Sleeping Newcomb

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

John W.E. Cremin
Aix-Marseille University, CNRS, AMSE
France-Spain Theory Meeting 2026

April 27, 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.

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*, as in the *Sleeping Beauty Problem* (wait two slides!)

Self-Locating Uncertainty

- In a game-theoretic setting, one can think of self-locating uncertainty as resulting from two scenarios we normally rule out:

1. Absent-mindedness.
2. Multiple agents sharing the same information set.

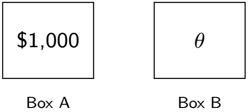
Self-Locating Uncertainty

- In a game-theoretic setting, one can think of self-locating uncertainty as resulting from two scenarios we normally rule out:

1. Absent-mindedness.
2. Multiple agents sharing the same information set.

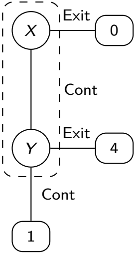
- Today I won't address how (and whether) one should choose to model such games in the first place: look out for *Sleeping Beauty's Dismal Day Out* for this (Yes... I like puns...)
- Instead I discuss a common feature of such games I call a *Newcomb Tension*

Three Puzzles



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

Newcomb's Problem



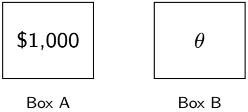
AMD

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

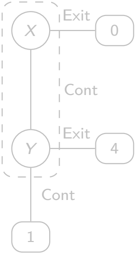
Sleeping Beauty

Different paradoxes, same structure: locally rational reasoning at the information set disagrees with the planning optimum: a *Newcomb Tension*.

Three Puzzles



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



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

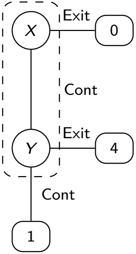
Newcomb's Problem **AMD** **Sleeping Beauty**

Different paradoxes, same structure: locally rational reasoning at the information set disagrees with the planning optimum: a *Newcomb Tension*.

Three Puzzles

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

Newcomb's Problem



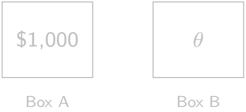
AMD

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

Sleeping Beauty

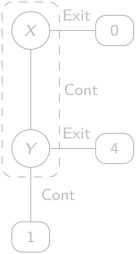
Different paradoxes, same structure: locally rational reasoning at the information set disagrees with the planning optimum: a *Newcomb Tension*.

Three Puzzles



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

Newcomb's Problem



AMD

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

Sleeping Beauty

Different paradoxes, same structure: locally rational reasoning at the information set disagrees with the planning optimum: a *Newcomb Tension*.

Preview of Results

- The point of this paper is to define this 'Newcomb Tension' and study when it will arise in (single info-set) games with SLU.

Preview of Results

- The point of this paper is to define this 'Newcomb Tension' and study when it will arise in (single info-set) games with SLU.

The main results are:

1. We can always find a 'one-boxer beliefs' representation such that an agent with these beliefs acts as if they were a Bayesian with commitment power.

Preview of Results

- The point of this paper is to define this ‘Newcomb Tension’ and study when it will arise in (single info-set) games with SLU.

The main results are:

1. We can always find a ‘one-boxer beliefs’ representation such that an agent with these beliefs acts as if they were a Bayesian with commitment power.
2. In single-agent games, randomisation always resolves this tension (this is not just a coincidental property of the AMD).

Preview of Results

- The point of this paper is to define this 'Newcomb Tension' and study when it will arise in (single info-set) games with SLU.

The main results are:

1. We can always find a 'one-boxer beliefs' representation such that an agent with these beliefs acts as if they were a Bayesian with commitment power.
2. In single-agent games, randomisation always resolves this tension (this is not just a coincidental property of the AMD).
3. In multi-agent games this is not the case (relevant to the duplicating Sleeping Beauty debate).

Model

Single-Information-Set SLU Games

A *single-information-set SLU game* Γ has:

- A finite set of agents N, prior ρ on states Θ, action set $A(h)$.
- A *single* information set $h \subseteq X$, with agent assignment $\iota : h \to N$.
- Co-cardinal vNM utilities $u_n : Z \to \mathbb{R}$.
- No perfect recall: an agent may occupy multiple nodes in h on the same play.
- For each state θ and action profile $\mathbf{a} : h \to A(h)$, let $D(\theta, \mathbf{a})$ denote the *dots* (nodes in h) visited.
- I assume $|D(\theta, \mathbf{a})|$ is deterministic given $(\theta, \mathbf{a})$.

More on the model

Thirder Beliefs

Definition (Thirder beliefs)

Under strategy σ, the interim agent at h assigns

$$\pi(\theta, \mathbf{a}) \;=\; \frac{\rho(\theta)\,\Pr_\sigma(\mathbf{a})|D(\theta, \mathbf{a})|}{\sum_{\theta', \mathbf{a}'} \rho(\theta')\,\Pr_\sigma(\mathbf{a}')|D(\theta', \mathbf{a}')|},$$

and distributes probability uniformly over dots within each $(\theta, \mathbf{a})$.

Thirder Beliefs

Definition (Thirder beliefs)

Under strategy σ, the interim agent at h assigns

$$\pi(\theta, \mathbf{a}) = \frac{\rho(\theta)\,\Pr_\sigma(\mathbf{a})|D(\theta, \mathbf{a})|}{\sum_{\theta', \mathbf{a}'} \rho(\theta')\,\Pr_\sigma(\mathbf{a}')|D(\theta', \mathbf{a}')|},$$

and distributes probability uniformly over dots within each $(\theta, \mathbf{a})$.

- Awakening counts scale the prior; follows Ross's Generalised Thirder Principle, except assuming an equilibrium.

Thirder Beliefs

Definition (Thirder beliefs)

Under strategy σ, the interim agent at h assigns

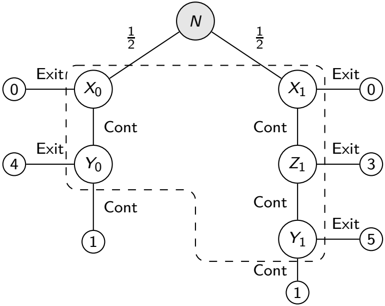
$$\pi(\theta, \mathbf{a}) = \frac{\rho(\theta)\,\Pr_\sigma(\mathbf{a})|D(\theta, \mathbf{a})|}{\sum_{\theta', \mathbf{a}'} \rho(\theta')\,\Pr_\sigma(\mathbf{a}')|D(\theta', \mathbf{a}')|},$$

and distributes probability uniformly over dots within each $(\theta, \mathbf{a})$.

- Awakening counts scale the prior; follows Ross's Generalised Thirder Principle, except assuming an equilibrium.
- When $|D(\theta, \mathbf{a})|$ is constant in $(\theta, \mathbf{a})$, thirder and halfer beliefs coincide.

Halfer beliefs

Halfer vs Thirder: Sleeping Beauty Behind the Wheel



AMD calculation

	$\theta=0$			$\theta=1$			
	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: $C = 1 + q + \frac{q^2}{2}$.

Planning-Optimal and Interim-Optimal

- Behavioural strategy $\sigma \in \Delta(A(h))$: independent randomisation at each dot.

Planning-Optimal and Interim-Optimal

- Behavioural strategy $\sigma \in \Delta(A(h))$: independent randomisation at each dot.

Planning-optimal

$$\sigma^* \ = \ \arg\max_{\sigma} V^{\text{plan}}(\sigma) = \arg\max_{\sigma} \sum_{\theta,\mathbf{a}} \rho(\theta) \Pr_{\sigma}(\mathbf{a}) U(\mathbf{a}, \theta).$$

Planning-Optimal and Interim-Optimal

- Behavioural strategy $\sigma \in \Delta(A(h))$: independent randomisation at each dot.

Planning-optimal

$$\sigma^* = \arg\max_{\sigma} V^{\text{plan}}(\sigma) = \arg\max_{\sigma} \sum_{\theta, \mathbf{a}} \rho(\theta) \Pr_{\sigma}(\mathbf{a}) U(\mathbf{a}, \theta).$$

Interim-optimal (fixed point)

$\hat{\sigma}$ is a behavioural strategy such that, when all temporal parts play $\hat{\sigma}$, each finds $\hat{\sigma}$ optimal given thirder beliefs.

Planning-Optimal and Interim-Optimal

- Behavioural strategy $\sigma \in \Delta(A(h))$: independent randomisation at each dot.

Planning-optimal

$$\sigma^* \;=\; \arg\max_{\sigma} V^{\text{plan}}(\sigma) = \arg\max_{\sigma} \sum_{\theta,\mathbf{a}} \rho(\theta) \Pr_{\sigma}(\mathbf{a}) U(\mathbf{a}, \theta).$$

Interim-optimal (fixed point)

$\hat{\sigma}$ is a behavioural strategy such that, when all temporal parts play $\hat{\sigma}$, each finds $\hat{\sigma}$ optimal given thirder beliefs.

Newcomb Tension

Γ exhibits a *Newcomb tension* if $\sigma^* \neq \hat{\sigma}$.
The *value of commitment* is $V^{\text{plan}}(\sigma^*) - V^{\text{plan}}(\hat{\sigma}) > 0$.

One-Boxer Representation

One-Boxer Beliefs

Definition

A joint distribution P^{OB} over $(\theta, \mathbf{a}, d)$ gives *one-boxer beliefs for* σ^* if an interim EU-maximiser holding P^{OB} finds σ^* optimal.

One-Boxer Beliefs

Definition

A joint distribution P^{OB} over $(\theta, \mathbf{a}, d)$ gives *one-boxer beliefs for* σ^* if an interim EU-maximiser holding P^{OB} finds σ^* optimal.

- **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 always do the job in this setting.

One-Boxer Beliefs

Definition

A joint distribution P^{OB} over $(\theta, \mathbf{a}, d)$ gives *one-boxer beliefs for* σ^* if an interim EU-maximiser holding P^{OB} finds σ^* optimal.

- **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 always do the job in this setting.
- **Halfer need not work.** In the AMD w. $\sigma = \mathrm{Cont}$, $D = \{X, Y\}$; one-boxer beliefs require $P^{\mathrm{OB}}(X) \geq \frac{3}{4}$, not $\frac{1}{2}$.

SBBtW example

Representation Theorem

Theorem (Representation)

A Bayesian thirder with commitment power is behaviourally equivalent to an uncommitted expected-utility maximiser with one-boxer beliefs P^{OB}.

Representation Theorem

Theorem (Representation)

A Bayesian thirder with commitment power is behaviourally equivalent to an uncommitted expected-utility maximiser with one-boxer beliefs P^{OB}.

- **Committed thirder**: plays σ^* because bound.

Representation Theorem

Theorem (Representation)

A Bayesian thirder with commitment power is behaviourally equivalent to an uncommitted expected-utility maximiser with one-boxer beliefs P^{OB}.

- **Committed thirder**: plays σ^* because bound.
- **Uncommitted one-boxer**: plays σ^* because beliefs make it interim-optimal.

Representation Theorem

Theorem (Representation)

A Bayesian thirder with commitment power is behaviourally equivalent to an uncommitted expected-utility maximiser with one-boxer beliefs P^{OB}.

- **Committed thirder**: plays σ^* because bound.
- **Uncommitted one-boxer**: plays σ^* because beliefs make it interim-optimal.
- In any game with a Newcomb tension, we can model an agent with 'commitment power' as having one-boxer beliefs.
- In some social learning games (e.g. BGYT, SBLTF), this amounts to assuming all agents are halfers.

Single-Agent Results

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 $|\Theta| \geq 1$.

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 $|\Theta| \geq 1$.

- Generalises Piccione–Rubinstein: not only does $q^* = 2/3$ work in the single-state AMD, but a planning-optimal σ^* *always* coincides with the interim fixed point in single-agent games.

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 $|\Theta| \geq 1$.

- Generalises Piccione–Rubinstein: not only does $q^* = 2/3$ work in the single-state AMD, but a planning-optimal σ^* *always* coincides with the interim fixed point in single-agent games.
- Two birds with one stone: randomisation creates the fixed point *and* aligns it with the planning optimum.

Intuition / pure-strategy FPs

Multi-Agent Games

Multi-Agent Social Welfare

Let $D(\theta, n, \mathbf{a})$ = dots occupied by agent n in compound state $(\theta, \mathbf{a})$. Under utilitarian SWF:

$$V^{\text{plan}}(\sigma) = \sum_{\theta,\mathbf{a}} \rho(\theta) \Pr_{\sigma}(\mathbf{a}) \sum_{n \in N} \frac{|D(\theta, n, \mathbf{a})|}{|D(\theta, \mathbf{a})|} U_n(\mathbf{a}, \theta).$$

Multi-Agent Social Welfare

Let $D(\theta, n, \mathbf{a})$ = dots occupied by agent n in compound state $(\theta, \mathbf{a})$. Under utilitarian SWF:

$$V^{\mathrm{plan}}(\sigma) = \sum_{\theta,\mathbf{a}} \rho(\theta) \Pr_{\sigma}(\mathbf{a}) \sum_{n \in N} \frac{|D(\theta, n, \mathbf{a})|}{|D(\theta, \mathbf{a})|} U_n(\mathbf{a}, \theta).$$

- Planner's social weight of agent n:

$$\lambda_n^{\mathrm{plan}}(\sigma) = \sum_{\theta,\mathbf{a}} \rho(\theta) \Pr_{\sigma}(\mathbf{a}) \frac{|D(\theta, n, \mathbf{a})|}{|D(\theta, \mathbf{a})|}.$$

Multi-Agent Social Welfare

Let $D(\theta, n, \mathbf{a})$ = dots occupied by agent n in compound state $(\theta, \mathbf{a})$. Under utilitarian SWF:

$$V^{\text{plan}}(\sigma) = \sum_{\theta,\mathbf{a}} \rho(\theta) \Pr_{\sigma}(\mathbf{a}) \sum_{n \in N} \frac{|D(\theta, n, \mathbf{a})|}{|D(\theta, \mathbf{a})|} U_n(\mathbf{a}, \theta).$$

- Planner's social weight of agent n:

$$\lambda_n^{\text{plan}}(\sigma) = \sum_{\theta,\mathbf{a}} \rho(\theta) \Pr_{\sigma}(\mathbf{a}) \frac{|D(\theta, n, \mathbf{a})|}{|D(\theta, \mathbf{a})|}.$$

- Interim agent: same formula but with $\pi(\theta, \mathbf{a})$ (thirder) in place of $\rho(\theta) \Pr_{\sigma}(\mathbf{a})$.

Multi-Agent Newcomb Tension

Theorem

(i) **Symmetric dot structure $\Rightarrow$ no tension.** If $|D(\theta, n, \mathbf{a})|$ is the same across agents n for every $(\theta, \mathbf{a})$, then per-dot continuation reasoning recovers σ^*.

Multi-Agent Newcomb Tension

Theorem

(i) **Symmetric dot structure $\Rightarrow$ no tension.** If $|D(\theta, n, \mathbf{a})|$ is the same across agents n for every $(\theta, \mathbf{a})$, then per-dot continuation reasoning recovers σ^*.

(ii) **Asymmetric dot structure $\Rightarrow$ generic tension.** If $|D(\theta_0, n_0, \mathbf{a}_0)| \neq |D(\theta_0, n_1, \mathbf{a}_0)|$ for some $n_0, n_1, \theta_0, \mathbf{a}_0$, then generically $\sigma^* \neq \hat{\sigma}$.

Duplicating Sleeping Beauty: The Weights

- Original exists in both states; clone only in Tails.

Duplicating Sleeping Beauty: The Weights

- Original exists in both states; clone only in Tails.
- Number of awakenings per agent in $(\theta, \mathbf{a})$:

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

Duplicating Sleeping Beauty: The Weights

- Original exists in both states; clone only in Tails.
- Number of awakenings per agent in $(\theta, \mathbf{a})$:

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

- **Planner weights** ($\rho \cdot$ dot share): $\lambda_{\text{orig}} = 3/4,\ \lambda_{\text{clone}} = 1/4.$

Duplicating Sleeping Beauty: The Weights

- Original exists in both states; clone only in Tails.
- Number of awakenings per agent in $(\theta, \mathbf{a})$:

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

- **Planner weights** ($\rho \cdot$ dot share): $\lambda_{\text{orig}} = 3/4$, $\lambda_{\text{clone}} = 1/4$.
- **Interim (thirder) weights**: $\lambda^{\text{int}}_{\text{orig}} = 2/3$, $\lambda^{\text{int}}_{\text{clone}} = 1/3$.

Duplicating Sleeping Beauty: The Tension

Brier score (or any strictly proper rule) $S(x, \theta)$:

Duplicating Sleeping Beauty: The Tension

Brier score (or any strictly proper rule) $S(x, \theta)$:

- Planning-optimal report maximises

$$V^{\mathrm{plan}}(x) = \tfrac{3}{8}\, S(x, H) + \tfrac{5}{8}\, S(x, T).$$

Duplicating Sleeping Beauty: The Tension

Brier score (or any strictly proper rule) $S(x, \theta)$:

- Planning-optimal report maximises

$$V^{\text{plan}}(x) = \tfrac{3}{8}\, S(x, H) + \tfrac{5}{8}\, S(x, T).$$

- Interim-optimal report maximises

$$V^{\text{int}}(x) = \tfrac{1}{3}\, S(x, H) + \tfrac{2}{3}\, S(x, T).$$

Duplicating Sleeping Beauty: The Tension

Brier score (or any strictly proper rule) $S(x, \theta)$:

- Planning-optimal report maximises

$$V^{\text{plan}}(x) = \tfrac{3}{8}\, S(x, H) + \tfrac{5}{8}\, S(x, T).$$

- Interim-optimal report maximises

$$V^{\text{int}}(x) = \tfrac{1}{3}\, S(x, H) + \tfrac{2}{3}\, S(x, T).$$

- $\frac{3}{8} \neq \frac{1}{3} \;\Rightarrow\; x^{\text{plan}} \neq x^{\text{int}}$.

Duplicating Sleeping Beauty: The Tension

Brier score (or any strictly proper rule) $S(x, \theta)$:

- Planning-optimal report maximises

$$V^{\text{plan}}(x) = \tfrac{3}{8}\, S(x, H) + \tfrac{5}{8}\, S(x, T).$$

- Interim-optimal report maximises

$$V^{\text{int}}(x) = \tfrac{1}{3}\, S(x, H) + \tfrac{2}{3}\, S(x, T).$$

- $\frac{3}{8} \neq \frac{1}{3} \;\Rightarrow\; x^{\text{plan}} \neq x^{\text{int}}.$

Randomisation is powerless

Each agent has at most one payoff-relevant dot. Mixing over reports is weakly dominated; the tension persists.

Duplicating AMD applications

Conclusion

Related Literature

- **SLU in Game Theory:** Principally Piccione and Rubinstein (1997) and Aumann et al. (1997); entire special issue of course related.

Related Literature

- **SLU in Game Theory:** Principally Piccione and Rubinstein (1997) and Aumann et al. (1997); entire special issue of course related.
- **SLU in Philosophy:** Kierland and Monton (2005), Ross (2010), Janda (2024), Spohn (2025), Schwarz (2015).
  - Elga (2000), Lewis (2001), Winkler (2017).

Summary

- A **Newcomb tension** arises in SLU games when planning-optimal and interim-optimal strategies diverge.

Summary

- A **Newcomb tension** arises in SLU games when planning-optimal and interim-optimal strategies diverge.
- **Representation Theorem:** committed thirder = uncommitted one-boxer.

Summary

- A **Newcomb tension** arises in SLU games when planning-optimal and interim-optimal strategies diverge.
- **Representation Theorem:** committed thirder = uncommitted one-boxer.
- **Single-agent:** randomisation always resolves the tension (and creates the fixed point where pure strategies fail).

Summary

- A **Newcomb tension** arises in SLU games when planning-optimal and interim-optimal strategies diverge.
- **Representation Theorem:** committed thirder = uncommitted one-boxer.
- **Single-agent:** randomisation always resolves the tension (and creates the fixed point where pure strategies fail).
- **Multi-agent:** the tension requires asymmetric dot structure; otherwise it generically vanishes.

Summary

- A **Newcomb tension** arises in SLU games when planning-optimal and interim-optimal strategies diverge.
- **Representation Theorem:** committed thirder = uncommitted one-boxer.
- **Single-agent:** randomisation always resolves the tension (and creates the fixed point where pure strategies fail).
- **Multi-agent:** the tension requires asymmetric dot structure; otherwise it generically vanishes.
- **Applications:** Duplicating Sleeping Beauty (tension that survives randomisation), Duplicating AMD (Persistent Tension Proposition).

Intuition

Sketch of the argument:

- Thirder belief under σ^*: $P(\theta, d) = \rho(\theta)\sigma_d(\theta, \sigma^*)/C(\sigma^*)$.

Intuition

Sketch of the 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':

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

Intuition

Sketch of the 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'.

Intuition

Sketch of the 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 $\Rightarrow \sigma^*$ is interim-optimal.

Pure-Strategy Fixed Points Are Planning-Optimal

Proposition

In any single-agent, single-information-set SLU game, if a pure strategy a is an interim fixed point then $a = a^{\mathrm{plan}}$ among pure strategies.

Pure-Strategy Fixed Points Are Planning-Optimal

Proposition

In any single-agent, single-information-set SLU game, if a pure strategy a is an interim fixed point then $a = a^{\text{plan}}$ among pure strategies.

- When a pure-strategy interim fixed point *exists*, it is already planning-optimal.

Pure-Strategy Fixed Points Are Planning-Optimal

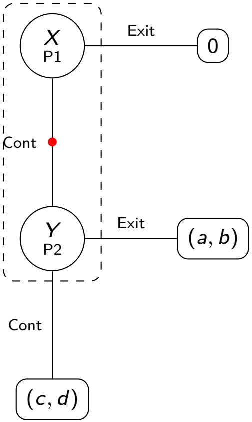
Proposition

In any single-agent, single-information-set SLU game, if a pure strategy a is an interim fixed point then $a = a^{\mathrm{plan}}$ among pure strategies.

- When a pure-strategy interim fixed point *exists*, it is already planning-optimal.
- Corollary: randomisation only *matters* when no pure fixed point exists (as in AMD). In such cases, it simultaneously delivers existence and planning-optimality.

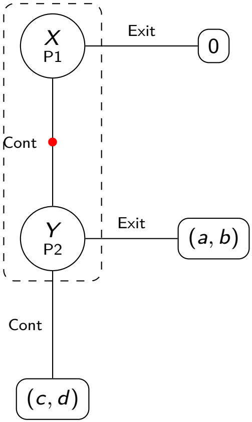
Back

Duplicating Absent-Minded Driver



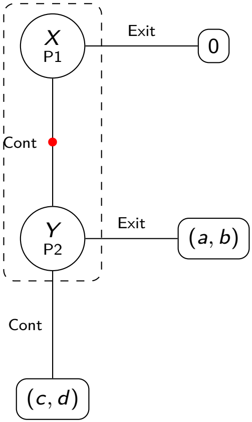
- P1 occupies X; if P1 continues, the clone (P2) is created (red dot) and occupies Y.

Duplicating Absent-Minded Driver



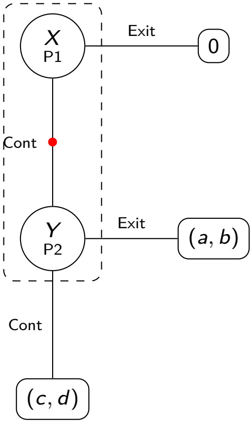
- P1 occupies X; if P1 continues, the clone (P2) is created (red dot) and occupies Y.
- $D(n = 1) = \{X\}$,
  $D(n = 2) = \{Y\}$.

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.

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.

Version (a): Identical Payoffs — No Tension

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

Version (a): Identical Payoffs — No Tension

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

- Planning payoff: $V^{\mathrm{plan}}(\sigma) = 4\sigma - 3\sigma^2$, so $\sigma^* = 2/3$.

Version (a): Identical Payoffs — No Tension

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

- Planning payoff: $V^{\mathrm{plan}}(\sigma) = 4\sigma - 3\sigma^2$, so $\sigma^* = 2/3$.
- Interim FOC: $\frac{4-6\sigma}{1+\sigma} = 0 \;\Rightarrow\; \hat{\sigma} = 2/3$.

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: $dV^{\text{plan}}/d\sigma = 4 - 6\sigma = C(\sigma) \cdot \text{FOC}^{\text{int}}$.

Version (a): Identical Payoffs — No Tension

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

- Planning payoff: $V^{\mathrm{plan}}(\sigma) = 4\sigma - 3\sigma^2$, so $\sigma^* = 2/3$.
- Interim FOC: $\frac{4-6\sigma}{1+\sigma} = 0 \;\Rightarrow\; \hat{\sigma} = 2/3$.
- Proportionality: $dV^{\mathrm{plan}}/d\sigma = 4 - 6\sigma = C(\sigma) \cdot \mathrm{FOC}^{\mathrm{int}}$.

- With $u_1 = u_2$, the compound-state welfare just equals the common payoff $\Rightarrow$ reduces to a single-agent optimisation $\Rightarrow$ the single-agent theorem applies.
- **Duplication alone does not create the tension.**

Version (b): Persistent Tension

Now Exit at Y gives (a, b), Continue past Y gives (c, d), with $a + b > c + d$.

Version (b): Persistent Tension

Now Exit at Y gives (a, b), Continue past Y gives (c, d), with $a + b > c + d$.

$$\sigma^* = \frac{a+b}{2(a+b-c-d)}, \qquad \hat{\sigma} = \frac{a}{a+b-c-d}.$$

Version (b): Persistent Tension

Now Exit at Y gives (a, b), Continue past Y gives (c, d), with $a + b > c + d$.

$$\sigma^* = \frac{a+b}{2(a+b-c-d)}, \qquad \hat{\sigma} = \frac{a}{a+b-c-d}.$$

Proposition (Persistent Tension)

The Newcomb tension is present if and only if $b \neq a$, with

$$\sigma^* - \hat{\sigma} = \frac{b-a}{2(a+b-c-d)}.$$

Version (b): Persistent Tension

Now Exit at Y gives (a, b), Continue past Y gives (c, d), with $a + b > c + d$.

$$\sigma^* = \frac{a+b}{2(a+b-c-d)}, \qquad \hat{\sigma} = \frac{a}{a+b-c-d}.$$

Proposition (Persistent Tension)

The Newcomb tension is present if and only if $b \neq a$, with

$$\sigma^* - \hat{\sigma} = \frac{b-a}{2(a+b-c-d)}.$$

- Wedge driven by the Exit-at-Y payoffs only: if $b > a$ the planner wants more continuation than the interim agent.

Version (b): Persistent Tension

Now Exit at Y gives (a, b), Continue past Y gives (c, d), with $a + b > c + d$.

$$\sigma^* = \frac{a+b}{2(a+b-c-d)}, \qquad \hat{\sigma} = \frac{a}{a+b-c-d}.$$

Proposition (Persistent Tension)

The Newcomb tension is present if and only if $b \neq a$, with

$$\sigma^* - \hat{\sigma} = \frac{b-a}{2(a+b-c-d)}.$$

- Wedge driven by the Exit-at-Y payoffs only: if $b > a$ the planner wants more continuation than the interim agent.
- Randomisation does **not** help.

Back

Duplicating Sleeping Beauty Behind the Wheel

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

- $\theta = 0$: $D(\text{P1}) = \{X_0\}$, $D(\text{P2}) = \{Y_0\}$.
- $\theta = 1$: $D(\text{P1}) = \{X_1\}$, $D(\text{P2}) = \{Z_1, Y_1\}$.

Duplicating Sleeping Beauty Behind the Wheel

Add a clone to each state of the two-state AMD (clone 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:

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

Duplicating Sleeping Beauty Behind the Wheel

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

- $\theta = 0$: $D(\text{P1}) = \{X_0\}$, $D(\text{P2}) = \{Y_0\}$.
- $\theta = 1$: $D(\text{P1}) = \{X_1\}$, $D(\text{P2}) = \{Z_1, Y_1\}$.

- 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 $\hat{\sigma} = \sigma^*$.

Duplicating Sleeping Beauty Behind the Wheel

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

- $\theta = 0$: $D(\text{P1}) = \{X_0\}$, $D(\text{P2}) = \{Y_0\}$.
- $\theta = 1$: $D(\text{P1}) = \{X_1\}$, $D(\text{P2}) = \{Z_1, Y_1\}$.

- 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

Two Types of Self-Locating Uncertainty

Two nodes $x, x' \in h$ are *co-reachable* if some play visits both.

Two Types of Self-Locating Uncertainty

Two nodes $x, x' \in h$ are *co-reachable* if some play visits both.

Intra-personal SLU (Absent-mindedness)

There exist co-reachable $x, x' \in h$ with $\iota(x) = \iota(x')$. A single agent visits h more than once on the same play.
Example: Absent-Minded Driver.

Two Types of Self-Locating Uncertainty

Two nodes $x, x' \in h$ are *co-reachable* if some play visits both.

Intra-personal SLU (Absent-mindedness)

There exist co-reachable $x, x' \in h$ with $\iota(x) = \iota(x')$. A single agent visits h more than once on the same play.
Example: Absent-Minded Driver.

Inter-personal SLU

There exist co-reachable $x, x' \in h$ with $\iota(x) \neq \iota(x')$. Different agents share the information set on the same play.
Example: Duplicating Sleeping Beauty.

Two Types of Self-Locating Uncertainty

Two nodes $x, x' \in h$ are *co-reachable* if some play visits both.

Intra-personal SLU (Absent-mindedness)

There exist co-reachable $x, x' \in h$ with $\iota(x) = \iota(x')$. A single agent visits h more than once on the same play.
Example: Absent-Minded Driver.

Inter-personal SLU

There exist co-reachable $x, x' \in h$ with $\iota(x) \neq \iota(x')$. Different agents share the information set on the same play.
Example: Duplicating Sleeping Beauty.

- Standard imperfect information — distinct branches, not co-reachable — is **not** SLU.

Planning stage

Example: Sleeping Beauty Behind the Wheel (No Randomisation)

- Restrict to pure strategies. Planning-optimal pure strategy is *Continue* (payoff 1 in both states, vs. 0 for Exit) — just like in AMD.

Example: Sleeping Beauty Behind the Wheel (No Randomisation)

- Restrict to pure strategies. Planning-optimal pure strategy is *Continue* (payoff 1 in both states, vs. 0 for Exit) — just like in AMD.
- Per-dot gain from continuing (with all other dots also continuing):

$$g_{X_0} = 1, \quad g_{Y_0} = -3, \quad g_{X_1} = 1, \quad g_{Z_1} = -2, \quad g_{Y_1} = -4.$$

Example: Sleeping Beauty Behind the Wheel (No Randomisation)

- Restrict to pure strategies. Planning-optimal pure strategy is *Continue* (payoff 1 in both states, vs. 0 for Exit) — just like in AMD.
- Per-dot gain from continuing (with all other dots also continuing):

$$g_{X_0} = 1, \quad g_{Y_0} = -3, \quad g_{X_1} = 1, \quad g_{Z_1} = -2, \quad g_{Y_1} = -4.$$

- P^{OB} is one-boxer iff $\sum_d P^{\mathrm{OB}}(d)\, g_d \geq 0$, i.e. $P^{\mathrm{OB}}(X_0) + P^{\mathrm{OB}}(X_1) \ \geq \ 3\, P^{\mathrm{OB}}(Y_0) + 2\, P^{\mathrm{OB}}(Z_1) + 4\, P^{\mathrm{OB}}(Y_1)$.

Example: Sleeping Beauty Behind the Wheel (No Randomisation)

- Restrict to pure strategies. Planning-optimal pure strategy is *Continue* (payoff 1 in both states, vs. 0 for Exit) — just like in AMD.
- Per-dot gain from continuing (with all other dots also continuing):

$$g_{X_0} = 1, \quad g_{Y_0} = -3, \quad g_{X_1} = 1, \quad g_{Z_1} = -2, \quad g_{Y_1} = -4.$$

- P^{OB} is one-boxer iff $\sum_d P^{\mathrm{OB}}(d)\, g_d \geq 0$, i.e.
  $P^{\mathrm{OB}}(X_0) + P^{\mathrm{OB}}(X_1) \ \geq \ 3\, P^{\mathrm{OB}}(Y_0) + 2\, P^{\mathrm{OB}}(Z_1) + 4\, P^{\mathrm{OB}}(Y_1).$
- Any distribution putting *enough* weight on X_0, X_1 does the job. Putting all weight there is sufficient; scattering weight to the continuation nodes is permitted up to this bound.

Halfer Beliefs

- A halfer updates on the event 'I am at *h*' using the *unconditional* probability of reaching *h* in each state-action profile:

$$\pi^{\mathrm{H}}(\theta, \mathbf{a}) \;=\; \rho(\theta) \Pr_{\sigma}(\mathbf{a} \mid h \text{ reached}).$$

Halfer Beliefs

- A halfer updates on the event 'I am at h' using the *unconditional* probability of reaching h in each state-action profile:

$$\pi^{\mathrm{H}}(\theta, \mathbf{a}) \;=\; \rho(\theta)\, \Pr_{\sigma}(\mathbf{a} \mid h \text{ reached}).$$

- In general, reaching h is more likely in some $(\theta, \mathbf{a})$ tuples than others, so halfer beliefs shift away from the prior.

Halfer Beliefs

- A halfer updates on the event 'I am at *h*' using the *unconditional* probability of reaching *h* in each state-action profile:

$$\pi^{\mathrm{H}}(\theta, \mathbf{a}) \;=\; \rho(\theta) \Pr_{\sigma}(\mathbf{a} \mid h \text{ reached}).$$

- In general, reaching *h* is more likely in some $(\theta, \mathbf{a})$ tuples than others, so halfer beliefs shift away from the prior.
- But in a *single*-information-set game, *h* is reached with probability 1 in every $(\theta, \mathbf{a})$ (at least 1 awakening occurs).

Halfer Beliefs

- A halfer updates on the event 'I am at *h*' using the *unconditional* probability of reaching *h* in each state-action profile:

$$\pi^{\mathrm{H}}(\theta, \mathbf{a}) \;=\; \rho(\theta)\ \Pr_{\sigma}(\mathbf{a} \mid h \text{ reached}).$$

- In general, reaching *h* is more likely in some $(\theta, \mathbf{a})$ tuples than others, so halfer beliefs shift away from the prior.
- But in a *single*-information-set game, *h* is reached with probability 1 in every $(\theta, \mathbf{a})$ (at least 1 awakening occurs).
- So conditioning on 'I am at *h*' is vacuous: the halfer's beliefs coincide with the unconditional joint $\rho(\theta)\,\Pr_{\sigma}(\mathbf{a})$.

Halfer Beliefs

- A halfer updates on the event 'I am at *h*' using the *unconditional* probability of reaching *h* in each state-action profile:

$$\pi^{\mathrm{H}}(\theta, \mathbf{a}) \;=\; \rho(\theta)\, \Pr_{\sigma}(\mathbf{a} \mid h \text{ reached}).$$

- In general, reaching *h* is more likely in some $(\theta, \mathbf{a})$ tuples than others, so halfer beliefs shift away from the prior.
- But in a *single*-information-set game, *h* is reached with probability 1 in every $(\theta, \mathbf{a})$ (at least 1 awakening occurs).
- So conditioning on 'I am at *h*' is vacuous: the halfer's beliefs coincide with the unconditional joint $\rho(\theta) \Pr_\sigma(\mathbf{a})$.
- Here, the halfer's beliefs are just the unconditional probabilities over $(\theta, \mathbf{a})$.

Back

The Planning Stage

Timing of the game:

$$h_0 \longrightarrow \text{Nature draws } \theta \longrightarrow h.$$

The Planning Stage

Timing of the game:

$$h_0 \longrightarrow \text{Nature draws } \theta \longrightarrow h.$$

- At h_0, either $A(h_0) = \{\text{Commit}, \text{Not Commit}\}$, or $A(h_0) = \varnothing$.
- Commit = bind to a strategy at h before θ is realised.

The Planning Stage

Timing of the game:

$$h_0 \longrightarrow \text{Nature draws } \theta \longrightarrow h.$$

- At h_0, either $A(h_0) = \{\text{Commit}, \text{Not Commit}\}$, or $A(h_0) = \varnothing$.
- Commit = bind to a strategy at h before θ is realised.
- Except in my representation of Newcomb's problem, θ is drawn *independently* of h_0.

The Planning Stage

Timing of the game:

$$h_0 \longrightarrow \text{Nature draws } \theta \longrightarrow h.$$

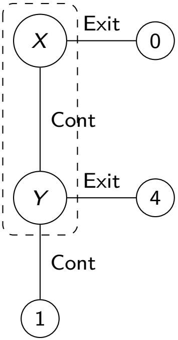
- At h_0, either $A(h_0) = \{\text{Commit}, \text{Not Commit}\}$, or $A(h_0) = \varnothing$.
- Commit = bind to a strategy at h before θ is realised.
- Except in my representation of Newcomb's problem, θ is drawn *independently* of h_0.

Newcomb's Problem as an SLU Game

$A(h_0) = \{\text{Commit}, \text{Not Commit}\}$, $A(h) = \{\text{One-box}, \text{Two-box}\}$,
$\theta \in \{\text{Full}, \text{Empty}\}$ *chosen by the predictor as a function of* h_0.
Then $a^{\text{plan}} = \text{One-box} \neq \text{Two-box} = a^{\text{int}}$.

Back

AMD: Setup



Under σ: Cont. w. $q \in (0, 1)$:

Profile	Prob	Dots
E	$1 - q$	$\{X\}$
CE	$q(1 - q)$	$\{X, Y\}$
CC	q^2	$\{X, Y\}$

Per-dot gain from Continue (vs. Exit) with others on σ:

$$g_X = 4 - 3q, \qquad g_Y = -3.$$

AMD: Planning and Thirder

Planning. $V^{\text{plan}}(q) = q(1-q)\cdot 4 + q^2 \cdot 1 = 4q - 3q^2$, so

$$\sigma^* = \frac{2}{3}.$$

AMD: Planning and Thirder

Planning. $V^{\text{plan}}(q) = q(1-q)\cdot 4 + q^2 \cdot 1 = 4q - 3q^2$, so

$$\sigma^* = \tfrac{2}{3}.$$

Thirder. Compound-state weights $\propto \rho(\theta)\,\mathrm{Pr}_\sigma(\mathbf{a})|D(\theta,\mathbf{a})|$; normalising constant $C(q) = 1 + q$. Uniform within compound states gives

$$P^T(X) = \tfrac{1}{1+q}, \qquad P^T(Y) = \tfrac{q}{1+q}.$$

Interim FOC:

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

AMD: Halfer

Halfer. Compound-state weights $\propto \rho(\theta)\,\mathrm{Pr}_\sigma(\mathbf{a})$ (no dot-count scaling). Uniform within compound states:

$$P^H(X) = 1 - \tfrac{q}{2}, \qquad P^H(Y) = \tfrac{q}{2}.$$

AMD: Halfer

Halfer. Compound-state weights $\propto \rho(\theta)\,\mathrm{Pr}_\sigma(\mathbf{a})$ (no dot-count scaling). Uniform within compound states:

$$P^H(X) = 1 - \tfrac{q}{2}, \qquad P^H(Y) = \tfrac{q}{2}.$$

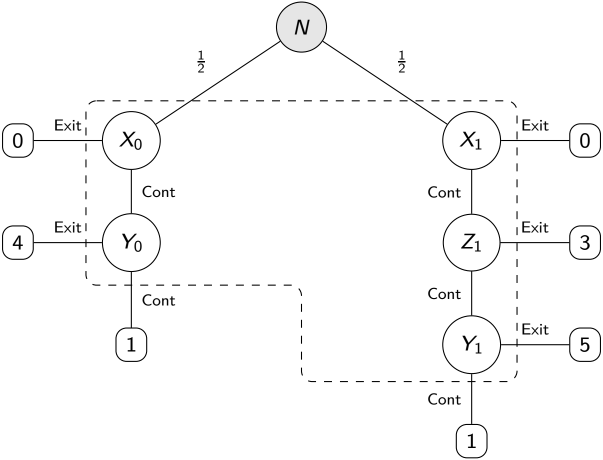
Interim FOC:

$$(1 - \tfrac{q}{2})(4 - 3q) - \tfrac{3q}{2} = 0$$

$$\Rightarrow\ 3q^2 - 13q + 8 = 0\ \Rightarrow\ \hat{q}_H = \tfrac{13-\sqrt{73}}{6} \approx 0.743.$$

Back

Sleeping Beauty Behind the Wheel



- $\theta = 0$: standard AMD. $\theta = 1$: an extra intersection Z_1 inflates the dot count.

SB Behind the Wheel: Planning = Interim

State payoffs under σ:

$$V_0(\sigma) = 4\sigma - 3\sigma^2,$$
$$V_1(\sigma) = 3\sigma + 2\sigma^2 - 4\sigma^3.$$

SB Behind the Wheel: Planning = Interim

State payoffs under σ:

$$V_0(\sigma) = 4\sigma - 3\sigma^2,$$
$$V_1(\sigma) = 3\sigma + 2\sigma^2 - 4\sigma^3.$$

Planning payoff:

$$V^{\text{plan}}(\sigma) = \tfrac{7\sigma - \sigma^2 - 4\sigma^3}{2}, \qquad \sigma^* = \frac{-1 + \sqrt{85}}{12} \approx 0.685.$$

SB Behind the Wheel: Planning = Interim

State payoffs under σ:

$$V_0(\sigma) = 4\sigma - 3\sigma^2,$$
$$V_1(\sigma) = 3\sigma + 2\sigma^2 - 4\sigma^3.$$

Planning payoff:

$$V^{\mathrm{plan}}(\sigma) = \tfrac{7\sigma - \sigma^2 - 4\sigma^3}{2}, \qquad \sigma^* = \frac{-1 + \sqrt{85}}{12} \approx 0.685.$$

Interim first-order condition gives the same quadratic:

$$12\sigma^2 + 2\sigma - 7 = 0 \;\Rightarrow\; \hat{\sigma} = \sigma^*.$$