Sleeping Newcomb

The 'Newcomb Tension' in Games with Self-Locating Uncertainty

John W.E. Cremin
Aix-Marseille University, CNRS, AMSE

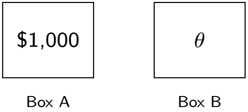
Stony Brook International Conference on Game Theory 2026

22 July 2026

Learning with Temporal Ignorance

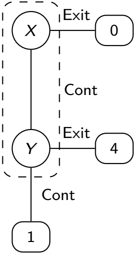
- I am broadly interested in Bayesian learning, and modelling this in the presence of 'temporal ignorance'.
  - Agents arrive in a game at time t and observe some signal whose distribution depends on this arrival time, but do not perfectly observe it.
  - My working papers *Sleeping Beauty Learns to Fish* and *Bot Got Your Tongue?* both consider settings with an element of this.
- This temporal ignorance is an instance of *self-locating uncertainty* (SLU): uncertainty as to which node one occupies *within* a given information set and play of a game.
- I study a dynamic inconsistency in such games that I call a *Newcomb tension*, and when it can (and cannot) be resolved.

Three Puzzles



$\theta \in \{\$0,\ \$10^{10}\}$

I. Newcomb



II. AMD

	H	T
$n{=}1,\ d{=}1$	•	•
$n{=}1/n{=}2,\ d{=}2$		•

III. Sleeping Beauty

Different paradoxes, same structure: locally rational reasoning at the point of decision disagrees with the planning optimum. I call this a *Newcomb tension*, and I will organise the talk around these three puzzles.

Preview of Results

- The paper defines the Newcomb tension and studies when it arises, and when it can be resolved, in (single information-set) games with SLU.

The main results are:

1. **Representation:** an uncommitted agent with 'one-boxer beliefs' acts exactly as a Bayesian with commitment power; I characterise exactly when such beliefs exist.
2. **Single agent:** randomisation always resolves the tension (this is not a coincidental property of the AMD).
3. **Multiple agents:** the tension can survive randomisation, and where it does, even the belief representation can fail: commitment becomes irreplaceable. The planner in Duplicating Sleeping Beauty bets like a halfer even though the correct credences are thirder.

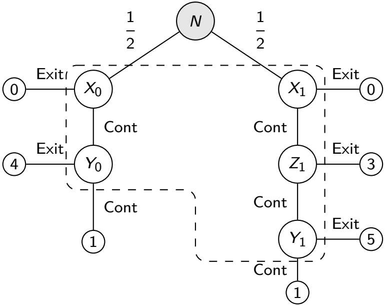
The Model

- **Game:** single information set h; agents N; prior ρ on Θ; for each $(\theta, \mathbf{a})$, $D(\theta, \mathbf{a})$ = the *dots* (nodes of h) visited. Behavioural $\sigma \in \Delta(A(h))$: independent randomisation at each dot.
- **Planning-optimal:** $\sigma^* = \arg\max_{\sigma} \sum_{\theta, \mathbf{a}} \rho(\theta) \Pr_{\sigma}(\mathbf{a})\, U(\mathbf{a}, \theta)$.
- A **halfer** (*Z-consistent*) updates on 'I am at h' using the *unconditional* probability of reaching h: $\pi^{\mathrm{H}}(\theta, \mathbf{a}) = \rho(\theta) \Pr_{\sigma}(\mathbf{a} \mid h \text{ reached})$.
- **Thirder** (*consistent assessment*) beliefs re-weight by the awakening count $|D(\theta, \mathbf{a})|$; halfers do not.
- **Interim fixed point:** $\hat{\sigma}$ is interim-optimal under the beliefs it induces: each temporal part's best response to $\hat{\sigma}$ is $\hat{\sigma}$.

Newcomb tension

Γ exhibits a *Newcomb tension* if $\sigma^* \neq \hat{\sigma}$.

Halfer vs Thirder: Sleeping Beauty Behind the Wheel



	$\theta=0$			$\theta=1$			
	E	CE	CC	E	CE	CCE	CCC
d=1	•	•	•	•	•	•	•
d=2		•	•		•	•	•
d=3						•	•
H	$\frac{1-q}{2}$	$\frac{q(1-q)}{2}$	$\frac{q^2}{2}$	$\frac{1-q}{2}$	$\frac{q(1-q)}{2}$	$\frac{q^2(1-q)}{2}$	$\frac{q^3}{2}$
T	$\frac{1-q}{2C}$	$\frac{q(1-q)}{C}$	$\frac{q^2}{C}$	$\frac{1-q}{2C}$	$\frac{q(1-q)}{C}$	$\frac{3q^2(1-q)}{2C}$	$\frac{3q^3}{2C}$

Eqm σ: Cont. w. $q \in (0,1)$.

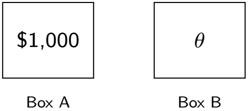
Halfer (H): $\pi = \rho(\theta)\,\Pr_\sigma(\mathbf{a})$.

Thirder (T): $\pi \propto \rho(\theta)\,\Pr_\sigma(\mathbf{a})|D(\theta,\mathbf{a})|$.

Normalising constant: $C = 1 + q + \dfrac{q^2}{2}$.

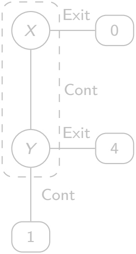
I. Newcomb's Problem

Three Puzzles



$\theta \in \{\$0,\ \$10^{10}\}$

I. Newcomb



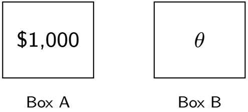
II. AMD

	H	T
$n{=}1,\ d{=}1$	•	•
$n{=}1/n{=}2,\ d{=}2$		•

III. Sleeping Beauty

Different paradoxes, same structure: locally rational reasoning at the point of decision disagrees with the planning optimum. I call this a *Newcomb tension*, and I will organise the talk around these three puzzles.

Three Puzzles



$$\theta \in \{\$0,\ \$10^{10}\}$$

I. Newcomb

- Nozick's Newcomb problem: a near-perfect predictor fills box B iff it predicts you take only box B. Two-boxing dominates; one-boxers get rich. (Also: strictly speaking ill-posed; a perfect predictor is incompatible with free randomisation.)
- **Newcomb's game**: add a planning stage h_0 before nature.

  $$h_0 \longrightarrow \text{predictor sets } \theta(h_0) \longrightarrow h$$

  Commit-to-One-box $\Rightarrow \theta =$ Full, payoff M; otherwise $\theta =$ Empty.

Newcomb tension (informal)

The planning-optimal and interim-optimal actions diverge: $a^{\text{plan}} =$ One-box, but at h Two-box dominates, so $a^{\text{int}} =$ Two-box. The *value of commitment* is $M - 1 > 0$.

One-Boxer Beliefs

Definition

A distribution P^{OB} over the state–profile–dot triples reached under σ^* gives *one-boxer beliefs for* σ^* if an interim EU-maximiser holding P^{OB} finds σ^* optimal.

- Whatever beliefs make the uncommitted agent behave as if committed.
- **Non-uniqueness (dot-independent payoffs).** If $v_d(a', \mathbf{a}_{-d}, \theta) = v(a', \theta)$, then $P^{\mathrm{OB}}(\theta) = \rho(\theta)$ with *any* conditional over $(\mathbf{a}, d)$ suffices; in particular halfer beliefs do the job.
- **Halfer need not work.** In the AMD with $\sigma = \mathrm{Cont}$, $D = \{X, Y\}$; one-boxer beliefs require $P^{\mathrm{OB}}(X) \geq \frac{3}{4}$, not $\frac{1}{2}$.

Representation Theorem, with Existence

Theorem (Representation with Existence)

(i) One-boxer beliefs for σ^* exist **iff** no mixture of actions beats σ^* in occupant continuation payoff at *every* dot reached under σ^*.

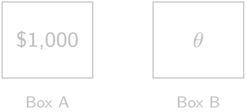
(ii) They always exist when σ^* is a behavioural planning optimum and the game is single-agent, payoff-independent (with exogenous dots or symmetric counts), or identical-payoff.

(iii) Whenever they exist: a Bayesian thirder *with* commitment power and an uncommitted EU-maximiser holding P^{OB} are behaviourally equivalent.

- So commitment can usually be modelled as a belief, with no explicit commitment technology.

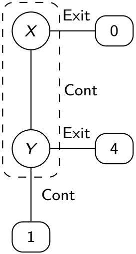
II. The Absent-Minded Driver: One Agent

Three Puzzles



$\theta \in \{\$0,\ \$10^{10}\}$

I. Newcomb



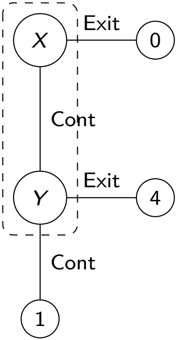
II. AMD

	H	T
$n{=}1,\ d{=}1$	•	•
$n{=}1/n{=}2,\ d{=}2$		•

III. Sleeping Beauty

Puzzle II: one agent, self-locating within a single play. Does allowing randomisation resolve the tension?

The Absent-Minded Driver (Piccione and Rubinstein, 1997)



- One drunk driver; intersections X and Y are indistinguishable (*absent-mindedness*); the same (behavioural) strategy applies at both.
- Pure strategies: $a^{\text{plan}} = \text{Cont}$ $(1 > 0)$. But at the wheel, with $P(X) = P(Y) = \frac{1}{2}$: Exit pays $2 > 1$.
- Each pure strategy makes the *other* interim-optimal: a Newcomb tension, and no pure fixed point at all.

Randomisation Resolves the Tension

Theorem (Randomisation Resolution)

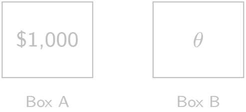
In any single-agent, single-information-set SLU game, the planning-optimal behavioural strategy σ^* is also interim-optimal. This holds for any $|\Theta| \geq 1$.

- In the AMD: $V^{\mathrm{plan}}(q) = 4q - 3q^2$, so $\sigma^* = \frac{2}{3}$; and $\frac{2}{3}$ is exactly the interim fixed point under thirder beliefs.
- Generalises Piccione and Rubinstein (1997): this is no accident of one state and two dots. Any number of states, any dot-dependence of payoffs.
- Two birds with one stone: randomisation *creates* the fixed point (none exists in pure strategies) and *aligns* it with the planning optimum.

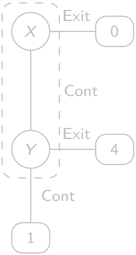
Proof idea

III. Duplicating Sleeping Beauty: Many Agents

Three Puzzles



$\theta \in \{\$0,\ \$10^{10}\}$



	H	T
$n{=}1,\ d{=}1$	•	•
$n{=}1/n{=}2,\ d{=}2$		•

I. Newcomb **II. AMD** **III. Sleeping Beauty**

Puzzle III: several agents sharing an information set. Here the tension can survive randomisation, and even belief representation can fail.

Three Puzzles

- Coin toss $\theta \in \{H, T\}$. Original wakes Monday in both states; *if Tails, a clone is created and wakes Tuesday*. Awakenings indistinguishable; each reports a credence $x \in [0, 1]$ in Tails, scored by quadratic loss.
- What should she *report*? A social choice problem behind the veil of ignorance (Harsanyi, 1955): balance the interests of original and clone.
- Assuming co-cardinal utilities, here the tension results from a shift in the social weights assigned to the planning and interim stages.

	H	T
$n=1,\ d=1$	•	•
$n=1/n=2,\ d=2$		•

III. Sleeping Beauty

Duplicating Sleeping Beauty

- Awakenings per agent:

	$\theta = H$	$\theta = T$
Original	1	1
Clone	0	1

- **Planner** (compound states, uniform over dots): weights $\frac{1}{2}, \frac{1}{4}, \frac{1}{4}$ on the three awakenings; by agent, $\lambda_{\text{orig}} = \frac{3}{4}$, $\lambda_{\text{clone}} = \frac{1}{4}$.
- **Awakened thirder**: $\frac{1}{3}$ each; $\lambda^{\text{int}}_{\text{orig}} = \frac{2}{3}$, $\lambda^{\text{int}}_{\text{clone}} = \frac{1}{3}$.

The Tension: Halfer Bets, Thirder Credences

Proposition (Newcomb Tension in Duplicating Sleeping Beauty)

With quadratic loss $S(x, \theta) = -(x - \mathbf{1}\{\theta = T\})^2$, the planning-optimal report is the **halfer credence** and the interim-optimal report the **thirder credence**:

$$V^{\mathrm{plan}}(x) = \frac{1}{2}S(x, H) + \frac{1}{2}S(x, T) \quad \Rightarrow \; x^{\mathrm{plan}} = \frac{1}{2},$$
$$V^{\mathrm{int}}(x) = \frac{1}{3}S(x, H) + \frac{2}{3}S(x, T) \quad \Rightarrow \; x^{\mathrm{int}} = \frac{2}{3}.$$

Randomisation over reports does not resolve the tension.

- Thirder credences are (I maintain) *correct*; but the right *bet* ex ante is the halfer's. Kierland and Monton (2005)'s halfer recommendation for this problem is exactly my planning stage; the disagreement is confined to credence.

When Does the Multi-Agent Tension Arise?

Theorem (Multi-Agent Newcomb Tension)

(i) **No tension** if the game is *payoff-independent* (each agent's payoff depends only on actions at their own dots) *and* dot counts are symmetric across agents.

(ii) **Generic tension** if payoffs are *interdependent*, or dot counts are asymmetric across agents.

- Two separate channels: the thirder re-weights compound states by dot counts (asymmetry channel), and the occupant ignores externalities the planner internalises (interdependence channel).
- Dup SB: payoff-independent but count-asymmetric. Duplicating AMD: both.

Duplicating AMD

Conclusion

Related Literature

- **SLU in game theory:** Piccione and Rubinstein (1997), Aumann et al. (1997), Grove and Halpern (1997) and the GEB special issue.
- **SLU in philosophy:** Elga (2000), Lewis (2001), Kierland and Monton (2005), Ross (2010), Janda (2024), Spohn (2025), Schwarz (2015).
- **Dynamic consistency:** Strotz (1955), Machina (1989); closest apparatus: Vergopoulos (2011)'s two-state-space framework.

Summary

- A **Newcomb tension** arises in SLU games when planning-optimal and interim-optimal strategies diverge.
- **Representation:** committed thirder = uncommitted one-boxer, exactly when no mixture dominates the plan dot-wise.
- **Single agent:** randomisation always resolves the tension (and creates the fixed point where pure strategies have none).
- **Multiple agents:** interdependence or asymmetric awakenings revive the tension; it survives randomisation, and can even defeat every belief. Commitment becomes irreplaceable.

Proof Idea

Sketch of the single-agent argument:

- Thirder belief under σ^*: $P(\theta, d) = \rho(\theta)\sigma_d(\theta, \sigma^*)/C(\sigma^*)$.
- First-order change in V^{plan} toward action a':

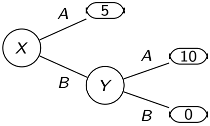
$$\left.\frac{dV^{\text{plan}}}{d\varepsilon}\right|_{\varepsilon=0} = C(\sigma^*) \sum_{\theta,d} P(\theta, d)[v_d(a', \sigma^*_{-d}, \theta) - v_d(\sigma^*, \sigma^*_{-d}, \theta)].$$

- This is $C(\sigma^*)$ times the *interim expected advantage* of a'.
- At a planning optimum the interim advantage is non-positive in every direction, so σ^* is interim-optimal.

Back

Non-existence of One-Boxer Beliefs

Two single-agent, pure-strategy failures (behavioural strategies rescue both):



Pure optimum A truncates Y: δ_B dominates at the only reached dot; no beliefs on reached dots.

	$a_2=0$	$a_2=1$	$a_2=2$
$a_1=0$	10	20	5
$a_1=1$	20	0	30
$a_1=2$	5	30	0

Both dots always reached; pure optimum 0, but action 1 dominates it at either dot: *no* belief, however supported, rationalises it.

- The criterion (theorem of the alternative): beliefs exist iff no dot-wise dominating mixture on reached dots. Behavioural planning optima escape both traps for a single agent; interdependence does not.

Back

AMD: the Halfer's Fixed Point

Thirder. $P^T(X) = \dfrac{1}{1+q}$, $P^T(Y) = \dfrac{q}{1+q}$; interim FOC

$$\frac{4-6q}{1+q} = 0 \;\Rightarrow\; \hat{q}_T = \frac{2}{3} = \sigma^*.$$

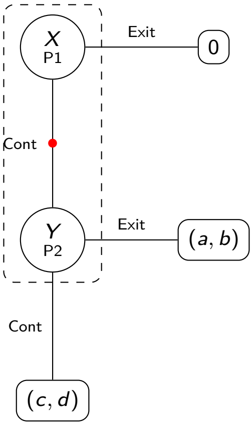
Halfer. $P^H(X) = 1 - \dfrac{q}{2}$, $P^H(Y) = \dfrac{q}{2}$; interim FOC

$$(1-\frac{q}{2})(4-3q)-\frac{3q}{2} = 0 \Rightarrow 3q^2-13q+8 = 0 \Rightarrow \hat{q}_H = \frac{13-\sqrt{73}}{6} \approx 0.743.$$

- The halfer's fixed point is *not* the planning optimum: halfer beliefs are not innocuous in dot-dependent games.

Back

Duplicating Absent-Minded Driver



- P1 occupies X; if P1 continues, the clone (P2) is created (red dot) and occupies Y.
- $D(n = 1) = \{X\}$,
  $D(n = 2) = \{Y\}$.
- Exit at X pre-dates the clone: scalar payoff to P1 only.
- The dot structure depends on the action profile, and payoffs are *interdependent*: P1's payoff depends on the action at P2's dot.

◂ Back

Version (a): Identical Payoffs, No Tension

Set $(a, b, c, d) = (4, 4, 1, 1)$.

- Planning payoff: $V^{\text{plan}}(\sigma) = 4\sigma - 3\sigma^2$, so $\sigma^* = 2/3$.
- Interim FOC: $\frac{4-6\sigma}{1+\sigma} = 0 \;\Rightarrow\; \hat{\sigma} = 2/3$.
- Proportionality: $dV^{\text{plan}}/d\sigma = 4 - 6\sigma = C(\sigma) \cdot \text{FOC}^{\text{int}}$.
- With $u_1 = u_2$, compound-state welfare equals the common payoff: the problem reduces to a single-agent optimisation and the single-agent theorem applies.
- **Duplication alone does not create the tension.**

Duplicating Sleeping Beauty Behind the Wheel

Add a clone to each state of the two-state AMD (created at the first Continue):

- $\theta = 0$: $D(\mathrm{P1}) = \{X_0\}$, $D(\mathrm{P2}) = \{Y_0\}$.
- $\theta = 1$: $D(\mathrm{P1}) = \{X_1\}$, $D(\mathrm{P2}) = \{Z_1, Y_1\}$.
- With $u_1 = u_2$ at every terminal, V^{plan} is the same as in the non-duplicating game:

$$\sigma^* = \frac{-1 + \sqrt{85}}{12} \approx 0.685.$$

- Interim: the per-dot FOC sums to the same polynomial, so $\hat{\sigma} = \sigma^*$.
- **Equal payoffs $\Rightarrow$ no tension, even with asymmetric dots across agents.**

Back