Dedicated to...

My family for supporting me, and to Professor Carnehl for guiding me

Contents

Chapter 1

Introduction

Teams, firms, and individual decision makers face a familiar choice: invest scarce time to learn, or exploit what currently looks best. In many environments with priority access, intellectual property, or capacity limits, early information can secure exclusive advantages in later periods. This thesis analyzes a minimal version of that trade off in a two player, short horizon bandit game and compares closed form predictions with behavior learned through self play reinforcement learning.

Baseline environment. Two players interact for $T = 2$ periods. At $t = 1$ each chooses from $\{\text{Do}, \text{ Think}\}$. Do exploits the ex ante more likely arm; Think (at cost c) discovers the good arm with probability r. Each period offers a single prize V with success probability λ when the good arm is pulled. Outcomes from period 1 are hidden until the end. If exactly one player thought, the player holds an exclusive right to the period 2 prize; if both thought the right is shared; otherwise the prize is contestable. These rules isolate the value of information from belief spillovers and make the Do versus Think trade off transparent.

Questions.

- How do the breakthrough rate r and the thinking cost c shift the period 1 Think frequency in the symmetric mixed equilibrium?
- How does extending exclusivity to additional future periods change incentives to Think?
- How closely can a simple self-play procedure reproduce the comparative statics of the model and the analytical thresholds?

Contributions.

- Derive a closed form symmetric mixed equilibrium for $T = 2$ with clear comparative statics in r and c.
- Propose a $T = 3$ extension (Extension A) where exclusivity covers both $t = 2$ and

$t = 3$, strengthening incentives to acquire information at $t = 1$.

- Implement tabular Double Q learning in self play that mirrors the payoff rules one for one and benchmark the learned Think frequencies against theory.

Preview of results. In the $T = 2$ baseline, the Think frequency rises with r and falls with c, with regime switches close to the closed form cutoffs $(r^\dagger, c^\dagger)$. Under Extension A, the Think curve shifts upward relative to $T = 2$ and reaches the Always Think region at lower r. Self play reproduces these signs and thresholds; levels can sit below theory when rewards from thinking are sparse or delayed, especially under Extension A.

Scope and modeling choices. I keep a single simultaneous decision at $t = 1$, state independent breakthroughs, one shared prize per period, and hidden period 1 outcomes. These choices keep the algebra simple and align the learning task with the model so behavior changes can be attributed to r, c, and the exclusivity rule.

Roadmap. The next chapter presents the model and payoffs. I then derive the $T = 2$ equilibrium and comparative statics, followed by the $T = 3$ exclusivity extension (Extension A). A subsequent chapter details the simulation design, and the results chapter reports learning outcomes with theory overlays. The discussion covers modeling choices, limitations, and implications. The conclusion summarizes and outlines next steps. Related work appears at the end for readability.

Chapter 2

Model

2.1 Environment and Primitives

We study a two-period interaction ($T = 2$) between two players. There are two arms, L and R, exactly one of which is good; the common prior is $p = \Pr(\theta = L) \in (0, 1)$. Pulling the good arm succeeds with probability $\lambda \in (0, 1)$ and yields prize $V > 0$, while the bad arm pays 0. Thinking carries a cost $c \geq 0$. Let $r \in (0, 1)$ denote the per-period breakthrough probability; one micro-foundation is a Poisson arrival with rate μ so that $r = 1 - e^{-\mu}$. There is no discounting. Importantly, discovery is independent of the state, so *no breakthrough means beliefs remain at p.*

2.2 Actions, Timing, and Observables

At $t = 1$, each player selects $a_i \in \{D, T\}$, where D stands for Do (exploit) and T for Think. These choices are publicly observed, whereas breakthroughs are private. Under D, a player chooses an arm. With a common prior and no signals, we focus on exploiting the more-likely arm in both periods, since deviations cannot raise expected payoffs in this environment. Under T, the player pays c and learns the good arm with probability r; otherwise the belief stays at p. Outcomes from period-1 exploitation are *not revealed* before $t = 2$. Any period-1 success is paid at the end and does not feed back into the $t = 2$ decision. The game always proceeds to $t = 2$, at which point only exploitation is available.

2.3 Prize Technology and Collision Rule

Each period delivers *at most one prize.* If both exploit in the same period, they divide V equally. Within a period, success is a single common event; across periods, success events are conditionally independent given θ.

Exclusivity (baseline). If exactly one player chose T at $t = 1$, that player obtains an exclusive claim to the period-2 prize. The non-thinker may still exploit at $t = 2$ but *cannot appropriate the prize of that period.* If both thought, they share the period-2 claim (each receiving half of the continuation value defined below). If neither thought, the period-2 prize is contested and split if both exploit. Realized period-1 payoffs are unaffected by these rules.

2.4 Continuation with One Period Left

For a player who chose T at $t = 1$, the expected period-2 value (before deducting c) is

$$\Lambda(p) = \lambda V\,[\,r + (1-r)p\,],$$

since with probability r the player knows the good arm and exploits it, and with probability $1 - r$ the player exploits based on the prior p.

2.5 Period-1 Expected Payoffs (ex ante over θ)

- (D, D): each earns $p\,\lambda V$ in total over the two periods.
- (D, T): Doer $= p\,\lambda V$; Thinker $= -c + \Lambda(p)$ (exclusive period-2 right).
- (T, D): symmetric to (D, T).
- (T, T): each $= -c + \frac{1}{2}\,\Lambda(p)$ (shared period-2 right).

Chapter 3

T=2 Equilibrium and Comparative Statics

3.1 Equilibrium (T=2)

Let $x \in [0,1]$ denote the opponents probability of choosing D at $t = 1$.

$$U_D(x) = p\,\lambda V, \qquad U_T(x) = -c + \Big[x \cdot 1 + (1-x)\cdot \tfrac{1}{2}\Big]\Lambda(p) = -c + \frac{1+x}{2}\,\Lambda(p).$$

(When the opponent Does, a unique thinker holds the full period-2 right; when the opponent Thinks, the right is shared.)

A symmetric mixed equilibrium requires $U_D(x^\star) = U_T(x^\star)$, which implies

$$x^\star = \frac{2\,(p\,\lambda V + c)}{\Lambda(p)} - 1.$$

This mixture is interior exactly when $\frac{1}{2}\,\Lambda(p) < p\,\lambda V + c < \Lambda(p)$; at the boundaries we obtain indifference with corner best responses (Always Think below, Always Do above).

3.2 Calibration Note

To target $x^\star = \frac{1}{3}$, set $c = \frac{2}{3}\Lambda(p) - p\,\lambda V$ (for instance, $p = 0.3,\ \lambda = 1,\ V = 1,\ r = 0.8 \Rightarrow \Lambda = 0.86$, yielding $c \approx 0.273$).

3.3 Comparative-Statics Figures (T=2)

Here we look at two diagrams that confirm our calculations for $x^\star$: the Think frequency versus thinking cost c, and the Think frequency versus breakthrough probability r.

Each diagram shows the equilibrium Think frequency in period 1, i.e., the probability that an agent chooses to Think in the symmetric mixed equilibrium, which equals $1 - x^\star$ where $x^\star$ is the Do probability that solves the Do Think indifference.

$$U_D \;=\; p\,\lambda V, \qquad U_T(x) \;=\; -c \;+\; \frac{1+x}{2}\,\Lambda(p), \qquad \Lambda(p) \;=\; \lambda V\,[\,r + (1-r)p\,].$$

Think frequency versus cost. Figure 3.1 displays the period-1 Think frequency as a function of c. Parameters: prior $= 0.6$, success probability when the good arm is pulled $= 0.7$, prize $= 1$, breakthrough probability $= 0.4$. The dashed line at $c^\dagger \approx 0.112$ is the threshold where the agent is indifferent between Doing and Thinking. To the left (blue shading) the equilibrium is a mixed strategy with positive Think frequency; to the right, the equilibrium is Always Do (Think frequency $= 0$). For this calibration, there is no Always Think region at non-negative costs.

The curve is a straight line because the continuation value is linear:

$$x^\star \;=\; \frac{2\,(p\,\lambda V + c)}{\Lambda(p)} - 1, \qquad 1 - x^\star \;=\; 2 - \frac{2\,(p\,\lambda V + c)}{\Lambda(p)}.$$

For fixed p, λ, V, r, $\Lambda(p)$ is constant, so Think frequency is linear in c and decreases as c rises.

Moreover, the cost threshold used for the dashed line is

$$c^\dagger \;=\; \Lambda(p) - p\,\lambda V \;=\; \lambda V\,(1-p)\,r \;\approx\; 0.112.$$

Think frequency versus breakthrough probability. Figure 3.2 sets the x-axis to r and keeps $c = 0.10$. The dashed line at $r^\dagger \approx 0.357$ is the threshold where the continuation value from thinking just equals Do + cost. To the left, the equilibrium is Always Do (pink); to the right, it is a mixed strategy with Think frequency rising in r.

We derive the threshold from $p\,\lambda V + c = \Lambda(p)$:

$$p\,\lambda V + c = \lambda V\big[p + r^\dagger(1-p)\big], \tag{3.1}$$

$$r^\dagger = \frac{c}{\lambda V(1-p)} \approx 0.357. \tag{3.2}$$

At the baseline $r = 0.40$ (dot), Think frequency is 0.046; at $r = 1$ it is about 0.51. These parameters were chosen for readability: setting $c = 0.10$ and $r = 0.40$ places the baseline inside the mixed region but close to the thresholds. The implied continuation value is $\Lambda(p) = 0.532$.

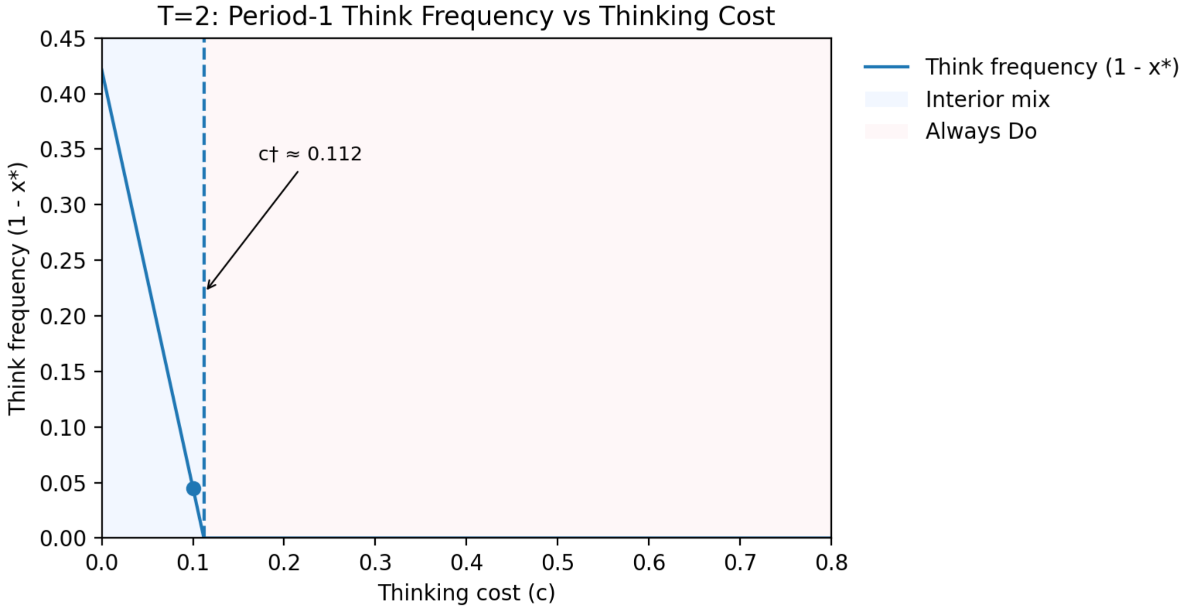

Figure 3.1: T=2: Think frequency versus thinking cost c. The dashed line marks $c^{\dagger}$.

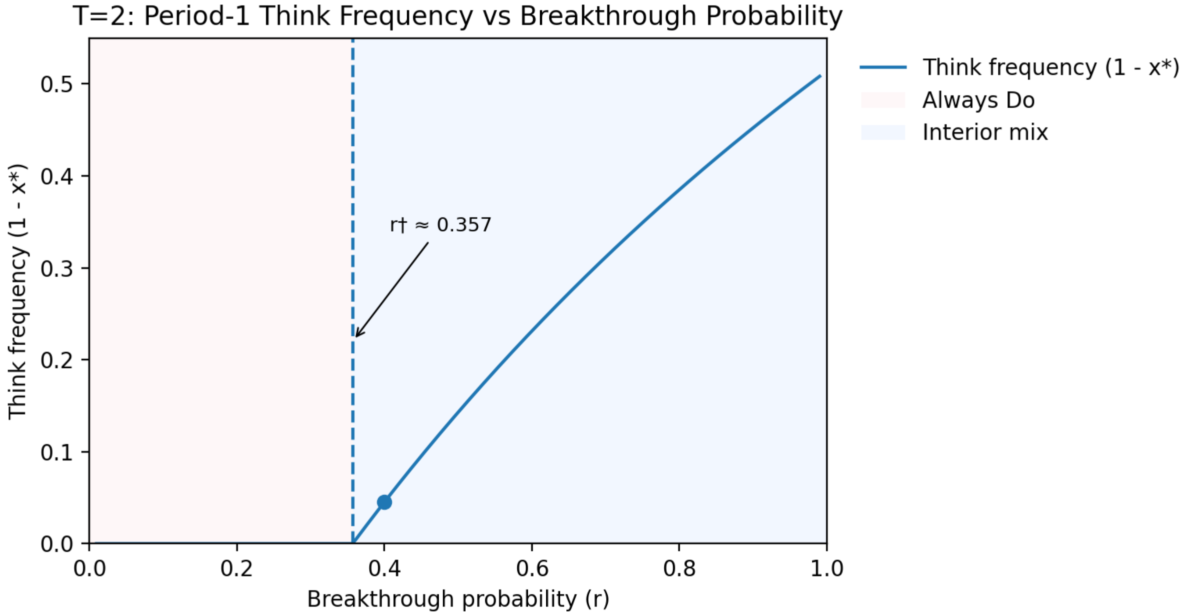

Figure 3.2: T=2: Think frequency versus breakthrough probability r. The dashed line marks $r^{\dagger}$.

3.4 Relation to the Literature

The original When to Think and When to Do paper is continuous-time with a time-dependent value of thinking; here we use a per-period breakthrough probability r and a two-period horizon for tractability.

Chapter 4

T=3 Extension: Exclusivity Extension (Extension A)

Motivation. Extension A grants exclusivity to a successful thinker over two future periods. This lets us test whether lengthening the exclusivity window raises the value of information acquisition and shifts the Do Think trade-off.

4.1 Baseline check: exclusivity only in period 2

Under $T = 3$ with a *single* decision at $t = 1$ and exclusivity only for period 2 (period 3 contestable, no outcomes revealed between periods), the indifference condition is unchanged relative to $T = 2$. The extra contestable period adds the same term $\frac{1}{2}p\,\lambda V$ to both actions:

$$U_D^{(3)} \;=\; p\,\lambda V \;+\; \tfrac{1}{2}p\,\lambda V, \qquad U_T^{(3)}(x) \;=\; -c + \tfrac{1+x}{2}\,\Lambda(p) \;+\; \tfrac{1}{2}p\,\lambda V, \tag{4.1}$$

so common terms cancel and the $T = 3$ curve coincides with the $T = 2$ benchmark.

4.2 Extension A: exclusivity in periods 2 and 3

Now assume a unique period-1 thinker holds exclusivity in both periods 2 and 3 (if both think, they share the right in both periods; if neither thinks, both periods are contestable). All other baseline features remain as before. The period-1 expected payoffs are:

$$U_D^A(x) = p\,\lambda V\Big(1 + \tfrac{x}{2}\Big), \qquad U_T^A(x) = -c + (1+x)\,\Lambda(p). \tag{4.2}$$

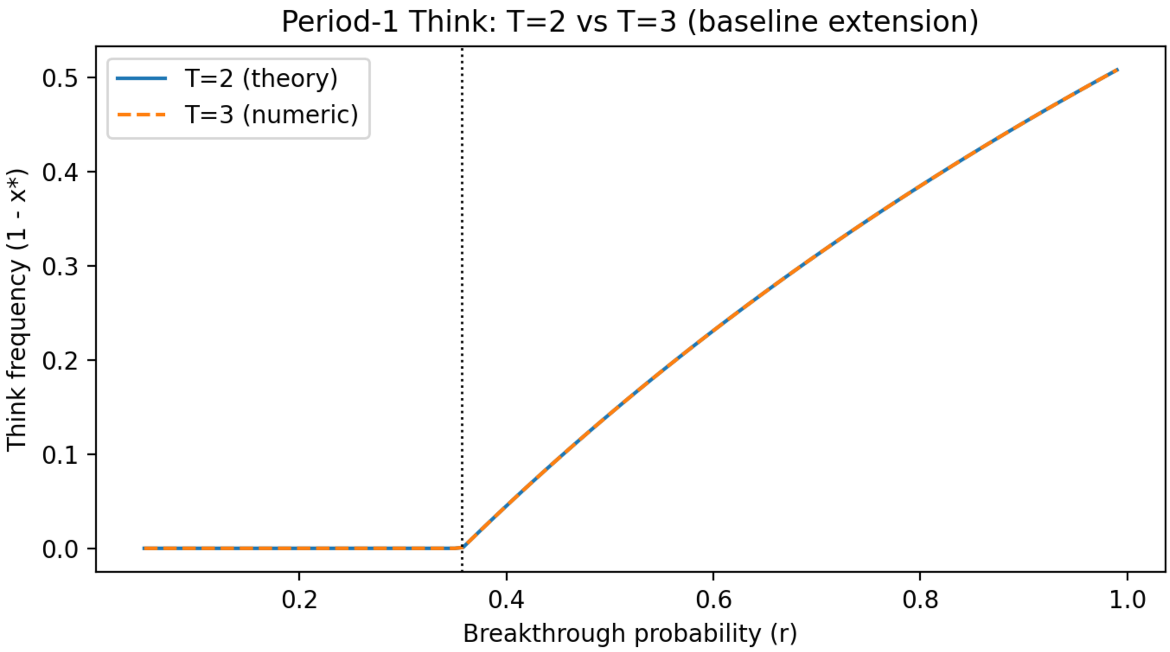

Figure 4.1: Think frequency vs. breakthrough probability r (solid: $T=2$; dashed: $T=3$ baseline with exclusivity only at $t=2$). The vertical dashed line is $r^{\dagger}=\frac{c}{\lambda V(1-p)}\approx 0.357$ for $p=0.6,\ \lambda=0.7,\ V=1,\ c=0.10$.

Indifference yields

$$c=(\Lambda-p\,\lambda V)+x\Big(\Lambda-\tfrac{1}{2}p\,\lambda V\Big), \tag{4.3}$$

$$x^{\star}=\frac{c-(\Lambda-p\,\lambda V)}{\Lambda-\frac{1}{2}p\,\lambda V}. \tag{4.4}$$

Region cutoffs:

$$c_{\text{lower}}=\Lambda-p\,\lambda V, \qquad c_{\text{upper}}=2\Lambda-\tfrac{3}{2}p\,\lambda V, \tag{4.5}$$

$$r_{\text{lower}}=\frac{c}{\lambda V(1-p)}, \qquad r_{\text{upper}}=\frac{c/(\lambda V)-0.5\,p}{2(1-p)}. \tag{4.6}$$

Overlay with $T=2$. Exclusivity extended to two future periods shifts the Think frequency curve *up* relative to $T=2$, and it reaches the Always Think region at lower r.

Comparative statics with fixed c (vary r). Think frequency increases smoothly and slightly concavely in r, hitting 1 at r_{upper} for sufficiently high r.

Comparative statics with fixed r (vary c). With r fixed, Think frequency declines with c and eventually hits zero as costs rise.

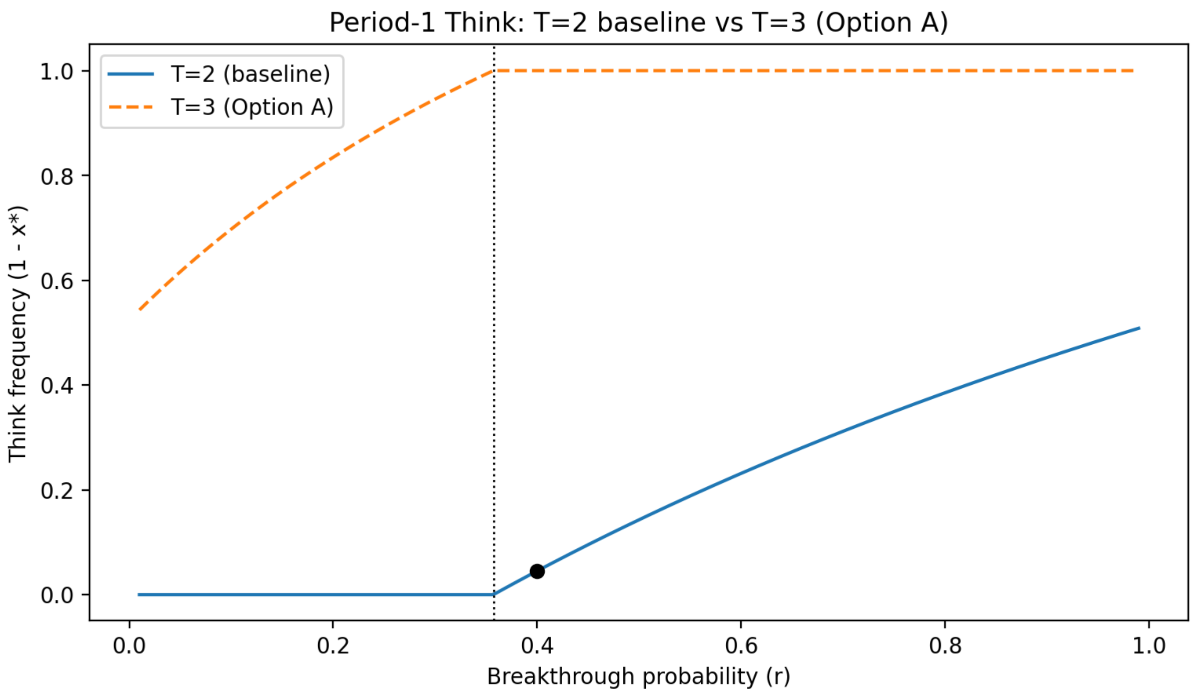

Figure 4.2: Think frequency vs. r (solid: $T = 2$; dashed: $T = 3$ Extension A). The Extension-A curve lies strictly above the $T = 2$ benchmark and reaches the Always Think region earlier.

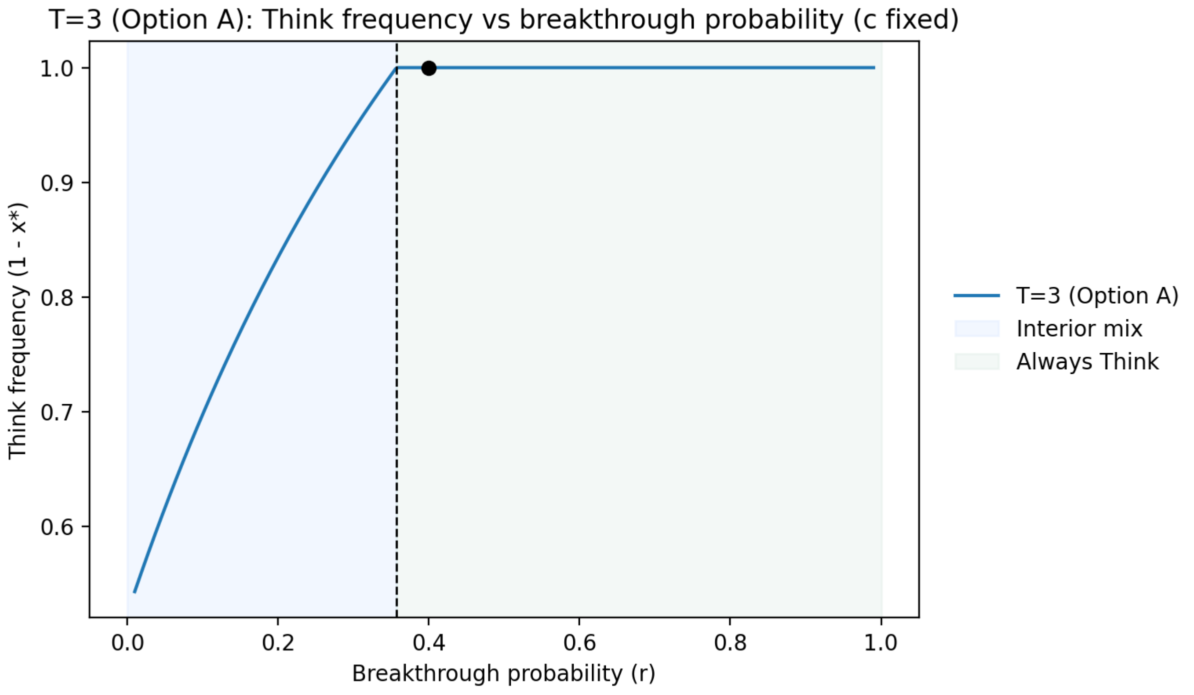

Figure 4.3: $T = 3$ (Extension A): Think frequency vs. breakthrough probability r (cost fixed).

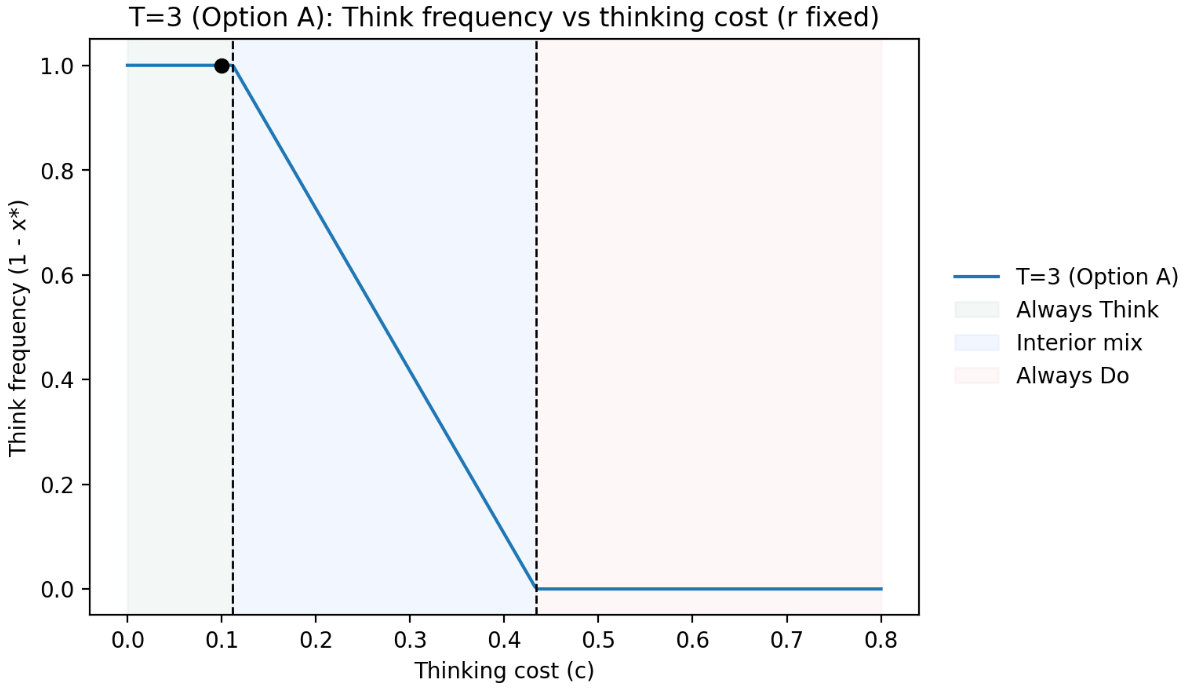

Figure 4.4: $T = 3$ (Extension A): Think frequency vs. thinking cost c (breakthrough probability fixed).

Chapter 5

Simulation Design

5.1 Environment

I implement the two-player bandit exactly as in the model with a single simultaneous move at $t = 1$ and forced exploitation thereafter. Each period draws one shared Bernoulli success and pays at most one prize V. Period-1 outcomes are hidden until the end. Actions at $t = 1$ are public; breakthroughs are private.

Period-1 success. If at least one player chooses Do, the shared success event occurs with probability λp. If success occurs, Doers split V if both chose Do; otherwise the unique Doer receives V. *(Code: `lam*p` and split logic in `run_episode`.)*

Rights at $t = 2$ (and at $t = 3$ under Extension A). If exactly one player thought, that player has exclusive rights: the period-2 prize, if it materializes, goes only to that player. If both thought, rights are shared and any prize is split. To match the continuation value $\frac{1}{2}\Lambda(p)$ per thinker, shared rights use a single virtual discovery draw at $t = 1$, implemented as a Bernoulli(r); this flag is reused in later periods. If neither thought, the prize is contestable and split upon success. *(Code: `rights` $\in \{$ `none`, `exclusive0`, `exclusive1`, `shared` $\}$ and a single `K_shared` draw; success probability via `period_success_prob()`.)*

Later-period success probabilities. If a rights holder knows (private breakthrough for an exclusive holder or `K_shared`$= 1$ for shared rights), success occurs with probability λ; otherwise with λp. *(Code: `period_success_prob()`.)*

Extension A ($T = 3$). When $T = 3$ and `option_A=True`, the same rights and success-probability logic apply again at $t = 3$ using the knowledge flags set at $t = 2$. *(Code: reuse of `b0`/`b1` and `K_shared`.)*

5.2 Observations, actions, and rewards

Action space. Only at $t = 1$: $\{\text{Do}, \text{Think}\}$.

State/observation. A single dummy state (no public signals); each agent privately draws its own breakthrough if it chooses Think. *(Code: `reset()` returns* $(0, 0)$*.)*

Costs and payoffs. Choosing Think subtracts c immediately; period prizes are added according to rights and success. *(Code: `R0/R1` accumulation.)*

5.3 Learning rule

Two independent Double Q-learning agents play in self-play (single-step problem; $\gamma = 0$). Each agent maintains two tables Q^A, Q^B for the single state and two actions. Updates alternate between tables; exploration is ε-greedy with slow decay. The Think action is given a small optimistic initialization so it is sampled early. *(Code: `DoubleQLearner`, `alpha`, `eps_start/eps_end/eps_decay`, `optimistic_init`.)*

5.4 Training, early stop, and evaluation

Training loop. For each parameter point, training runs up to N episodes (typically 3×10^5 to 8×10^5, depending on the sweep). Periodically the agents greedy actions are checked; if both stabilize for several checks, training stops early.
(Code: `train_self_play(..., patience_checks=3)`.)

Evaluation metric. After training, ε is set to 0 and 150k episodes are simulated to measure the greedy period-1 Think frequency, averaged across agents.
(Code: `evaluate_greedy_think_freq`.)

Seeds and warm-start. For each grid point I average over seeds $\{0, 17, 39, 71\}$. When sweeping a parameter, each grid point warm-starts its Q-tables from the previous point to reduce variance and runtime. *(Code: `avg_over_seeds_with_warmstart`.)*

Extension A knobs. Because Thinks reward is more delayed when $T = 3$ with Extension A, I use slightly stronger exploration and longer training (higher `episodes`, slower `eps_decay`, slightly larger `alpha`, and modest `optimistic_init`). *(Code: `set_r_and_boost` in the* $T = 3$ *sweep.)*

5.5 Sweeps and outputs

$T = 2$. (i) r-sweep on $r \in [0.05, 0.95]$ with c fixed at 0.10; (ii) c-sweep on $c \in [0, 0.35]$ with r fixed at 0.40.

$T = 3$ Extension A. Same r-sweep as above, using the Extension-A training schedule.

For each sweep I save a CSV with learned frequencies and the theory overlays, and render a PNG plot. *(Code: `sweep_vs_r_and_plot`, `sweep_vs_c_and_plot`, `sweep_optionA_vs_r_and_plot`.)*

5.6 Theory overlays

Dashed overlays come from the analytic mixes used in the paper:

$$\text{T=2:} \quad 1 - x^\star, \quad x^\star = \frac{2(p\lambda V + c)}{\Lambda(p)} - 1, \qquad \Lambda(p) = \lambda V\,[r + (1-r)p].$$

$$\text{T=3 Extension A:} \quad 1 - x^\star_A, \quad x^\star_A = \frac{c - (\Lambda - p\lambda V)}{\Lambda - \frac{1}{2}p\lambda V}.$$

(Code: `theory_think_freq_T2`, `theory_think_freq_T3_OptionA`.)

5.7 Baseline parameters

Unless stated otherwise: $p = 0.6$, $\lambda = 0.7$, $V = 1$, $c = 0.10$, $r = 0.40$, exclusivity on. For $T = 3$ without Extension A, the period-1 indifference is the same as for $T = 2$ (the third period adds the same contestable value to both actions), so I do not plot it separately.

5.8 Why learned curves can deviate from theory

Self-play Double Q-learning faces sparse or delayed rewards for Think (especially at low r), and the opponent is nonstationary. With finite episodes this can induce under-investment in Think at small r even when the payoff rules are implemented exactly. Longer runs, slower ε decay, and warm-starts reduce (but do not fully eliminate) the gap.

Chapter 6

Results

6.1 Empirical patterns from the learning curves

6.1.1 $T = 2$ Think frequency vs. r (with c fixed at 0.10)

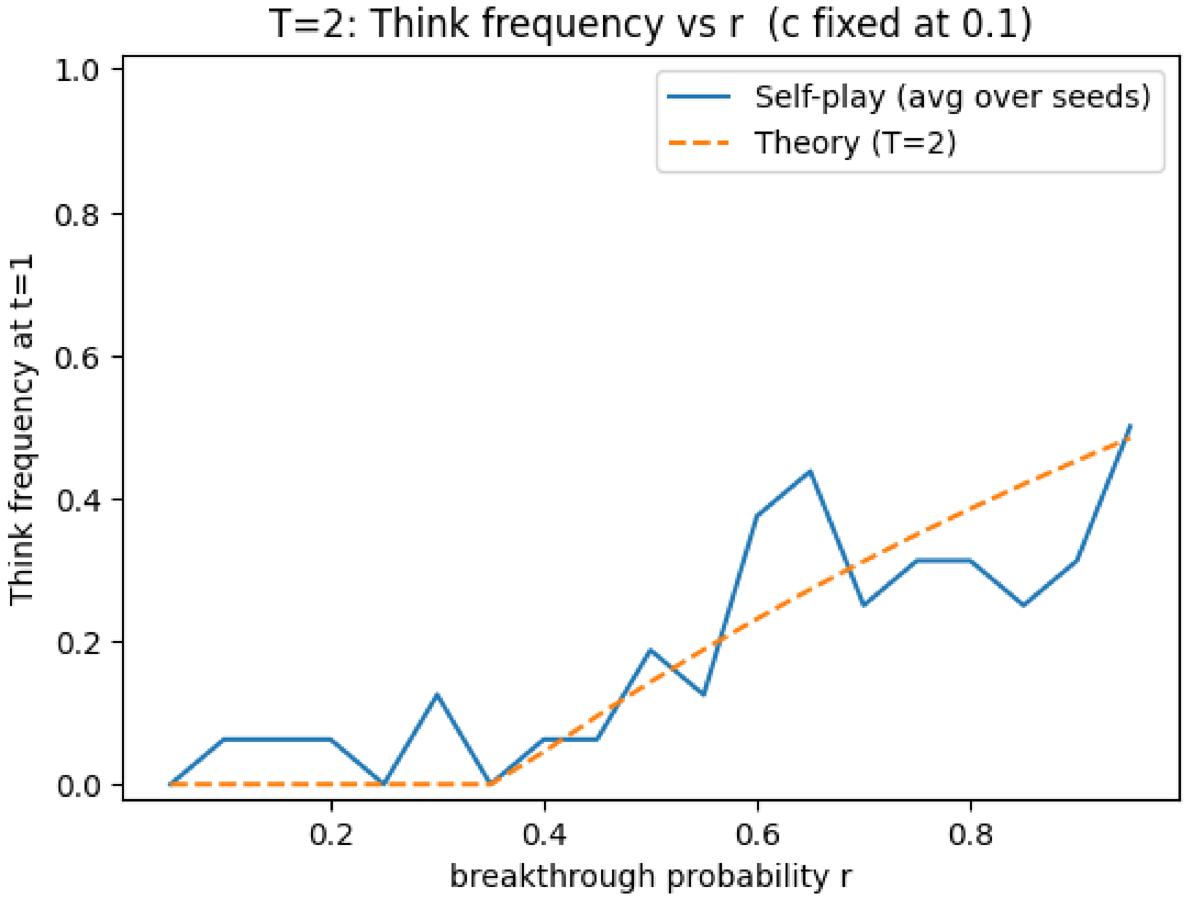

Figure 6.1: Self-play (avg. over seeds) vs. theory: $T = 2$, Think frequency as a function of r with $c = 0.10$.

- **Shape.** The learned Think rate is essentially 0 for small r and then increases monotonically once r passes the neighborhood of the theoretical cutoff $r^{\dagger} \approx 0.357$.
- **Level vs. theory.** Across much of the grid the curve sits *below* the $T = 2$ benchmark (under-investment in Think at moderate r), but it tracks the positive slope and enters the mixed region toward the right-hand side.
- **Local variance.** Small fluctuations reflect stochastic self-play and finite training. Averaging over seeds $\{0, 17, 39, 71\}$ reduces but does not eliminate this noise.
- **Interpretation.** When breakthroughs are rare, Think yields sparse/delayed feedback and is under-sampled by ε-greedy Double-Q. As r becomes large enough, the value signal strengthens and agents tilt toward Think, consistent with the comparative statics of the model.

6.1.2 $T = 2$ Think frequency vs. c (with r fixed at 0.40)

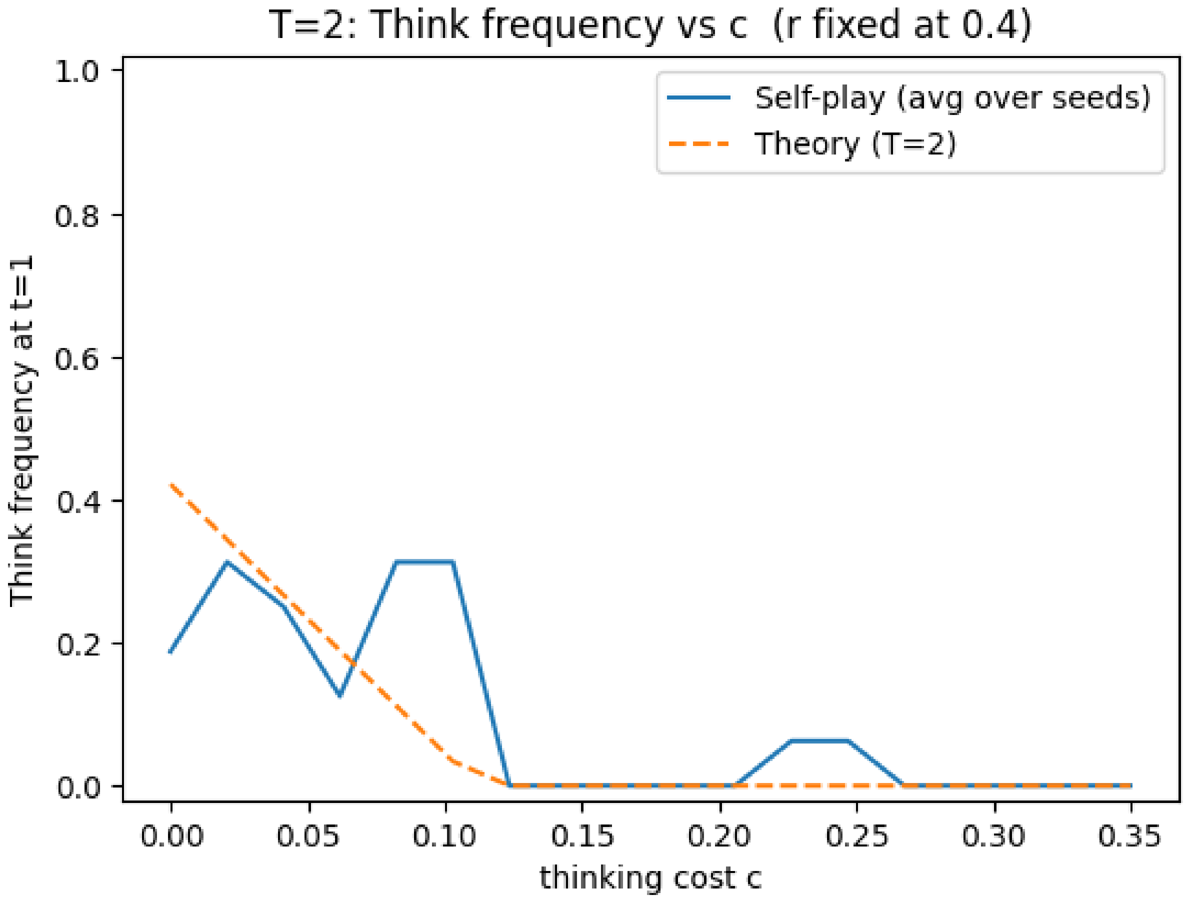

Figure 6.2: Self-play (avg. over seeds) vs. theory: $T = 2$, Think frequency as a function of c with $r = 0.40$.

- **Shape.** The learned Think frequency falls with cost and approaches (or reaches) zero slightly before/around the theoretical threshold $c^{\dagger} \approx 0.112$.

- **Match quality.** This panel aligns most closely with theory: both the slope and the crossing are in good agreement; remaining gaps are small and largely due to sampling noise.
- **Takeaway.** With r held fixed and reasonably high, Thinks reward is less sparse; learners recover the Do Think trade-off implied by the indifference condition.

6.1.3 $T = 3$ with Extension A Think frequency vs. r (with c fixed)

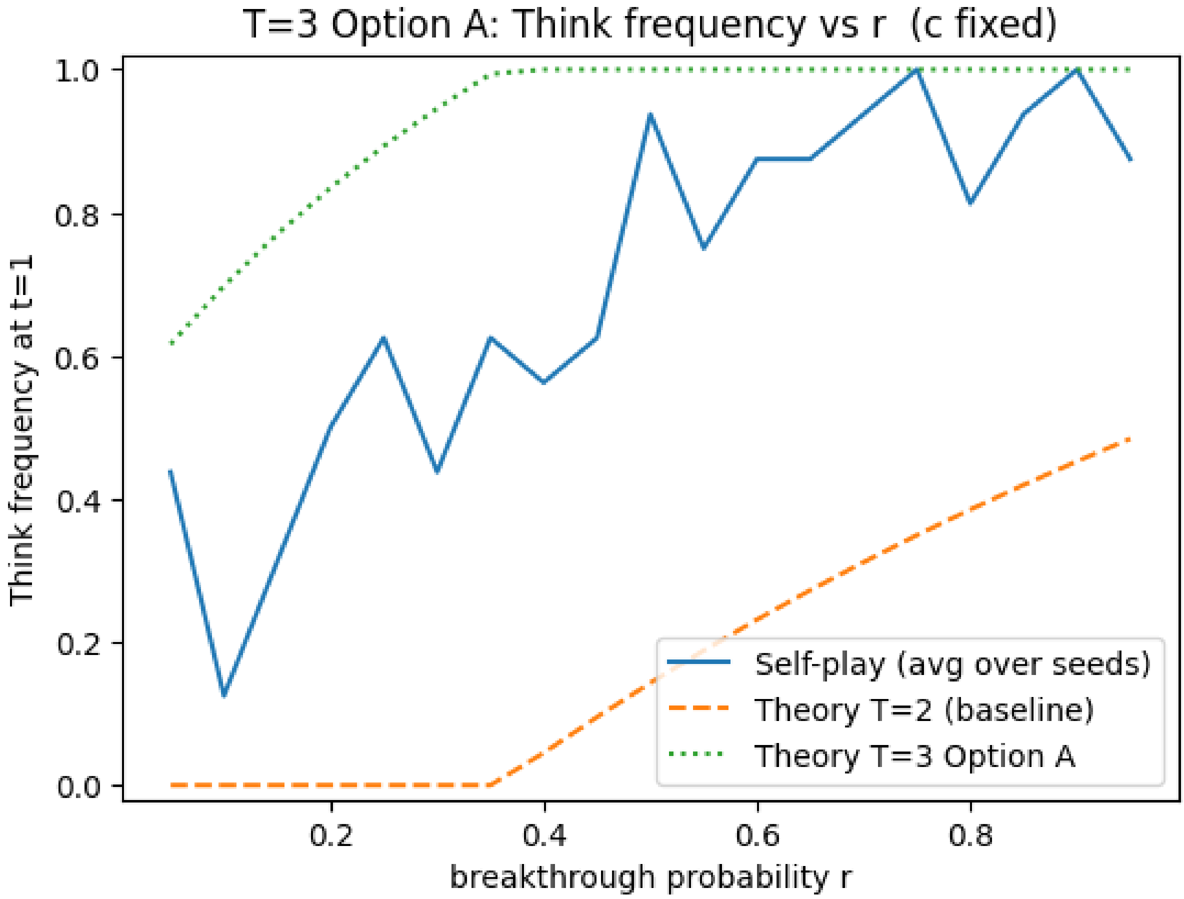

Figure 6.3: Self-play (avg. over seeds) vs. theory: $T = 3$ Extension A, Think frequency as a function of r (with c fixed). Dashed overlay is the $T = 2$ baseline; dotted overlay is the Extension-A theory.

- **Qualitative agreement.** Think frequency rises steeply with r and lies well *above* the $T = 2$ baseline overlay, exactly as predicted under Extension A (exclusivity over two future periods increases the appeal of Think).
- **Level vs. Extension-A theory.** The curve still falls *below* the $T = 3$ Extension-A

benchmark at mid-range r (about 0.2–0.5). Rewards from Think arrive later; payoffs accrue in both $t = 2$ and $t = 3$; so exploration must propagate through two future periods.

- **Mitigation used.** Longer training, slower ε-decay, and a modest optimistic bump for Think improve alignment but do not fully close the gap expected with tabular self-play and nonstationary opponents.

6.1.4 Cross-figure takeaways

- **Correct comparative-statics signs.** Increasing r raises Think; increasing c lowers Think; extending exclusivity to $t = 3$ (Extension A) shifts the curve upward relative to $T = 2$.
- **Thresholds visible in data.** Regime changes occur near the analytical cutoffs ($r^\dagger$, $c^\dagger$), with a consistent bias toward *too little* Think at low/intermediate r.
- **Systematic deviation pattern.** Bias is largest precisely when Thinks rewards are sparse or delayed (low r; Extension A). This reflects learning dynamics, not payoff implementation.
- **Robustness knobs (kept light).** Averaging across seeds and warm-starting by grid point stabilize shapes; modestly higher episode counts or slower ε shrink gaps further at additional compute cost.
- **Sanity check.** In a $T = 3$ variant *without* Extension A (exclusivity only at $t = 2$), the theory coincides with $T = 2$. Empirically, that sweep produces curves visually indistinguishable from $T = 2$ (not shown), confirming the wiring of the environment.

6.1.5 Summary of Results

Under the $T = 2$ baseline, self-play reproduces the models comparative statics: Think increases with r and decreases with c, with changes occurring near the closed-form thresholds. Quantitatively, the learned policy typically *understates* the theoretical Think rate at moderate r, reflecting sparse/delayed feedback from Think and nonstationary best responses in self-play. Extending exclusivity to $t = 2$ *and* $t = 3$ (Extension A) shifts the curve upward as predicted; the learned curve remains below the Extension-A benchmark at mid-range r for the same reason; payoffs arrive later and call for more exploration. Increasing training length and softening ε-decay improve the match but do not remove these gaps, which is in line with known behavior of tabular Double-Q in two-player settings with delayed rewards.

Chapter 7

Discussion and Modeling Choices

7.1 What the results say

Across all panels, the models comparative statics show up in self-play: period-1 Think rises with the breakthrough rate r, falls with the cost c, and shifts up when exclusivity lasts two periods (Extension A). The regime switches occur close to the closed-form cutoffs ($r^{\dagger}$, $c^{\dagger}$). Quantitatively, the learned Think rates land a bit below theory at lowmid r and under Extension A, which lines up with the learning frictions below.

7.2 Why learned curves can sit below theory

Two frictions matter in self-play:

1. **Sparse/delayed payoffs for Think.** At small r discoveries are rare; under Extension A value arrives over two future periods. With finite episodes, ε-greedy can under-sample Think even when it is optimal in expectation.
2. **Nonstationary opponents.** Each agent learns against another learner, so payoffs come from a drifting policy pair. Double Q-learning helps with overestimation but does not remove nonstationarity effects.

Averaging over seeds, warm-starting along the sweep grid, slightly longer training, and slower ε decay narrow (but do not fully close) the gaps, typical for tabular self-play with delayed rewards.

7.3 Key modeling choices (and why)

- **Single decision at $t = 1$.** Keeps the algebra tractable and matches the one-step simulation, making the Do / Think indifference easy to see.
- **State-independent breakthroughs.** No news leaves beliefs at p; this isolates the

value of exclusivity from belief dynamics and makes the continuation term simple.

- **One shared success per period.** A single prize (split on collision) captures capacity/priority stories and yields clean expected payoffs.
- **Exclusivity rule.** If exactly one player thought, the player gets the next-period right (and under Extension A, also the $t = 3$ right). This mimics IP/priority-slot setups and creates the race-to-learn incentives we study.

7.4 Simplifications vs. natural variants

We hide period-1 outcomes until $t = 2$, do not update beliefs from failures, and do not allow midstream arm switches. These choices avoid path-dependent beliefs and long action histories that would blur the comparative statics of exclusivity and discovery intensity. Baselines without exclusivity are straightforward but less informative about the strategic value of information.

7.5 Robustness: Extension B (appendix)

Allowing a second chance to think at $t = 2$ (Extension B) drives period-1 thinking to (near) zero at the baseline parameters: agents can take the $t = 1$ expected prize and keep the option to secure exclusivity for $t = 3$. The appendix overlay shows a flat, near-zero period-1 Think rate across r, matching the idea that postponement dominates when late learning is on the table.

7.6 Limitations and external validity

This study sticks to short horizons $T \in \{2, 3\}$, tabular Q-learning with ε-greedy exploration, and symmetric primitives. Finite-episode effects, exploration-schedule sensitivity, and nonstationarity can move levels (though not the signs) of learned frequencies. Moving to function approximation, asymmetric costs/priors, or public outcome revelation may change the quantitative picture.

7.7 Implications and next steps

Exclusivity plus moderate discovery intensity can generate a race to think even when exploitation looks tempting. From a policy angle, stronger exclusivity (or priority access) tilts incentives toward information acquisition. Natural extensions: partial revelation between periods, heterogeneous costs, longer horizons with function approximation, and equilibrium refinements for richer action sets.

Chapter 8

Conclusion

This thesis analyzes a two-player explore exploit game on a short, fixed horizon. In the baseline $T = 2$ setup with a single move at $t = 1$, one shared prize per period, state-independent breakthroughs, and exclusivity at $t = 2$; the closed-form indifference pins down the symmetric mixed strategy and its comparative statics. I then extend the baseline to $T = 3$ by letting exclusivity span two future periods (Extension A) and check the expected shift in incentives. Finally, a minimal self-play Double Q-learning experiment mirrors the payoff rules one-for-one and lets me compare learned behavior to the theoretical benchmarks.

Main findings. (i) For $T = 2$, period-1 Think rises with the breakthrough rate r and falls with the cost c; the regime changes line up with the analytic cutoffs ($r^{\dagger}$, $c^{\dagger}$). (ii) For $T = 3$ with Extension A, two-period exclusivity strictly raises the value of thinking and shifts the Think-frequency curve up relative to $T = 2$; the always Think region is reached at lower r. (iii) Self-play matches these signs and visibly tracks the thresholds; absolute levels can sit below theory at low or mid r (and under Extension A), consistent with sparse/delayed rewards for Think and nonstationary opponents.

Robustness. A simple check (Extension B) that adds a second chance to think at $t = 2$ pushes period-1 thinking to (near) zero at the baseline parameters: agents take the period-1 expected prize and postpone learning, in line with the idea that late exclusivity makes early experimentation dominated.

Interpretation. The results show how exclusivity and discovery intensity together shape a race to think. When breakthroughs are likely or exclusive access covers more than one future period, agents tilt toward information acquisition even if immediate exploitation looks good. When discoveries are rare or can be delayed at low cost, exploitation wins out.

Limitations. The analysis sticks to short horizons ($T \in \{2, 3\}$), symmetric primitives, hidden period-1 outcomes, and tabular Double Q-learning with ε-greedy exploration. These choices keep the algebra clean and the simulation close to the model, but they can understate learned Think rates when rewards are sparse or arrive late.

Directions for future work. Natural next steps include partial outcome revelation between periods, asymmetric costs or heterogeneous priors, longer horizons with function approximation, and richer action sets (e.g., arm switching or public beliefs). On the learning side, policy-gradient or opponent-aware methods could reduce nonstationarity and narrow the remaining gap to theory.

Takeaway. Exclusivity is a strong lever: extending rights into future periods can move behavior from exploitation toward experimentation. In short, stronger priority access combined with moderate discovery intensity creates a measurable race to acquire information.

Appendix A

Reproducibility and Code Availability

All figures produced by the self play simulations can be reproduced from the Jupyter notebooks included with the submission (or available at the project repository). The environment requires Python 3.10, `numpy` and `matplotlib` only. Each notebook writes CSVs and PNGs to `./out_*` and `./figs/`. (References to option A are for Extension A and references to option B are for Extension B)

- **01_ML_final.ipynb**: generates the three ML panels used in the Results: *ML output 1.png* (T=2, Think vs. r), *ML output 2.png* (T=2, Think vs. c), and *ML output 3.png* (T=3, Extension A, Think vs. r with theory overlays).
- **02_T2_benchmarks.ipynb**: computes and plots the closed-form T=2 benchmarks (*t2_think_freq_vs_r.png*, *t2_think_freq_vs_cost.png*) used in Section 3.1 and 3.2.
- **03_T3_OptionA_compstat.ipynb**: produces the Extension A comparative-statics overlays (e.g., *t3A_vs_t2_think_diff.png* and *t3A_think_vs_r.png*).
- **03b_T3_baseline_equals_T2.ipynb**: sanity check that the T=3 baseline (exclusivity only at $t = 2$) coincides with T=2 (overlay figure in Section 4.1).
- **04_T3_OptionB_appendix.ipynb**: generates Appendix figures for Extension B (*t3B_think_vs_c.png*, *t3B_think_vs_r.png*, *t3B_vs_t2_think_vs_r.png*).

Default seeds are {0,17,39,71}; changing seeds alters levels slightly but not the qualitative patterns. To speed up on a laptop, reduce episode counts; final plots in the thesis were produced with the settings listed in Section “Training, early stop, and evaluation”.

Appendix B

Extension B (Appendix)

Setup and intuition

Extension B changes a single feature of the baseline. At $t = 2$ agents may choose to *Think* again, paying cost c. A unique thinker at $t = 2$ (if any) receives exclusivity only for period 3. Outcomes from period 1 remain hidden, period 2 is fully contestable, and each period still pays at most one prize.

Postponement dominates. From the perspective of $t = 1$, there is no benefit to thinking early:

- Payoffs in periods 1 and 2 do not depend on a $t = 1$ breakthrough, because period 1 outcomes are hidden and period 2 is contestable.
- The option to buy exclusivity for period 3 is available at $t = 2$ at the same cost c.

Thus, Think at $t = 1$ is (weakly) dominated by Do at $t = 1$ or by waiting and, if desired, Think at $t = 2$. The implied period 1 Think frequency is 0 for all r and c.

Figures

The plots show the predicted flat, near zero Think rate at $t = 1$.

Takeaway

Granting a second opportunity to Think at $t = 2$ eliminates the exclusivity motive at $t = 1$. With no informational benefit before $t = 2$, early experimentation is dominated, and the learned policy that sits at zero matches the theoretical prediction.

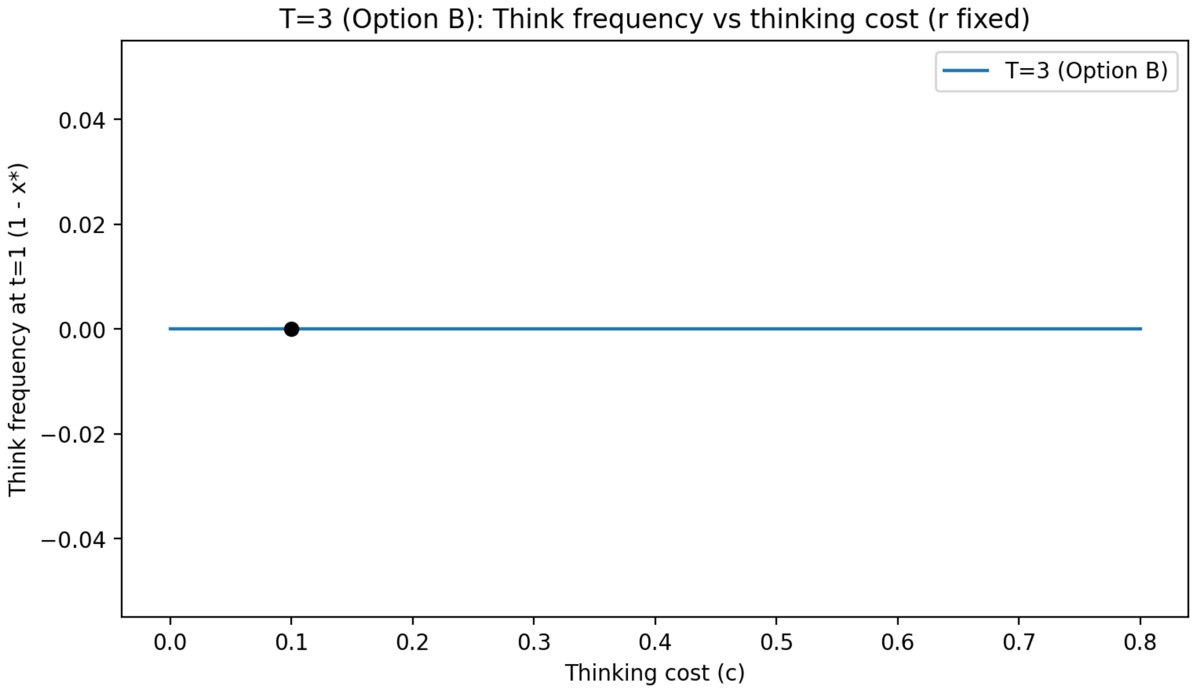

Figure B.1: $T = 3$ Option B: period 1 Think frequency versus thinking cost c for fixed r. The curve is flat at 0. The dot marks the baseline calibration.

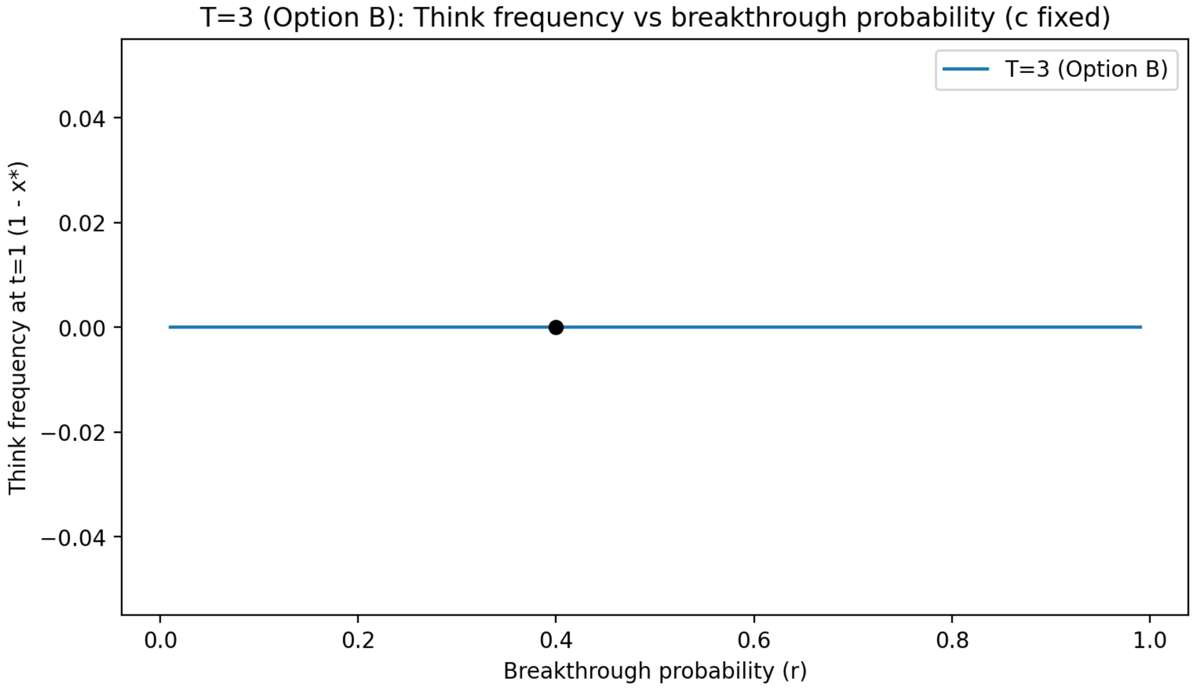

Figure B.2: $T = 3$ Option B: period 1 Think frequency versus breakthrough probability r for fixed c. Again flat at 0. The dot marks the baseline calibration.

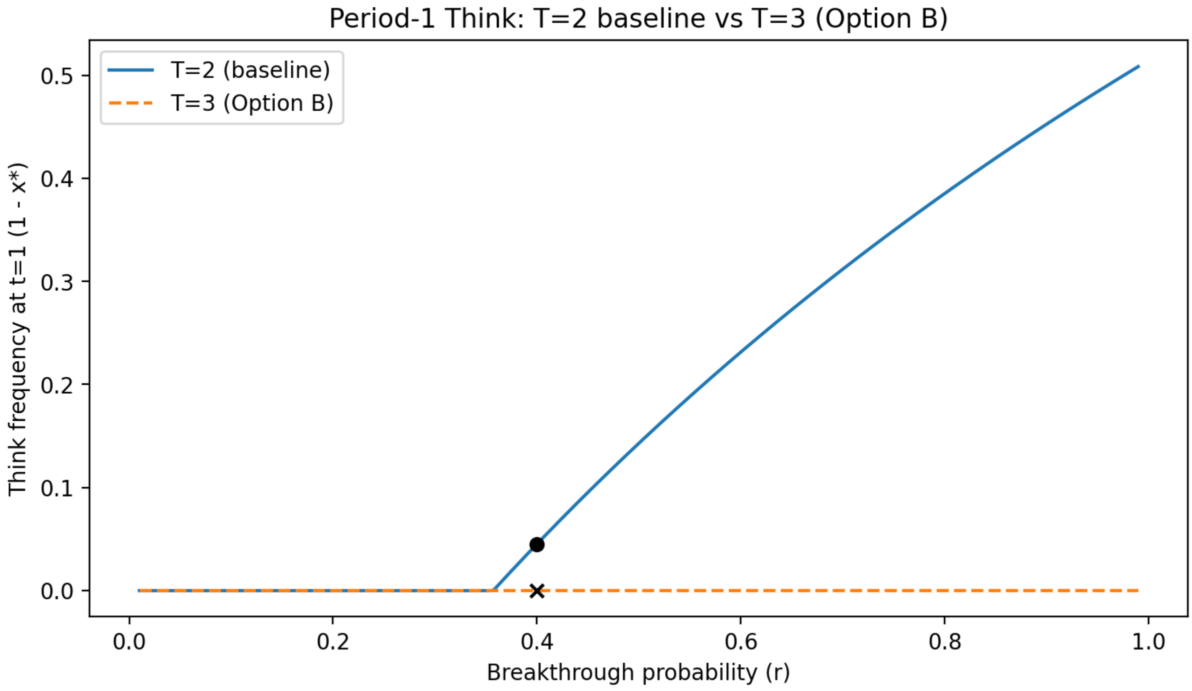

Figure B.3: Overlay of the $T = 2$ baseline (solid) and $T = 3$ Option B (dashed). Under $T = 2$, Think rises with r once $r^\dagger$ is crossed. Under Option B it remains at 0 for all r. The dot and cross indicate the baseline calibration.

Appendix C

Related Work

This thesis builds on the analysis of Christoph Carnehl, regarding the trade-off between information acquisition and exploitation. I recast the problem in discrete time, introduce a two-player exclusivity rule, and derive closed-form benchmarks that guide the simulations. Strategic experimentation in multi-agent settings is classically studied by Bolton and Harris, who show how incentives to learn depend on the actions of others; exclusivity in my model amplifies those incentives and generates a race to think.

Methodologically, the self-play experiments use tabular Q-learning with ε-greedy exploration following Sutton and Barto, and adopt Double Q-learning in the spirit of van Hasselt to mitigate value overestimation in a stochastic, nonstationary self-play environment. Relative to the statistical bandit literature surveyed by Lattimore and Szepesvári, my objective is not regret minimization but equilibrium comparative statics under a specific economic payoff structure. The role of exclusivity also parallels priority rights in patent-race models, where early discovery secures future returns.

Bibliography (with links)

1. Carnehl, C. Schneider, J., (2022). *On Risk and Time Pressure: When to Think and When to Do.* Working paper. Available at: https://arxiv.org/abs/2111.07451.
2. Bolton, P., & Harris, C. (1999). Strategic experimentation. *Econometrica*, 67(2), 349–374. DOI: 10.1111/1468-0262.00022.
3. Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (2nd ed.). MIT Press. Online draft: http://incompleteideas.net/book/the-book-2nd.html.
4. van Hasselt, H., Guez, A., & Silver, D. (2016). Deep reinforcement learning with Double Q-learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1). Preprint: https://arxiv.org/abs/1509.06461.

5. Lattimore, T., & Szepesvári, C. (2020). *Bandit Algorithms.* Cambridge University Press. Book page: https://www.cambridge.org/core/books/bandit-algorithms/8E39FD004E6CE036

6. Fudenberg, D., Gilbert, R., Stiglitz, J., & Tirole, J. (1983). Preemption, leapfrogging and competition in patent races. Available at: https://www.sciencedirect.com/science/article/abs/pii/0014292183900879.