward.

Notation lock-in

- **trajectory** τ — the path of states and actions we observe
- **goal** g — the hidden thing we infer; **reward** R, **cost** C
- **posterior** $P(g \mid \tau)$ — our belief over goals after seeing τ
- **likelihood** $P(\tau \mid g)$ — how a g-seeking agent would move
- **rationality** β — how decisively the agent picks the best-valued action ($\beta \to \infty$ greedy / pure exploit, $\beta \to 0$ random); formula next slide
- carried from Week 8: state s, action a, policy π, value Q, discount γ

The inversion is Bayes' rule

$$P(\text{goal} \mid \text{actions}) \propto P(\text{actions} \mid \text{goal}) P(\text{goal})$$

posterior
(what we infer)

likelihood = a policy
(how a goal-seeker acts:
softmax over Q), in reverse

prior
over goals

Posterior (what we infer) $\propto$ **likelihood** (how a goal-seeker acts — a softmax policy, read in reverse) $\times$ **prior** (over candidate goals). Next: build each piece.

Build the inversion

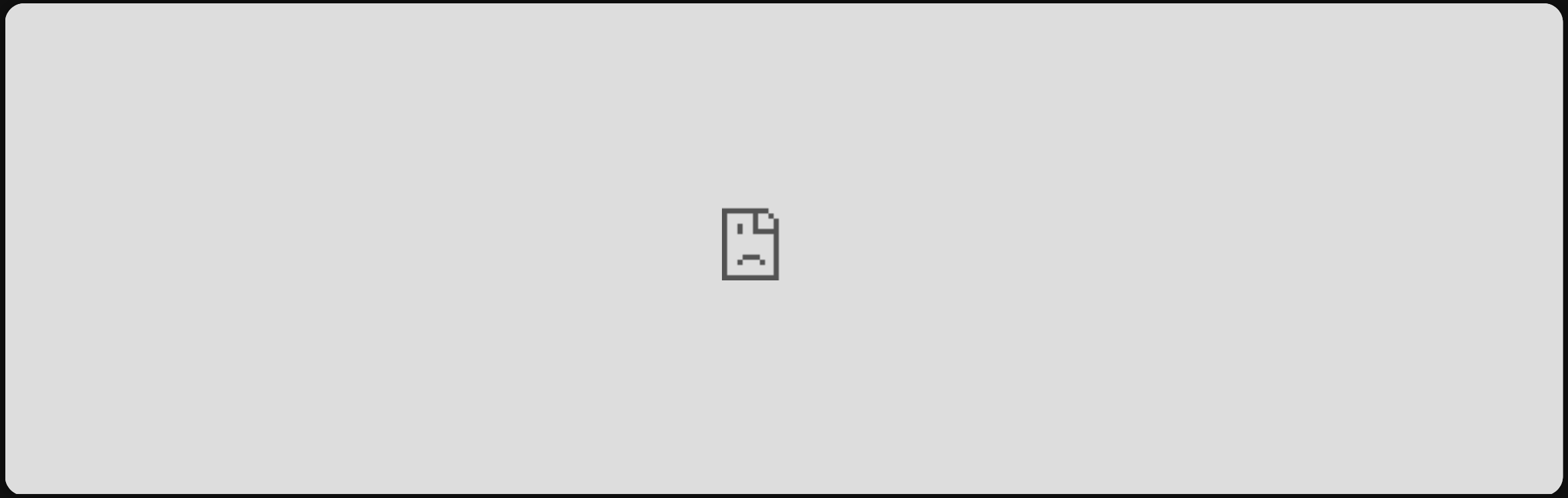
1. **Forward** — value iteration (Week 8) gives the Q-values Q_g; wrap them in a **softmax** (best action most likely, the rest still possible) → a noisy-rational policy $\pi_g(a \mid s) \propto \exp(\beta\, Q_g(s, a))$ *(new this week; one per candidate goal)*
2. **Likelihood** — $P(\tau \mid g) = \prod_t \pi_g(a_t \mid s_t)$
3. **Prior** — $P(g)$ over candidate goals
4. **Posterior** — $P(g \mid \tau) \propto P(\tau \mid g)\, P(g)$

In GenJAX — invert the MDP

```
LOGITS = beta * Q_per_goal                    # β·Q logits; categorical
@gen
def observer(states):                         # observed actions = the
    g = categorical(jnp.log(goal_prior)) @ "goal"     # the hidde
    for t in range(T):
        categorical(LOGITS[g, states[t]]) @ f"a_{t}"  # softmax p
post = normalize([exp(observer.assess(cm(g, acts), (states,))[0])
```

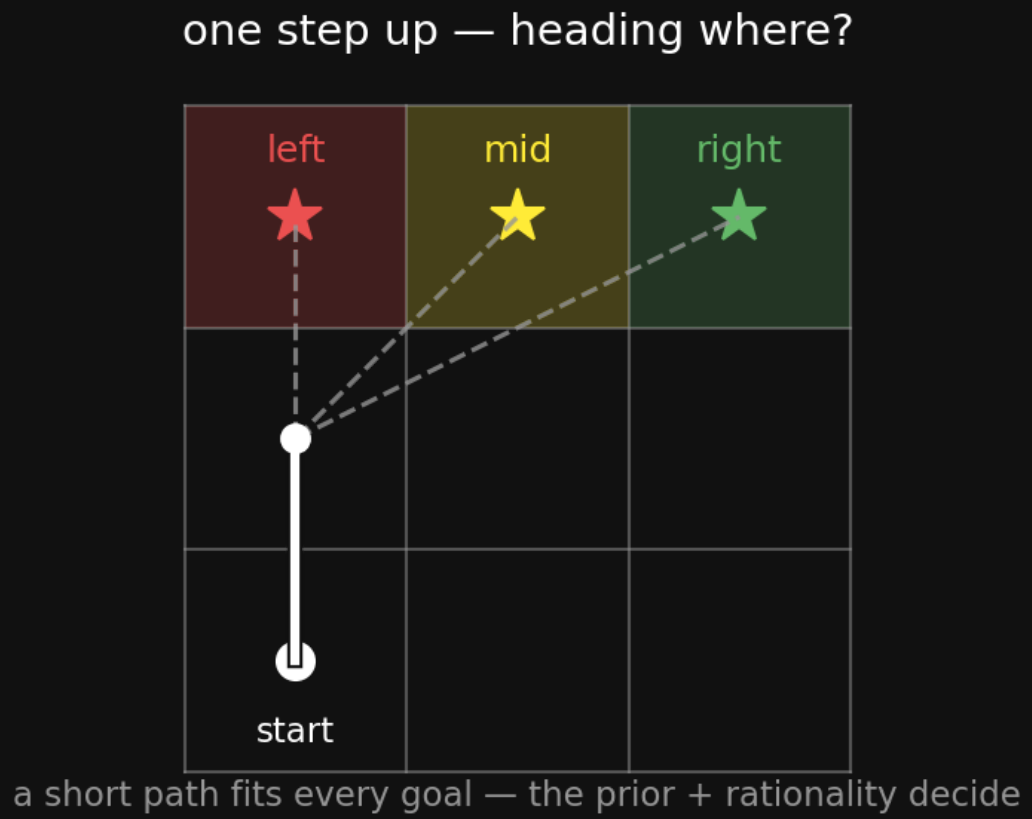
The hidden variable we recover **is “which MDP are we in?”**

Watch it infer — live



Drive the agent (arrow keys / 🎮) — the posterior over goals updates each step; drag β to feel the rationality assumption bite.

The inversion is ill-posed



A few steps fit **many** goals. The single confident answer comes from the **prior** + the **rationality assumption** — it's a *posterior*, not a unique solution.

Baker, Saxe & Tenenbaum (2009): the *changing-goal* model tracks human goal judgments at **r = 0.97** (vs 0.82 single-goal).

People reason about cost vs reward

Naive utility calculus (Jara-Ettinger et al. 2016): we assume agents **maximize reward minus cost**, and we infer *both* from behavior — present from infancy through adulthood. Inferring the reward is the same move as inferring the goal.

Watch **Chibany** trek all the way across town for a bento box and open it beaming — you infer two things at once: the trip was **costly**, so the contents must have been **worth it**. You'd be *surprised* if it weren't his beloved **tonkatsu**.

Poll — which is *not* inverse RL?

A. Find an optimal policy by acting in an MDP

B. Learn a reward from an agent’s behavior

C. Learn an agent’s preferences from behavior

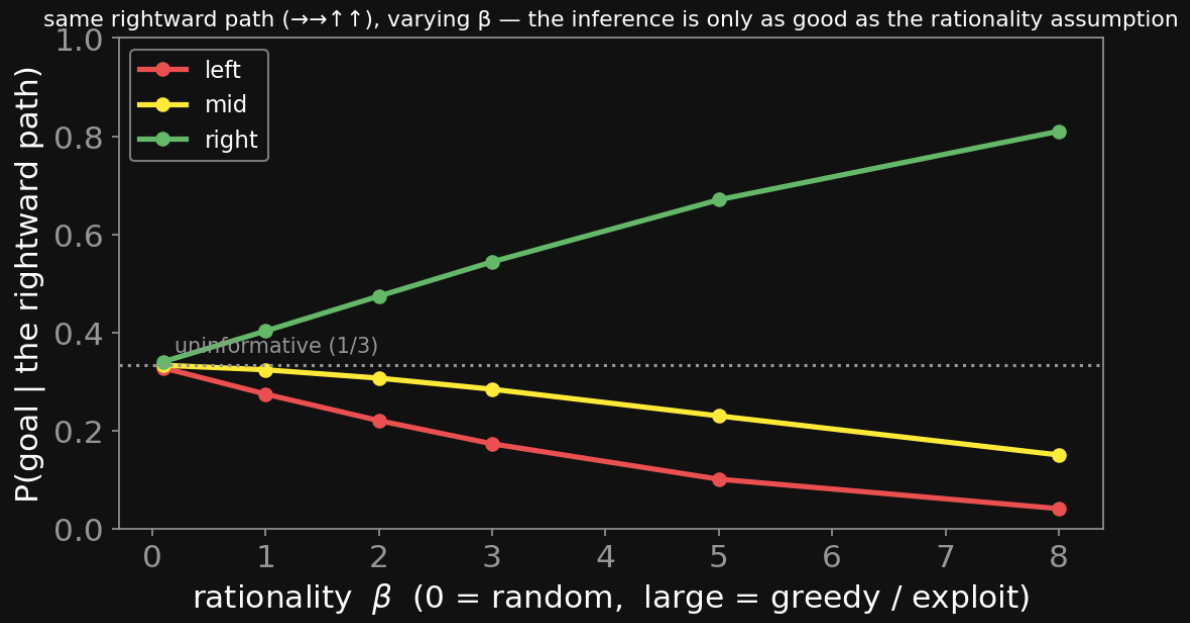
D. Learn an MDP’s transitions from behavior

Poll — reveal

A — that’s *forward* RL.

Acting *in* an MDP to find a policy is the forward problem (Week 8). B, C, D all **recover something hidden from behavior** — that’s the inverse problem.

Noisy rationality



$\pi(a \mid s) \propto \exp(\beta\, Q(s,a))$ — favors high-Q actions, imperfectly. Take the **rightward path** (→ → ↑ ↑) from the goal-inference demo: at high β it's decisive evidence the goal is **top-right**; at $\beta \to 0$ (a random agent) the *same path* says nothing.

So when an agent **blunders or detours**, the model rationalizes it by inferring a *weirder goal* — or, soon, an **unseen obstacle**. "Assume rationality, blame the goal."

What is theory of mind?

Where we are

Run the camera backwards	0:00
Goal inference as inverse planning	0:07
What is theory of mind?	0:37
POMDPs — inferring beliefs	0:45
Break	1:07
Inverse RL — recovering rewards	1:12
Teaching, flipped	1:26
Alignment & LLM Theory of Mind	1:38
Close & Week 10	1:52

Theory of mind: the false-belief test

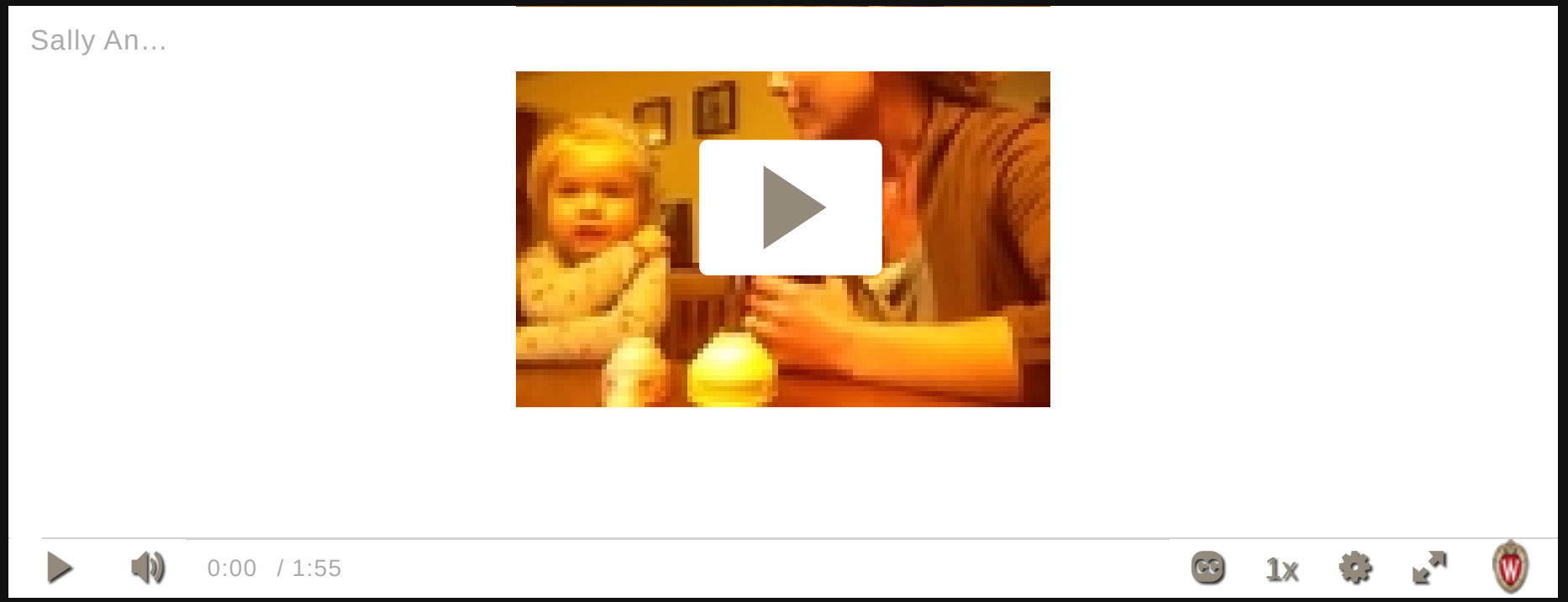
Theory of mind = attributing mental states — beliefs, desires, goals — to others (Premack & Woodruff 1978).

The classic probe is the **false-belief task** — *Sally-Anne* (Baron-Cohen, Leslie & Frith 1985):

Sally hides her marble in a **basket** and leaves. Anne moves it to a **box**. Sally returns. **Where will Sally look?**

Passing means representing a belief that **differs from reality** (she looks in the basket). Most children pass around **age 4**.

Watch: the Sally-Anne task



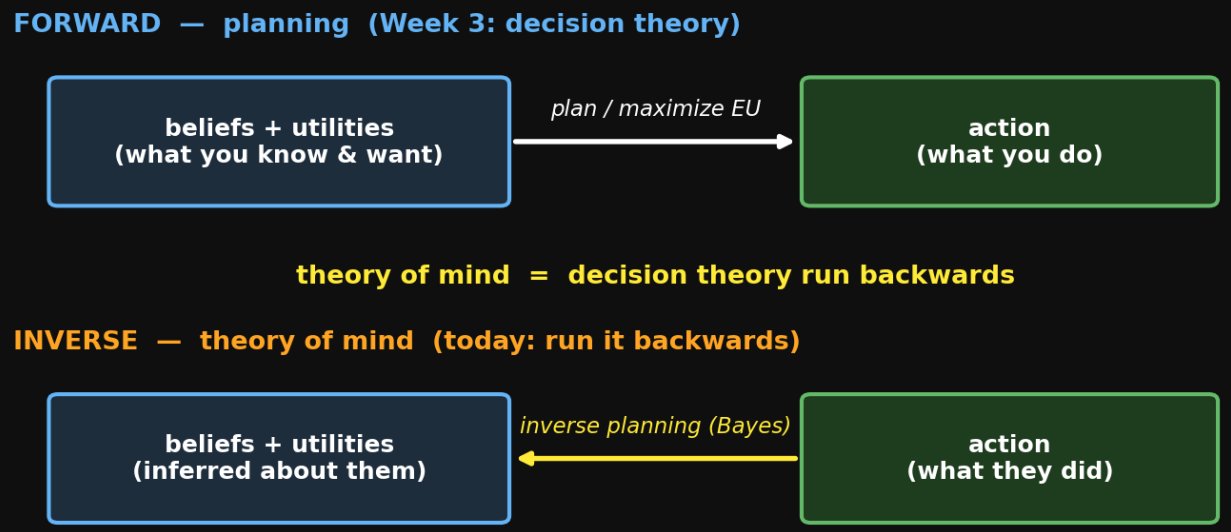
Watch when the child's answer flips from "the box" (where it *really* is) to "the basket" (where Sally *thinks* it is).

Reading a mind *is* inverse planning

How does mature theory of mind work? **Baker, Saxe & Tenenbaum (2009)** — *action understanding as inverse planning*:

- forward: beliefs + goals → action (planning)
- ToM: action → beliefs + goals (inverse planning)

Theory of mind is **your own decision theory, run backwards** on someone else — reviewed as "ToM = inverse RL" (Jara-Ettinger 2019).



The machine is old; the evidence is new

“Infer preferences from choices, assuming utility maximization” is just **revealed preference** — economics run backwards. The *theory* isn’t new. What is striking is the **empirical** reach:

- **Toddlers** infer others’ costs and rewards from a single choice (Jara-Ettinger et al. 2016; Liu et al. 2017).
- Inverse-planning models **quantitatively predict** adults’ moment-to-moment belief and desire attributions (Baker et al. 2017).

A harder test: the faux pas

A **faux pas**: someone says something they *don't realize* they shouldn't — and it lands badly. Recognizing one needs **two mental states at once**:

- the speaker holds a **false belief** (they don't know the hurtful fact), **and**
- the listener **knows** what the speaker doesn't, so the remark stings.

A faux pas is a *nested* false belief — harder than Sally-Anne, passed later (~age 9–11). Hold this exact structure; the LLM debate turns on it.

Theory of mind & autism — and a reframe

Sally-Anne was built to study **autism** (Baron-Cohen et al. 1985, “Does the autistic child have a theory of mind?”): autistic children passed control questions but **failed the false-belief question** more often — launching the **“mindblindness”** hypothesis.

Milton (2012), the double-empathy problem reframes it: the mismatch is **bidirectional** — neurotypical people are **no better** at reading autistic minds than the reverse; autistic-to-autistic communication is **as effective** as neurotypical-to-neurotypical.

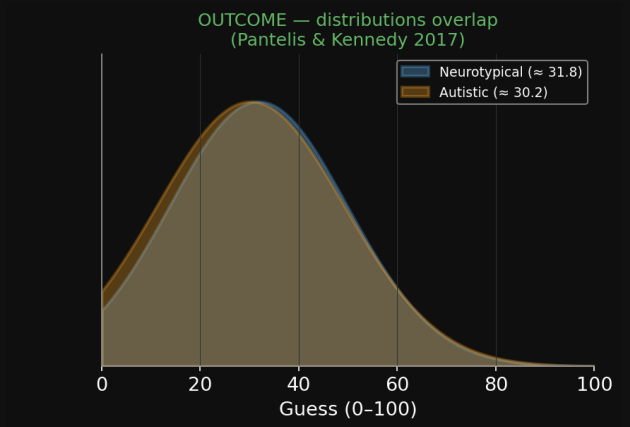
Not a broken module — **two differently-tuned inverse planners** reading each other.

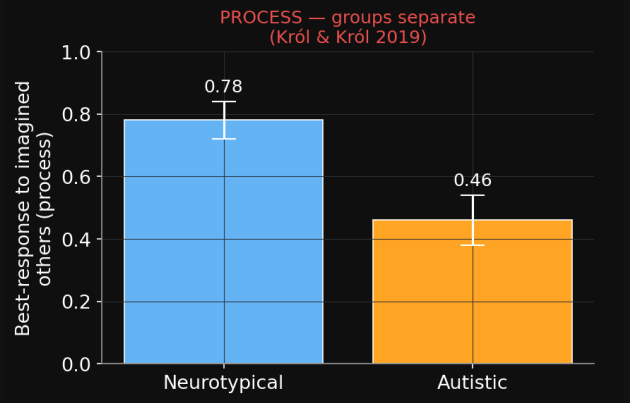
What you measure decides what you conclude

A clean prediction: deep strategic reasoning ("I think that you think…") needs theory of mind, so autistic players should reason *less* deeply. Two studies:

- **Pantelis & Kennedy (2017):** *outcomes* are **statistically indistinguishable** — ASD mean ≈ 30.2, controls ≈ 31.8. Moderate evidence *for* the null.
- **Król & Król (2019):** same outcomes — but a process tracer shows autistic players were **less strategic in process**.

Look only at the **outcome** → "no difference." Trace the **process** → a difference appears. The lesson that decides the LLM debate.



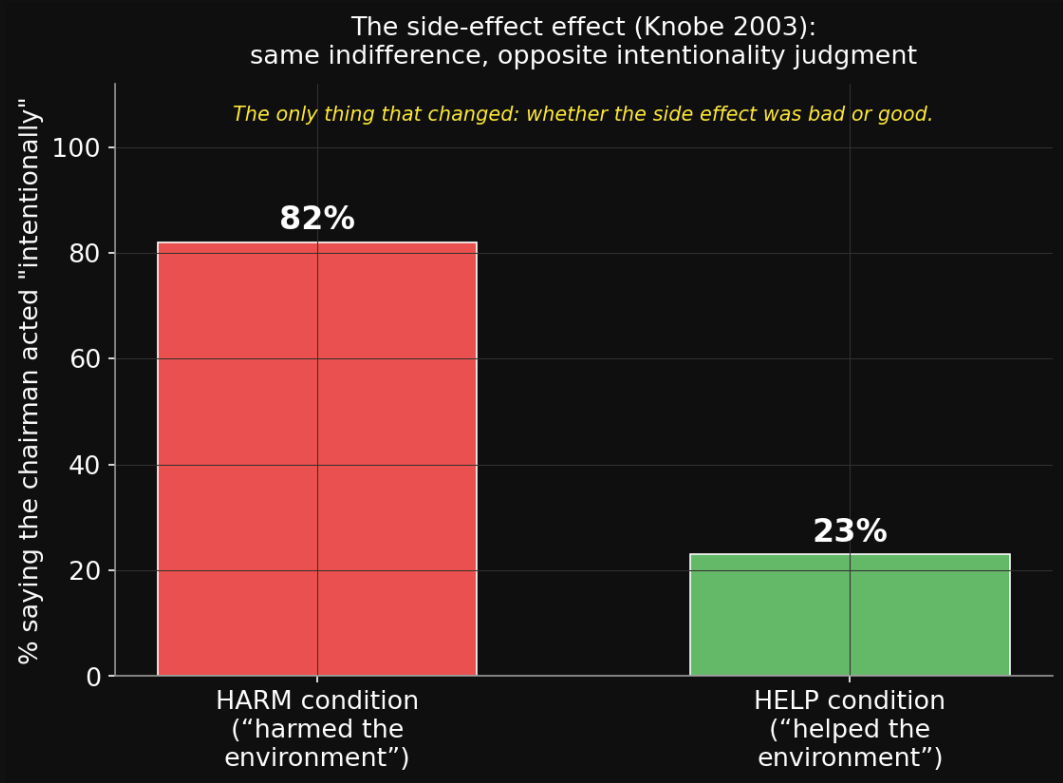


Blame bends the inference: the side-effect effect

A chairman is told a program will raise profits and **harm the environment** as a side effect. He says: *“I don’t care about the environment — I just want profit.”* It runs; the environment is harmed. **Did he harm it *intentionally*?**

Knobe (2003): 82% say *intentionally*. Flip one word to **help** — same indifferent chairman — and only **23%** say he helped intentionally.

Same mental state both times. Yet “intentional” tracks **bad vs. good**, not his actual intent — moral valence **bends** the mind-reading.

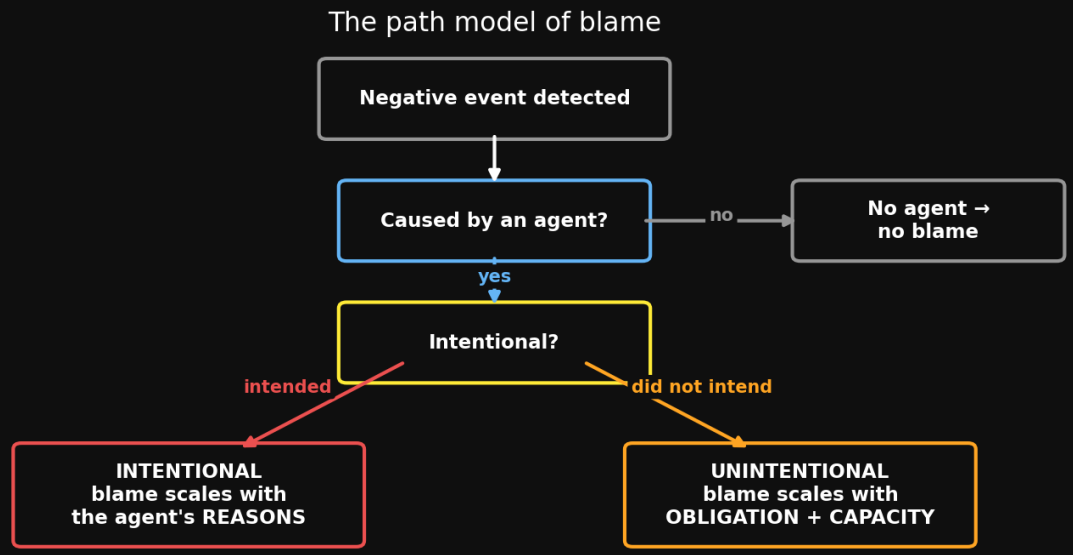


Blame is structured — and needs an agent

Much blame is **structured** (Malle et al. 2014): negative event → causal agent? → intentional? → if so, blame scales with **reasons**; if not, with **could they have prevented it**. The inverse-planning machinery, with a moral layer.

Mind perception (Gray, Gray & Wegner 2007) has two axes: **agency** (the capacity to *do*) and **experience** (the capacity to *feel*).

Blame needs an agent with **agency**. *That's* why we argue over blaming a **company**, an **animal**, or an **AI** — the alignment question, two blocks early.



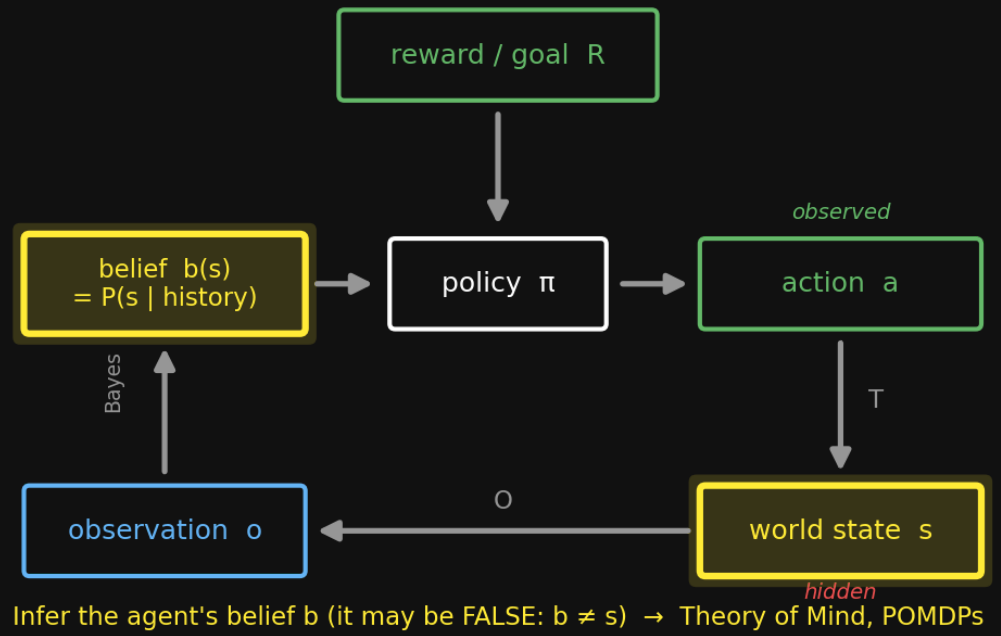
Blame is a structured inference, not a gut reaction (Malle, Guglielmo & Monroe 2014)

False belief needs a hidden belief

Sally's mistake is the key. In an **MDP** the agent sees the true state s — so it can **never** be wrong. A **false belief** means the belief b disagrees with the world: $b \neq s$.

To represent that, belief must be its **own hidden variable**, separate from the world. That model is a **POMDP** — and **Sally-Anne is a POMDP**.

Next: infer the belief b, not just the goal.



POMDPs — inferring beliefs

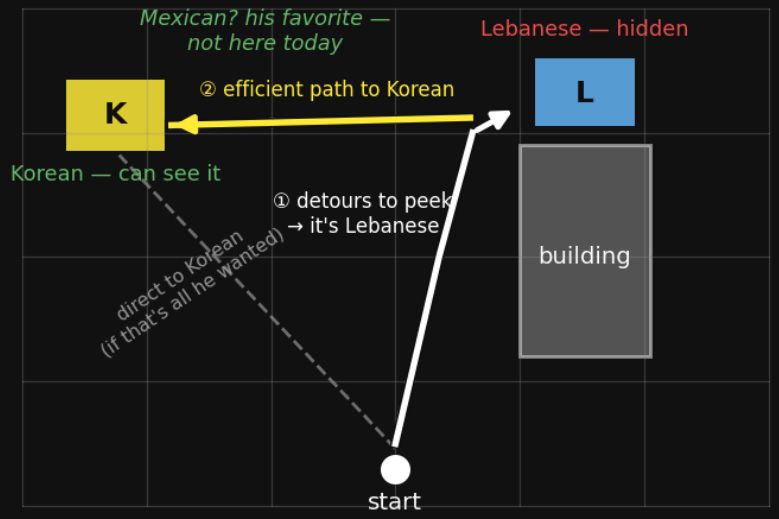
Where we are

Run the camera backwards	0:00
Goal inference as inverse planning	0:07
What is theory of mind?	0:37
POMDPs — inferring beliefs	0:45
Break	1:07
Inverse RL — recovering rewards	1:12
Teaching, flipped	1:26
Alignment & LLM Theory of Mind	1:38
Close & Week 10	1:52

The food-truck puzzle

He skips the visible Korean to peek at the hidden spot — only worth it if he hoped for Mexican

Chibany wants: Mexican (favorite) > Korean > Lebanese

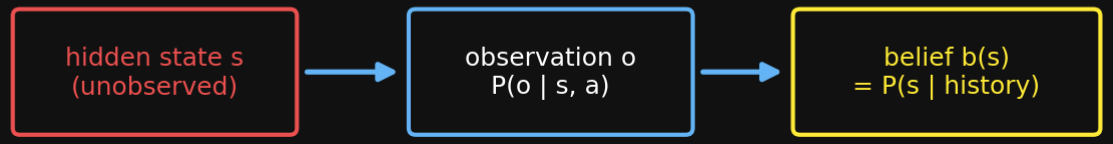


Chibany likes **Mexican > Korean > Lebanese**. He can **see** a Korean truck nearby; a building **hides the far spot**.

Instead of going straight to the **visible Korean**, he detours to peek at the **hidden** spot — the moment he sees it's **Lebanese** (not the Mexican he hoped for), he takes the **direct** path to the Korean.

Passing up a perfectly good Korean only makes sense if he was **hoping for Mexican** and *didn't know* what was hidden. So you infer a desire for a truck **not even in the scene** — impossible without modeling his **belief** (Baker et al. 2017).

Forward POMDP — the belief state



$$b'(s') \propto P(o \mid s', a) \sum_{s} P(s' \mid s, a)\, b(s)$$

belief is a sufficient statistic ⇒ a POMDP is an MDP over beliefs

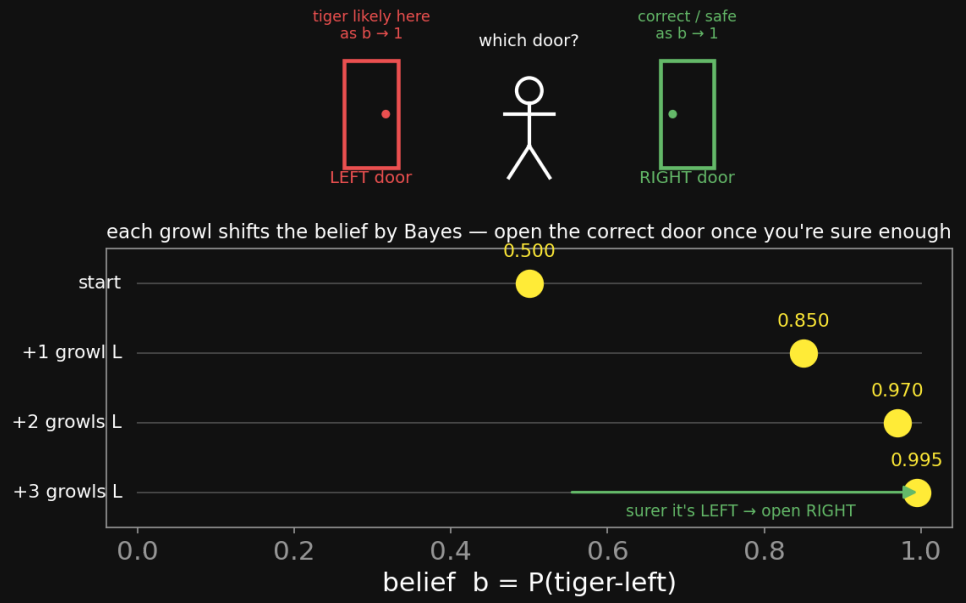
The agent never sees state s, only an **observation** o. It keeps a **belief** $b(s) = P(s \mid \text{history})$, updated by **Bayes** after each (a, o).

Don't let the new symbol fool you: $b(s)$ is **nothing exotic** — it's the same Bayesian posterior we've worked with all term, just over the *hidden* state. We write b because it's the field's standard, and because it's the **one probability a POMDP agent must track**.

The belief is a **sufficient statistic** → a POMDP is just an **MDP over beliefs**.

The Tiger problem

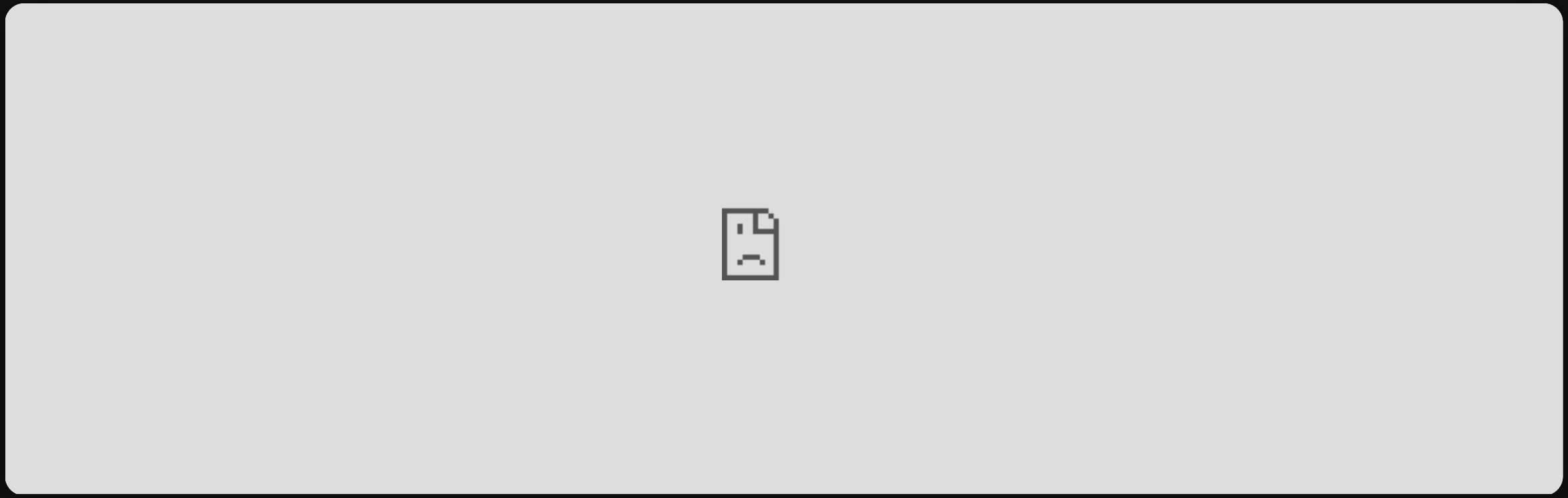
Two doors; a tiger behind one. **Listen** (cost −1) — the growl is right only **85%** of the time. Open the wrong door: **−100**; the **correct** door: **+10**. Your belief slides by Bayes:



In GenJAX — belief update = conditioning

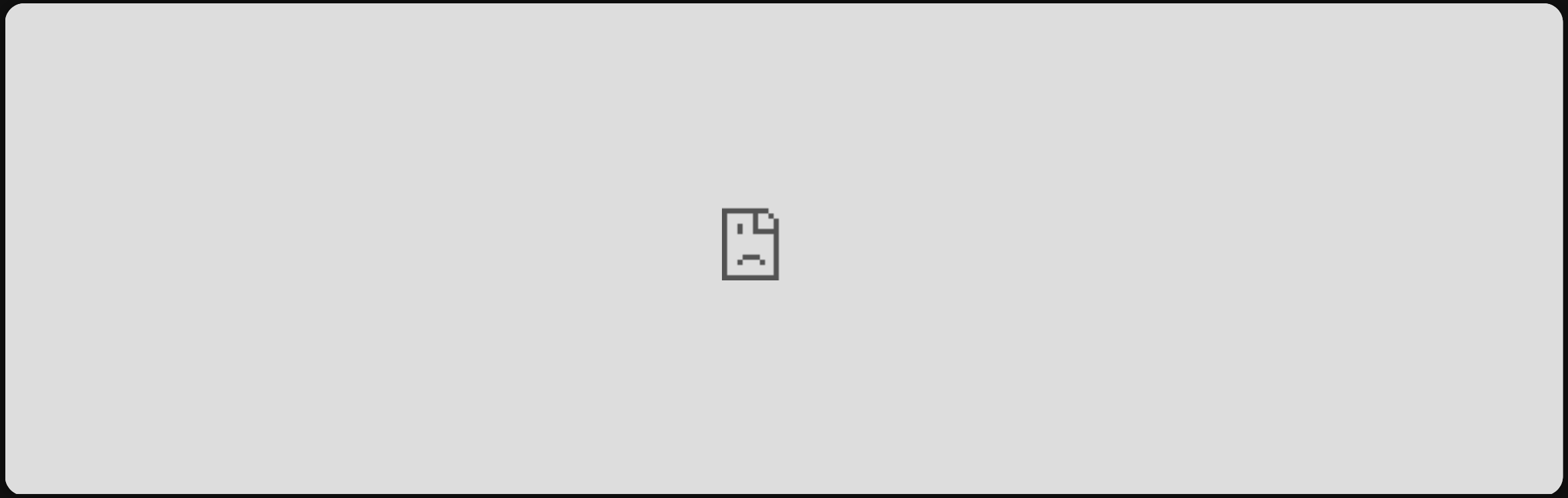
```
OBS = jnp.array([[0.85, 0.15],     # tiger-left  -> hear-left .85
                 [0.15, 0.85]])    # tiger-right -> ...

@gen
def listen(belief):
    s = categorical(jnp.log(belief)) @ "s"       # the hidden stat
    categorical(jnp.log(OBS[s])) @ "o"           # the growl we ac

# b'(s) ∝ P(o | s) b(s): enumerate the 2 states, score with asses
def update(belief, obs):
    return normalize([exp(listen.assess(cm(s, obs), (belief,))[0]
# 0.5 -> 0.85 -> 0.9698 -> 0.9945  as the growls agree
```

Track the belief — live



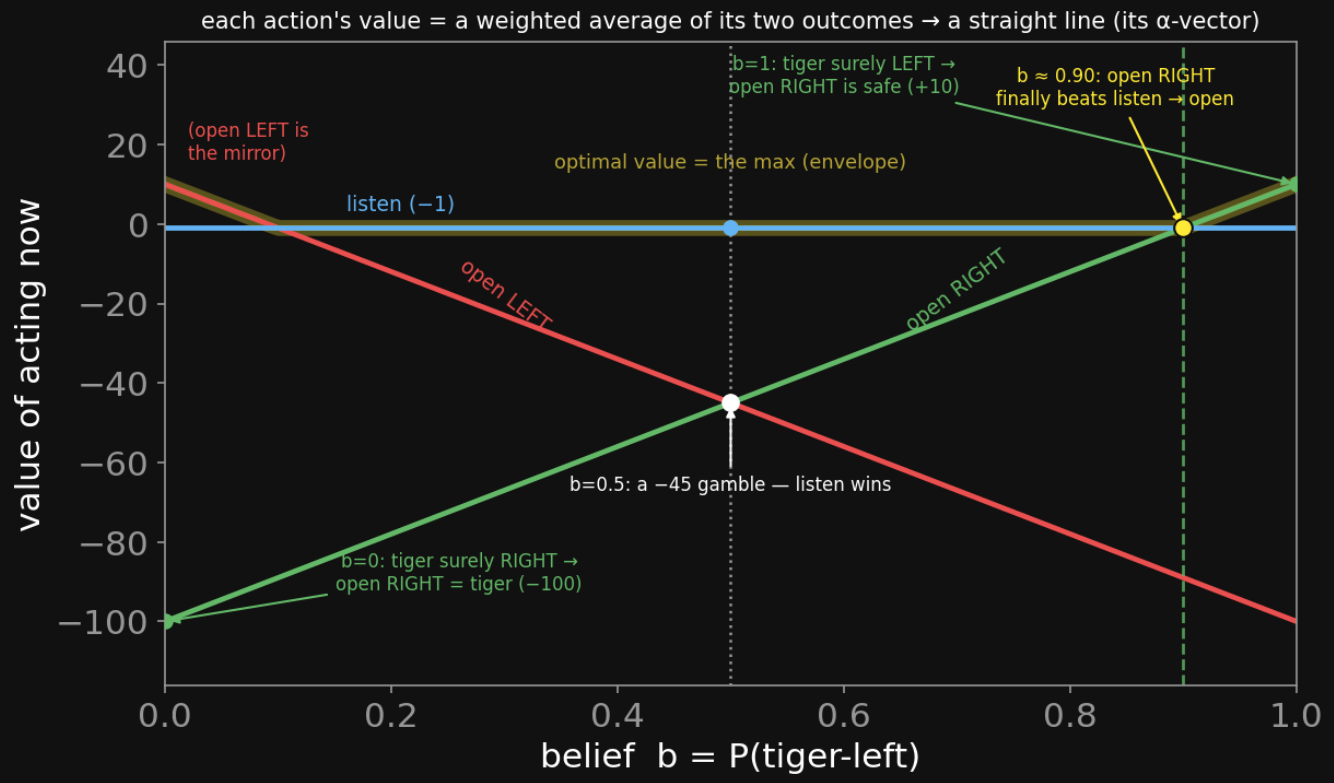
Listen → the belief slides 0.5 → 0.85 → 0.97; hear-left then hear-right cancels back to 0.5; the α-vectors show when opening beats listening.

Belief tracking, in general



The Tiger’s two-number belief is the simplest case; in general a belief is a *cloud* over states, updated by the same Bayes rule — here, a particle filter.

Value is a function of belief

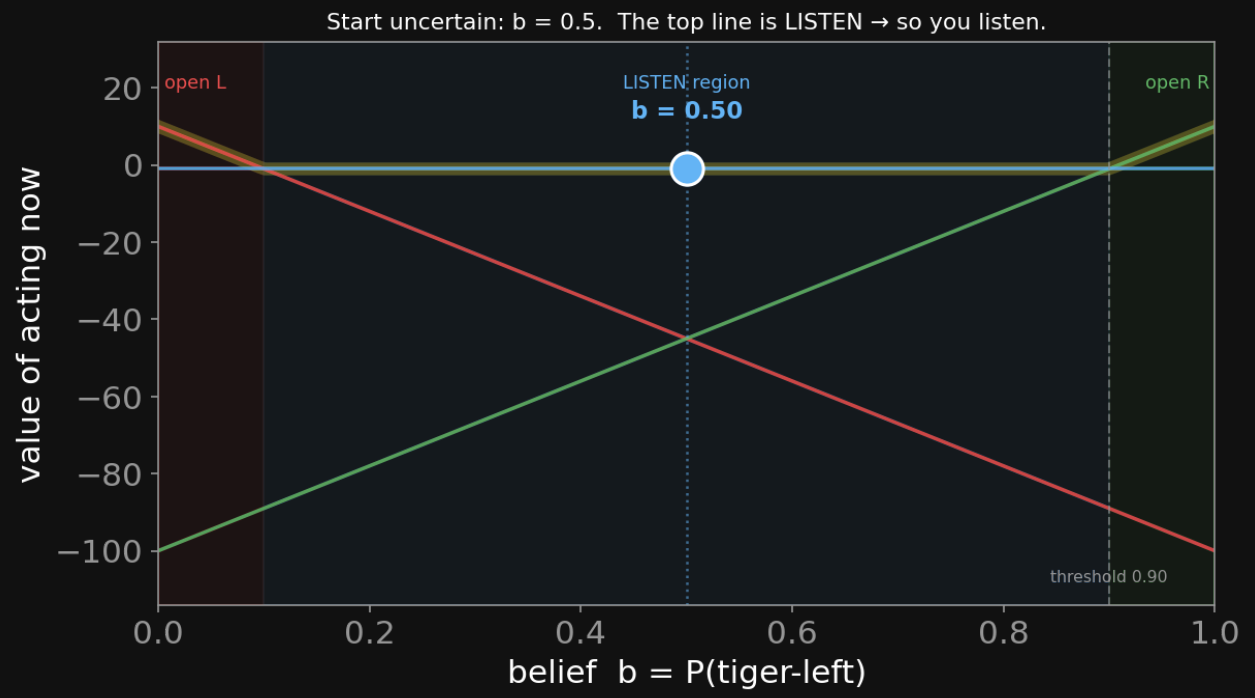


Each action's value is a **weighted average** of its two outcomes → a **straight line** in b (its **alpha-vector**). *Open right* = **+10** if tiger-left (prob b), **−100** if right: that's $110b - 100$. At $b = 0.5$ it's a **−45 gamble** — *listen* (-1) wins.

The optimal value is their **upper envelope** (the max). Its breakpoints are the **decision thresholds**: open the correct door once b passes **≈ 0.90** — where open-right's line finally beats *listen* (-1).

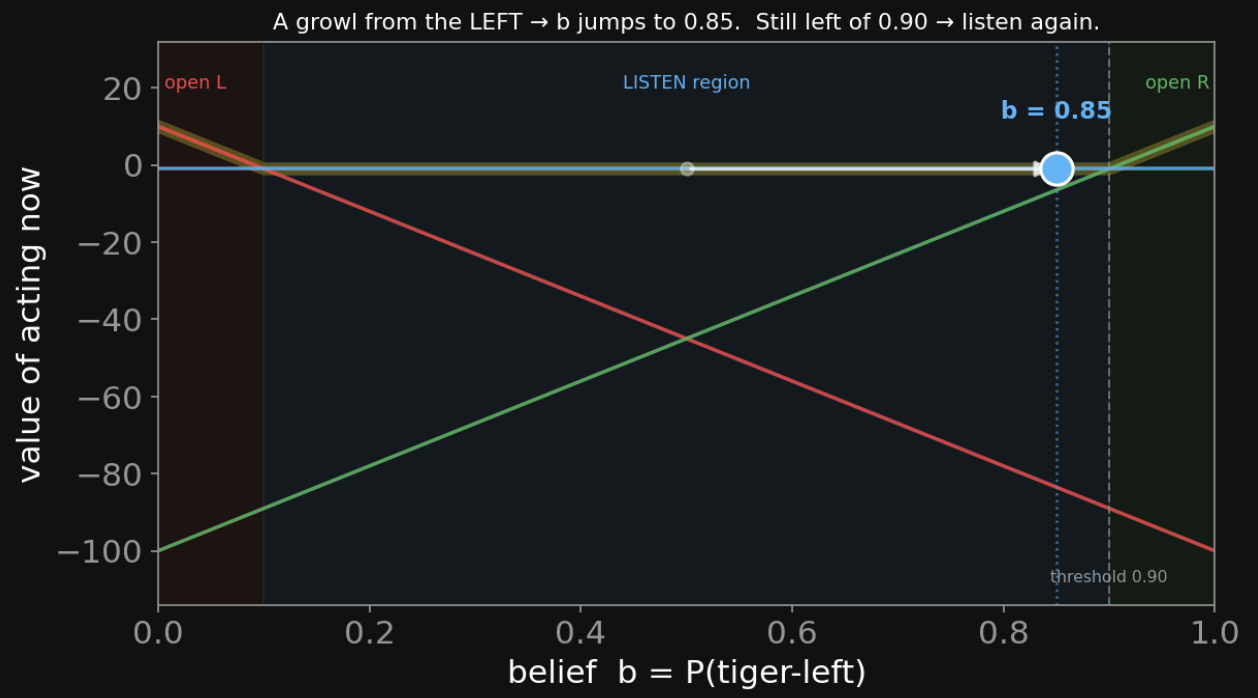
Solvers scale this up: exact α-vector VI → offline **QMDP / SARSOP** → online **POMCP / DESPOT**.

How the decision unfolds — ① uncertain



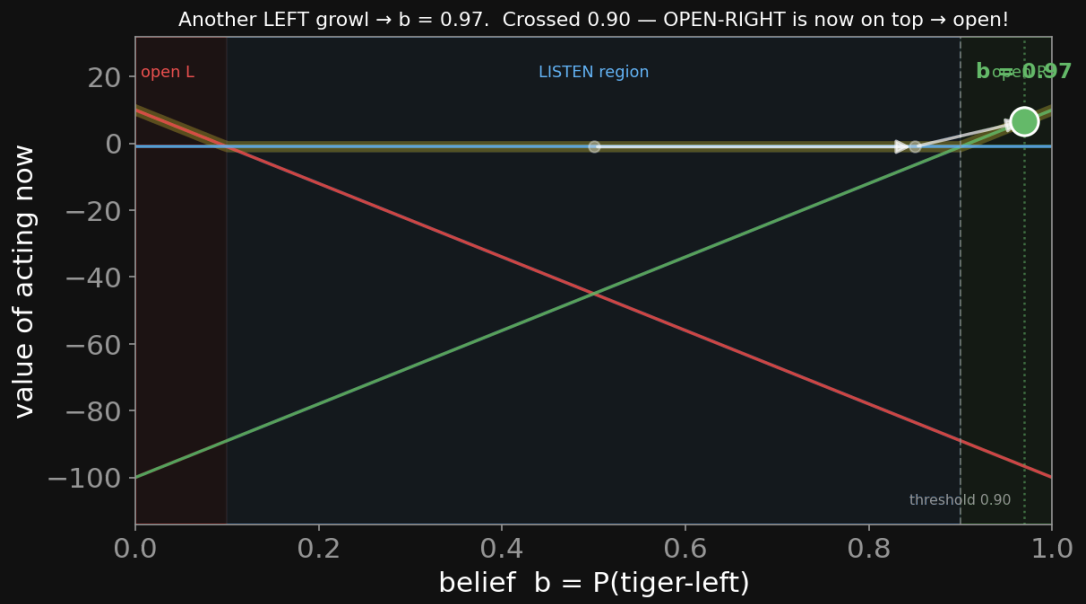
You start at **50/50**. The top line is **listen** (−1) — much better than a **−45 gamble**. So you listen, and a growl arrives.

How the decision unfolds — ② still listening



Each growl **slides the belief right**. After one “left” growl, $b = 0.85$ — still in the **listen region** (left of 0.90). So you listen again.

How the decision unfolds — ③ open!



A second agreeing growl $\rightarrow b = 0.97$, **past the 0.90 threshold**. The marker **hops up onto the open-right line** — now opening beats listening. **Open the right door.**

Poll — open, or listen again?

You’ve heard **one** growl from the left. Do you open the right door now, or listen again?

A. Open the right door now

B. Listen again

Poll — reveal

B — listen again.

Opening right wins (+10) if the tiger is left, loses (−100) if right: $E[\text{open-right}] = 10\,b + (-100)(1-b) = 110b - 100$. One growl → $b = 0.85$ → $E = \mathbf{-6.5}$, *worse* than the −1 of listening. **Two** agreeing growls → $b = 0.97$ → $E = \mathbf{+6.7}$ → open.


Break

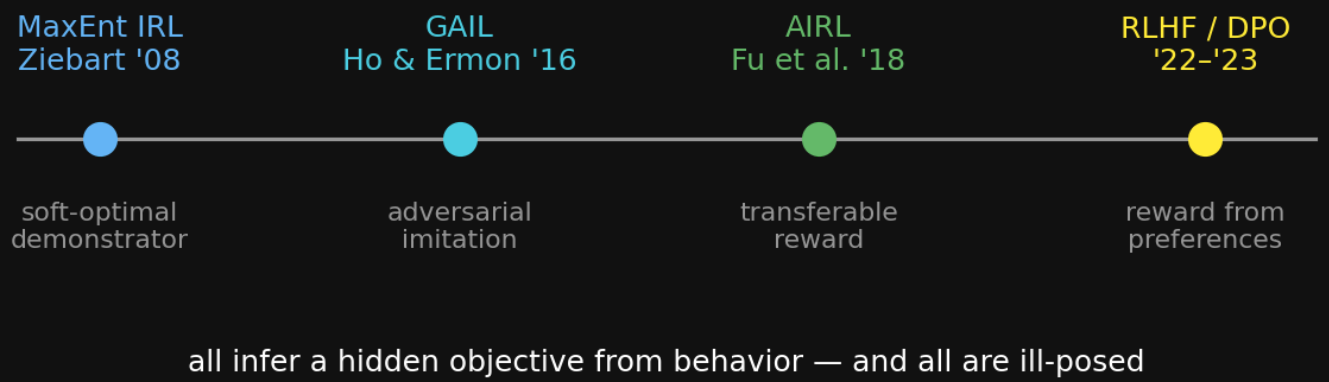
Inverse RL — recovering rewards

Where we are

Run the camera backwards	0:00
Goal inference as inverse planning	0:07
What is theory of mind?	0:37
POMDPs — inferring beliefs	0:45
Break	1:07
Inverse RL — recovering rewards	1:12
Teaching, flipped	1:26
Alignment & LLM Theory of Mind	1:38
Close & Week 10	1:52

IRL methods = learning the reward

“IRL methods” all do one thing: **recover the reward function** behind observed behavior. The catch — it’s **underconstrained**: many rewards explain the same actions. So each method adds an assumption to pin one down.

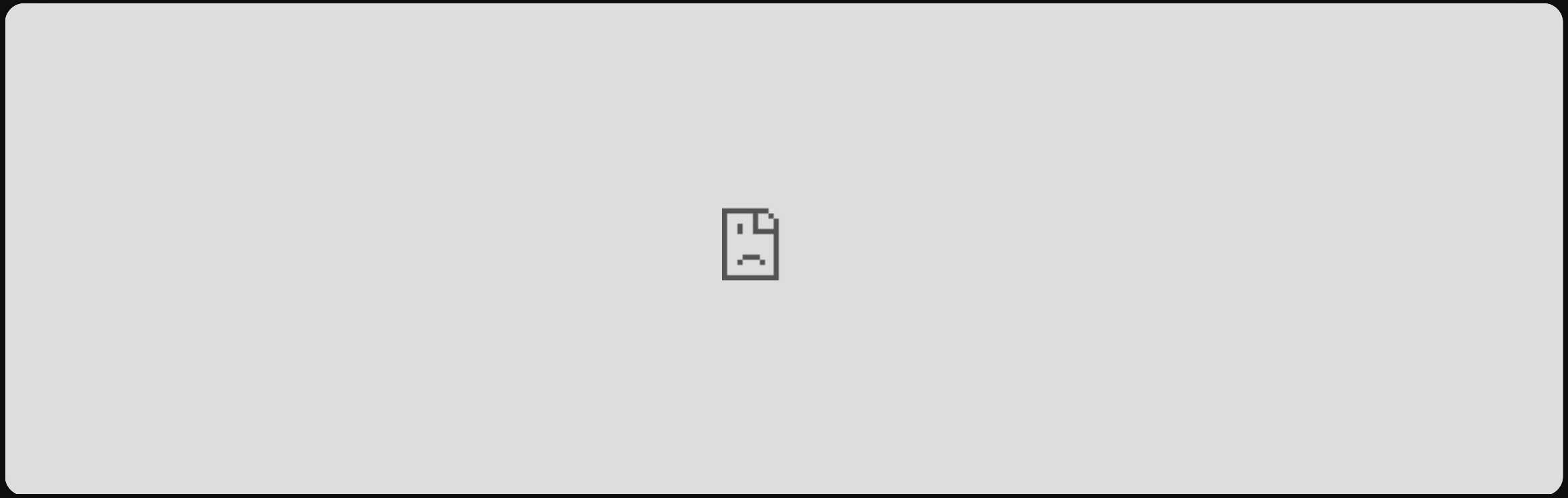


MaxEnt IRL (Ziebart 2008) — among all rewards consistent with the data, pick the one that keeps **maximum uncertainty** (highest entropy): $P(\tau) \propto e^{\text{reward}}$, a soft-optimal demonstrator (the *same* softmax as Block 2). **Why a milestone:** a principled fix for the ill-posedness — commit no more than the data forces → *one* well-defined reward by max-likelihood, and the bridge from IRL to soft-RL and RLHF.

…from imitation to transferable rewards to alignment

- **GAIL** (Ho & Ermon 2016) — *adversarial imitation*: a **critic** learns to tell the expert's behavior from the imitator's, and the imitator improves until the critic **can't tell them apart**. **Milestone:** imitation that **scales** to deep, high-dim control — but it *skips* recovering a reward, so no transferable objective. *(≈ a GAN, if you've met those.)*
- **AIRL** (Fu et al. 2018) — the same critic-vs-imitator game, but the critic is **structured so a reward falls out**, disentangled from the dynamics. **Milestone:** GAIL's scale **plus** a **transferable** reward you can re-optimize under *new* dynamics.
- **RLHF / DPO** ('22–'23) — learn the reward from human **preferences** (pairwise comparisons), not demonstrations; then optimize the policy (DPO folds the reward *into* the policy). **Milestone:** this is how **LLMs are aligned** — preference-based IRL at frontier scale.

Recover a reward — live



Paint the true reward on the right grid (+1 / 0 / −1), add demonstrations, watch the middle panel recover it — and light up neutral cells on the way (ill-posed).

Teaching = inverse planning, flipped

Where we are

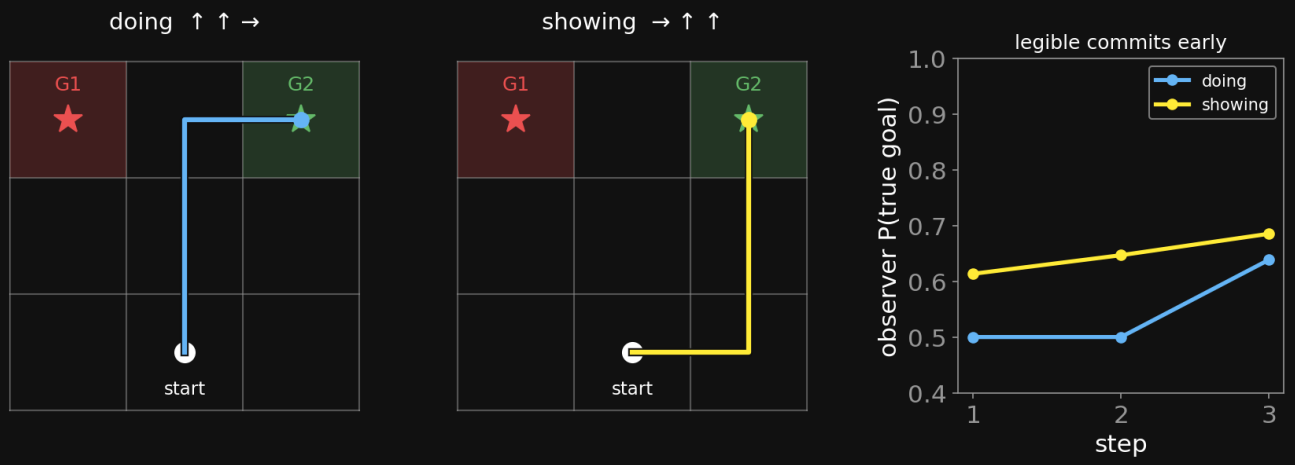
Run the camera backwards	0:00
Goal inference as inverse planning	0:07
What is theory of mind?	0:37
POMDPs — inferring beliefs	0:45
Break	1:07
Inverse RL — recovering rewards	1:12
Teaching, flipped	1:26
Alignment & LLM Theory of Mind	1:38
Close & Week 10	1:52

Flip the inversion

If an observer infers your goal from your actions, you can **choose** actions that make your goal **obvious**. That's **teaching** — and it's inverse planning run the other way.

The forward planner asks *"what should I do to reach my goal?"* The teacher asks *"what should I do so you can tell what my goal is?"* — same machinery, one more level of recursion.

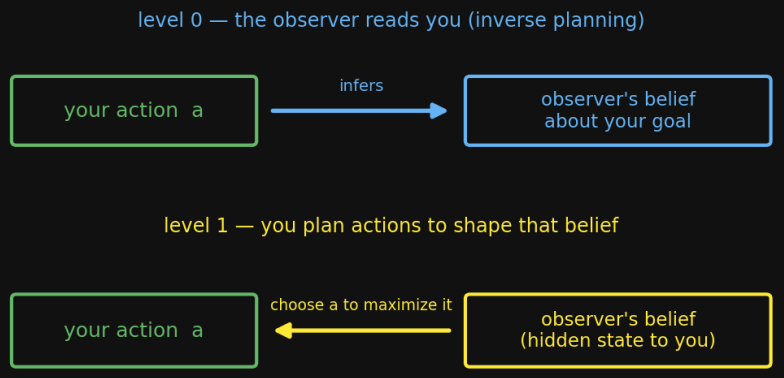
Showing vs. doing



People **doing** a task and people **showing** it move differently — demonstrators deviate from the efficient path to make the goal **unambiguous** (Ho, Littman, MacGlashan, Cushman & Austerweil 2016).

Both paths here are optimal-length, but the legible one commits **on the first move**.

Showing as recursive inference



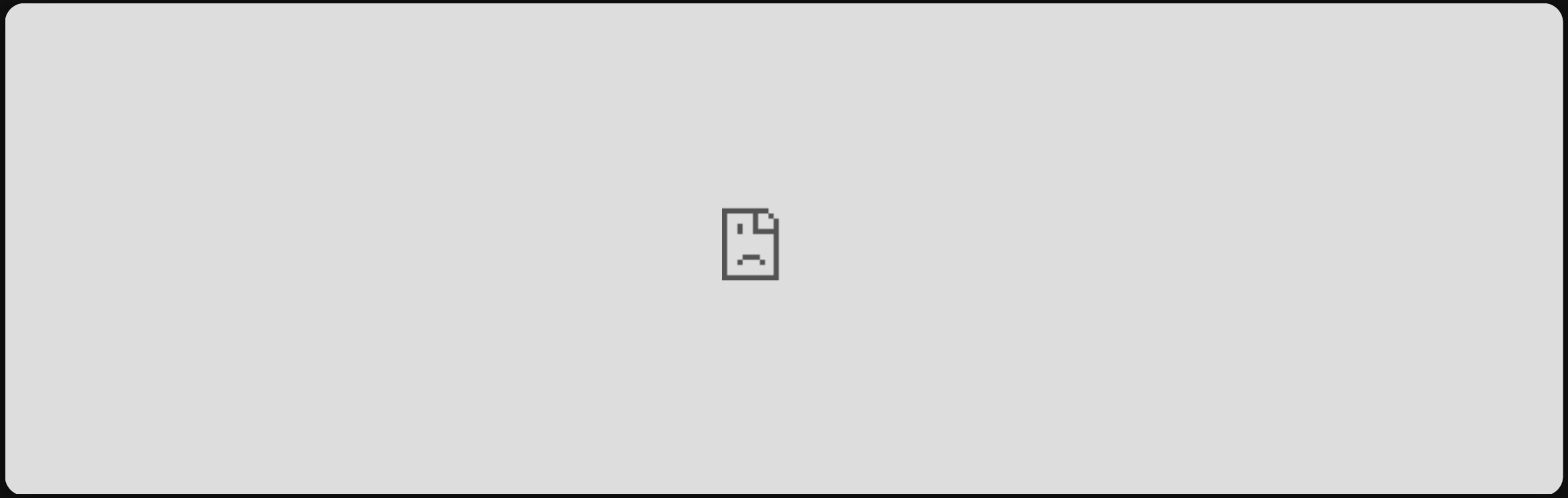
you are uncertain about the observer's belief and act to move it → a POMDP whose hidden state is the observer's belief (Ho et al. 2021)

Why does the legible path work? **Ho et al. (2021)** make it precise. Two levels:

- **the observer** inverts your actions into a belief about your goal (inverse planning);
- **you**, knowing this, pick actions to *move that belief* toward the truth.

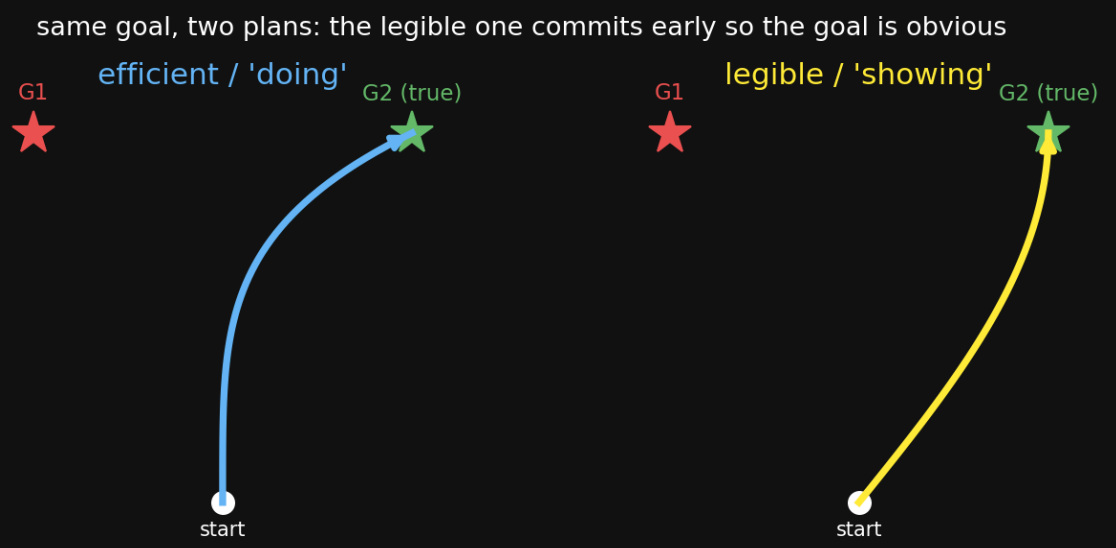
Your target — the observer's belief — is **hidden to you**, so the demonstration is **itself a POMDP**. Teaching = inverse planning, **one level up**.

Toggle legible vs. efficient — live



Toggle efficient (doing) ↔ legible (showing): the observer’s P(goal) stays ~0.50 under doing but jumps to 0.61 on the first move under showing — both paths optimal-length.

Legibility vs. predictability



Dragan, Lee & Srinivasa (2013): the **efficient** reach is ambiguous until the end; the **legible** reach veers toward the goal early. Two inferences in opposite directions — goal → trajectory vs trajectory → goal.

Reward can be a **signal**, not just a reinforcer (Ho et al. 2015 — the GardenPath paper).

Alignment as a teaching game

- **Cooperative IRL** (Hadfield-Menell et al. 2016): human + robot, both rewarded by the *human's* reward — but only the human knows it.
- **Efficient expert demonstration is provably suboptimal** — you should *deviate* to teach.
- It **reduces to a POMDP** whose hidden state is the human's reward.
- *The unifying frame: every beat today is a POMDP over which-MDP-we're-in.*

Poll — show the exit

You want to show a friend which of two exits to take. Do you walk the **shortest** path, or **exaggerate** toward it?

A. Walk the shortest, most efficient path

B. Exaggerate toward the exit

Poll — reveal

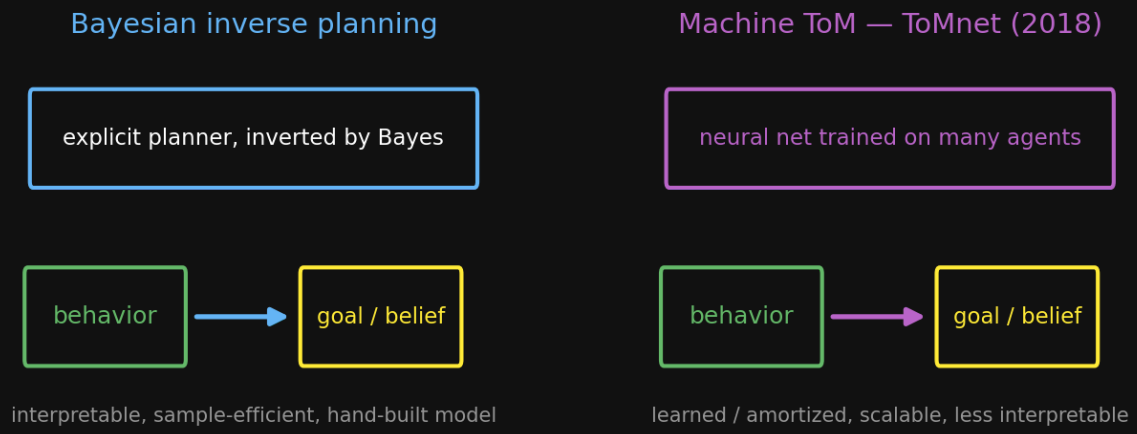
B — exaggerate (be legible).

The efficient path is ambiguous early; the legible path makes the observer’s posterior commit sooner — exactly what the widget showed (61% vs 50% on the first move).

Alignment & LLM Theory of Mind

Where we are

Two ways to read a mind



ToMnet (Rabinowitz et al. 2018) — a neural net watches many agents and **learns** to predict their behavior and false beliefs. The **amortized** cousin (learned in one forward pass) of Baker's explicit **Bayesian** inversion: same inverse problem, learned vs. hand-built.

RLHF & DPO — a reward from preferences

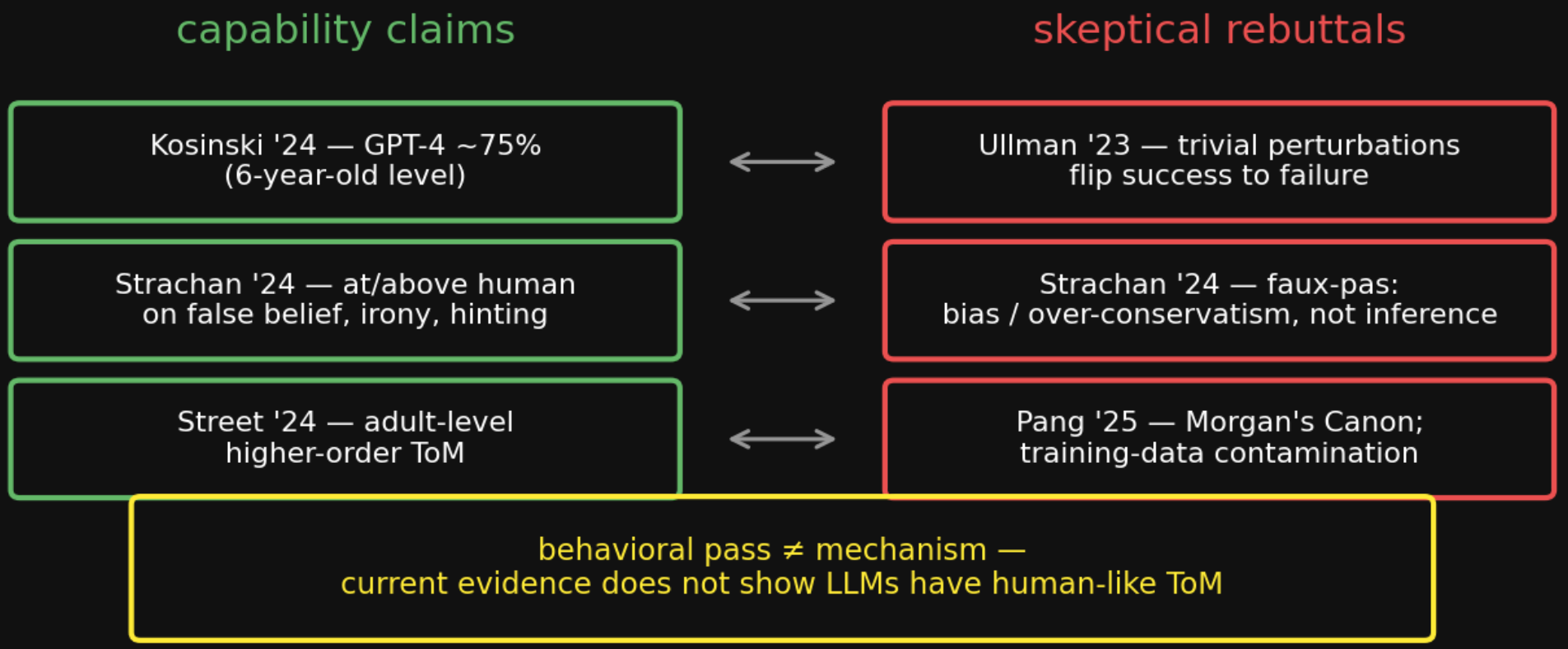
- **RLHF** (RL from Human Feedback): don't hand-write a reward — show people **pairs** of outputs, record **which they prefer**, fit a **reward model** to the preferences, then RL the policy against it.
- **DPO** (Direct Preference Optimization): skip the separate reward model — optimize the policy **directly** on the preferences (the reward is *implicit* in the policy).
- The choice model is **Bradley–Terry**: the chance you prefer A is a **logistic** function of the reward gap, $P(A \succ B) = \sigma(r_A - r_B)$ — σ is the S-curve squashing any real number into $(0, 1)$. Fitting r from human **choices** is exactly **preference-based inverse RL**.

…recovering that reward in GenJAX

```
@gen
def prefs(pairs):                                          # the Bradle
    r = jnp.array([normal(0., 2.) @ f"r_{k}" for k in range(K)])
    for n, (i, j) in enumerate(pairs):
        flip(sigmoid(r[i] - r[j])) @ f"pref_{n}"
```

Sampled data — 90 preferences from a hidden quality (A=+2, B=+0.5, C=−1): `A>B, A>C, B>C, A>B, C>B, A>C, …` **Inference** — condition on them, recover the reward: **A=+1.36, B=0.00, C=−1.36 → A > B > C ✓** (only up to an additive constant — IRL's ill-posedness).

Do LLMs have a Theory of Mind?



What the tests actually look like

Standard Sally-Anne — *Anne hides a toy in a box, leaves, the toy is moved.* A strong model reports Anne's **false** belief correctly. ✓ *(Kosinski 2024: GPT-4 ~75%.)*

Ullman's perturbation (2023) — make the box **transparent** so Anne can see in. A belief-tracker says "Anne now knows," yet models often **still report the old false belief**. ✗

Faux pas — needs the nested structure from before: the speaker's false belief **+** the listener's knowledge. Models pass some batteries, but the wins may be **response bias**, not the inference (Strachan 2024).

The autism lesson, live for machines: **behavioral pass ≠ mechanism**.

The honest answer: contested

- **Pass** — many false-belief, irony, and hinting vignettes, at adult level.
- **Fail** — minimal, *belief-preserving* perturbations flip success to failure (Ullman 2023): make the container *transparent* so the agent can see in, yet the model still reports the false belief.
- **Confounds** — also stumble on faux pas; the wins may be response bias or training-data contamination (Pang 2025).
- **A behavioral pass is not the mechanism.**

Defensible 2026 position: current evidence does not show LLMs have human-like Theory of Mind.

Poll — does GPT-4 have ToM?

GPT-4 passes ~75% of false-belief vignettes — a 6-year-old's level. Does it **have** a theory of mind?

A. Yes — it passes the tests

B. No — passing tests isn't having the mechanism

C. We can't tell yet

Poll — reveal

B / C — behavioral pass ≠ mechanism.

This is the recurring epistemics question of the course — “fits the judgments” and “this is the computation it runs” are different claims. The honest read of the current evidence is skeptical.

Where we landed

One model, every piece inferred

Reward R — goal inference, invert the MDP

Theory of mind — defined (false belief), then = inverse planning

Belief b — invert a POMDP (Sally-Anne, Tiger, food-truck)

Reward at scale — MaxEnt → adversarial imitation → RLHF

Teaching — steer the observer's belief (Ho 2021; CIRL)

Alignment — RLHF as IRL; LLM theory of mind is contested

Next week

We've inferred *discrete* hidden causes — goals, beliefs, rewards. **Next: Bayesian nonparametrics** — inferring *how many* causes there are when you don't know in advance.

⚠️ **Next week (Jul 3) is async** — no in-person class. Watch the recording I post; Slack discussion runs as normal. (Optional AI Safety workshop that week — in English.)

Reminders: project proposal due **Sun Jun 28, 8 PM**; **3 reflection chances left** (Weeks 10–12); RL + Monte Carlo both due **Jul 10**.