Algorithmic Advice Human Compliance & Learning

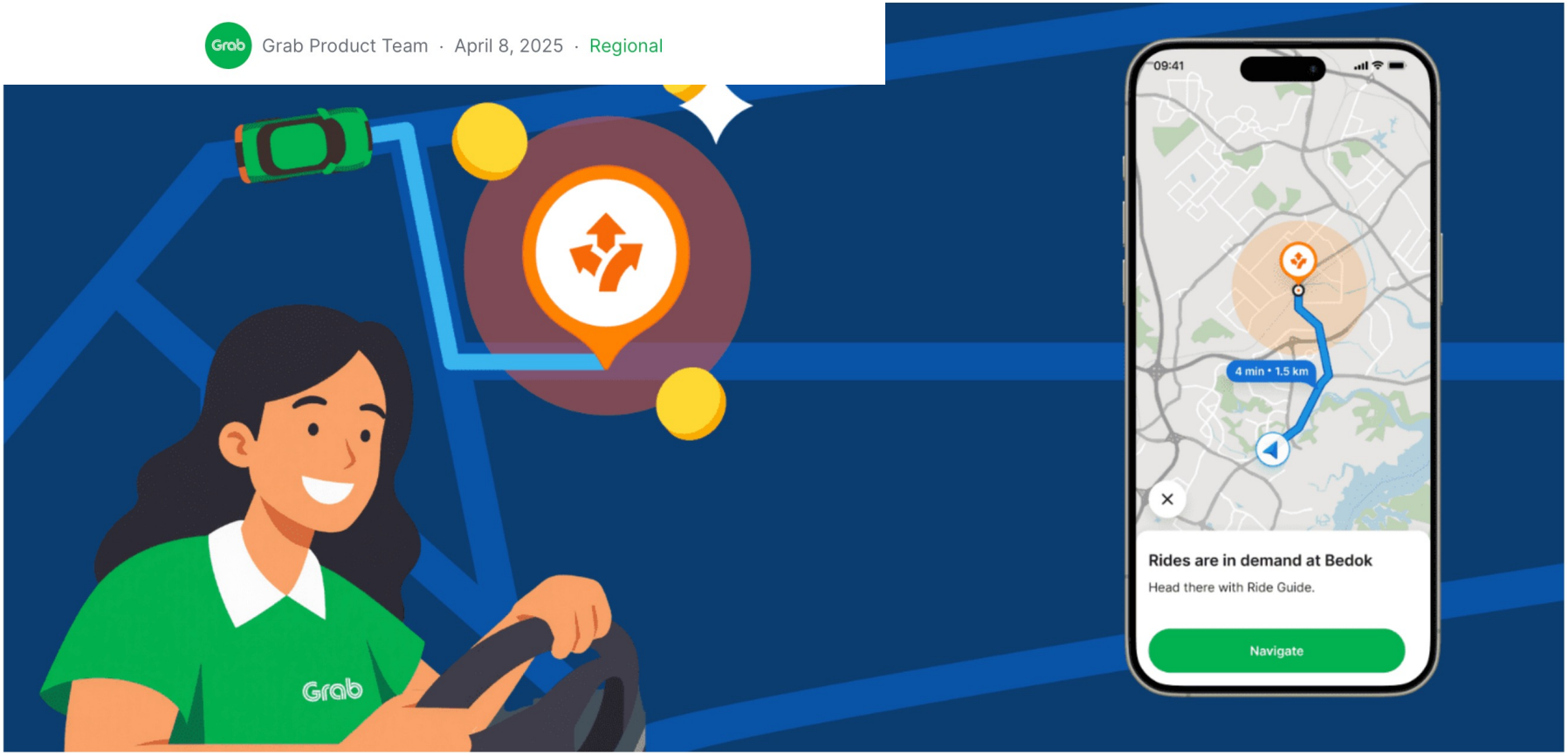
Park Sinchaisri
UC Berkeley Haas

University of Michigan SBEE 2026

AI systems are increasingly telling workers what to do next

New AI ride guidance feature predicts ride demand areas

Grab Product Team · April 8, 2025 · Regional

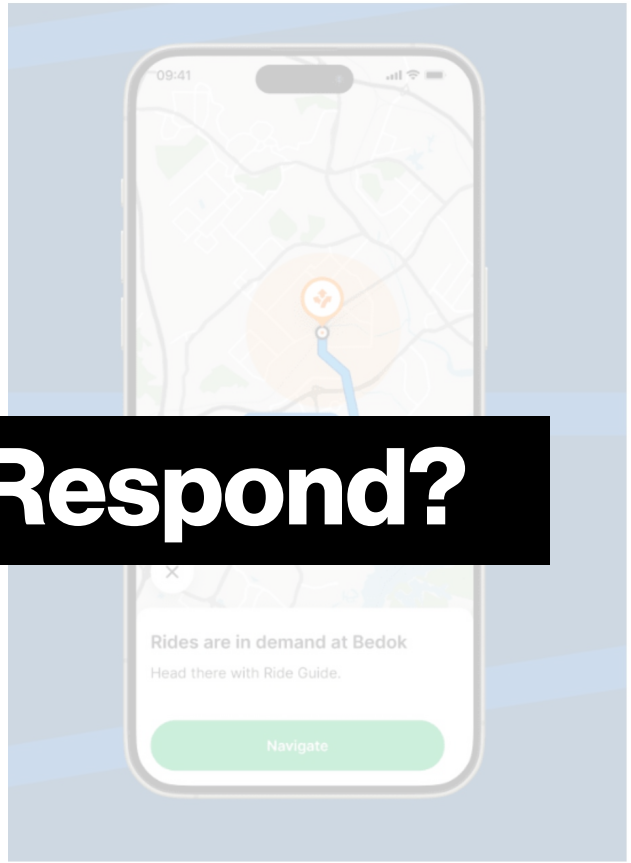


AI systems are increasingly telling workers what to do next

instacart

amazon

Grab



$51.22

1.9 mi

$49.22 batch earnings + $2 tip

Accept

But how Do Humans Respond?

Alibaba

AI assigns each worker a task sequence, every minute.

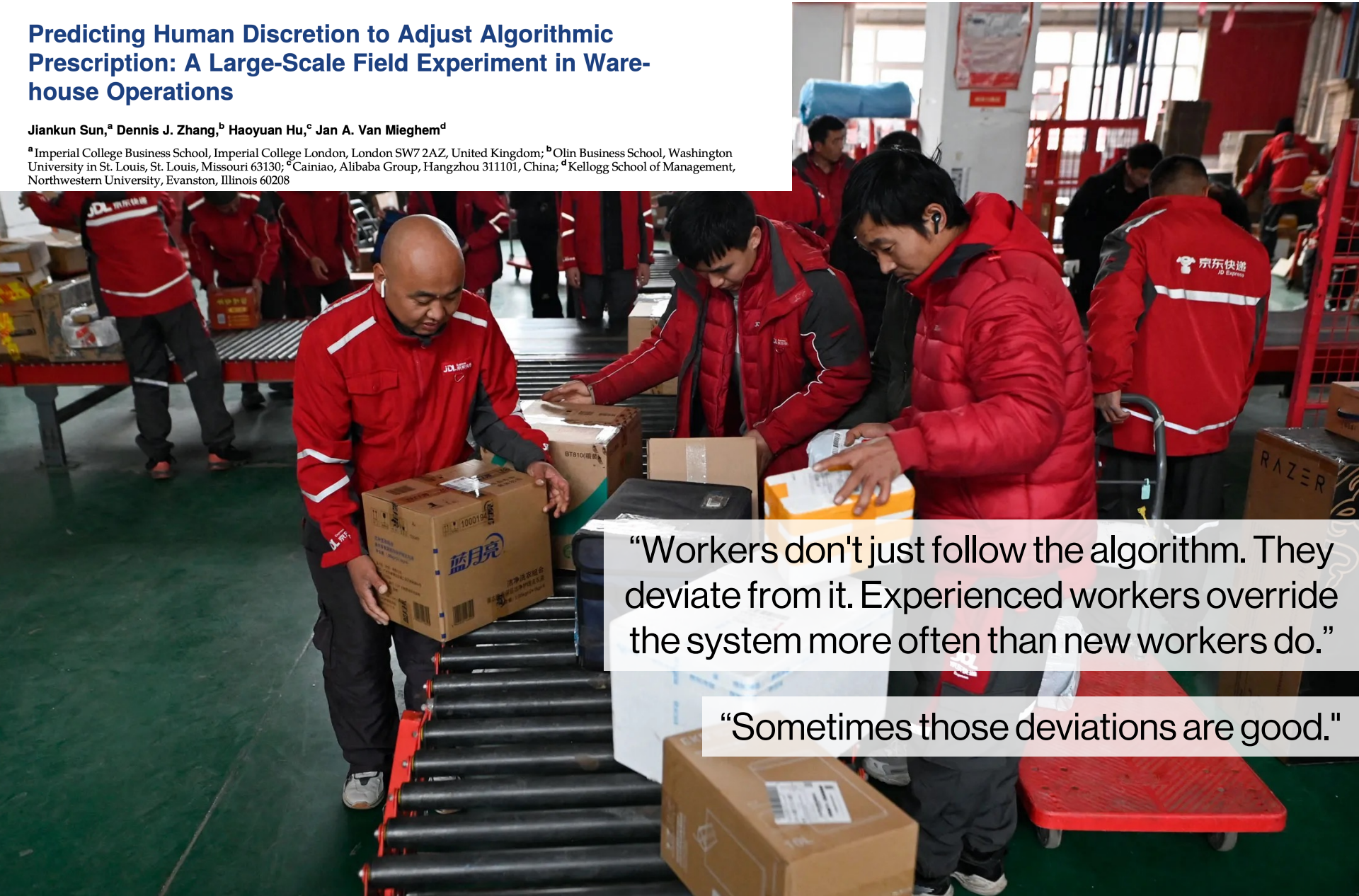
informs
http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 68, No. 2, February 2022, pp. 846–865
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Predicting Human Discretion to Adjust Algorithmic Prescription: A Large-Scale Field Experiment in Warehouse Operations

Jiankun Sun,[a] Dennis J. Zhang,[b] Haoyuan Hu,[c] Jan A. Van Mieghem[d]

[a] Imperial College Business School, Imperial College London, London SW7 2AZ, United Kingdom; [b] Olin Business School, Washington University in St. Louis, St. Louis, Missouri 63130; [c] Cainiao, Alibaba Group, Hangzhou 311101, China; [d] Kellogg School of Management, Northwestern University, Evanston, Illinois 60208



"Workers don't just follow the algorithm. They deviate from it. Experienced workers override the system more often than new workers do."

"Sometimes those deviations are good."

Many Reasons...

"Radiologists exercise more discretion as they accumulate experience."

(Ibanez, Clark, Huckman, Staats 2017)

"Humans choose human forecasters over AI, esp. after seeing AI performs.

They lose confidence in AI more quickly than humans after same mistake"

(Dietvorst, Simmons, Massey 2015)

Effect of seeing model: $\chi^2(1, N = 361) = 57.48, p < .001$

Effect of seeing human: $\chi^2(1, N = 361) = 0.14, p = .706$

"Managers follow pricing AI only when products are close to stockout."

(Caro & Saez de Tejada Cuenca 2023)

"Humans may struggle to operationalize AI advice into their workflow, and optimal AI advice tend to be counterintuitive."

(Bastani, Bastani, Sinchaisri 2026)

Also Grand-Clément & Pauphilet 2024; Balakrishnan et al. 2026; McLaughlin & Spiess 2022/2024; Caro & Saez de Tejada Cuenca 2023; Bastani & Cachon 2025; Chen, Kirshner, Ovchinnikov, et al 2025; DosSantos DiSorbo, Ferreira, Balakrishnan, Tong 2025; Keppler, Sinchaisri, Snyder 2025/2026; Snyder, Keppler, Leider 2025; Bastani, Bastani, Sinchaisri 2026; Delarue, Lian, Martin 2026; Dreyfuss & Hoong 2026; Poulidis, Bastani, Bastani 2026; ...

Some Rely on AI Too Much...

'Automation Addiction': Are Pilots Forgetting How to Fly?

Is auto-pilot weakening response time to emergency situations?

The Dangers of Overreliance on Automation

Safety Concerns and Mitigation Strategies for Pilots

FAA Safety Briefing Magazine Follow 8 min read · May 2, 2025

GPS use negatively affects environmental learning through spatial transformation abilities

Ian T. Ruginski[a,b,*], Sarah H. Creem-Regehr[a], Jeanine K. Stefanucci[a], Elizabeth Cashdan[b]

[a] *University of Utah, Department of Psychology, United States*
[b] *University of Utah, Department of Anthropology, United States*

When systems are always on, skills can quietly decay.

Two failure modes of AI advice:

Today's Talk

Should AI tell people exactly what to do or teach them how to think?

Two failure modes of AI advice:

- **Humans don't follow it enough** → skill preserved but performance left on the table
- **Humans follow it too much** → performance today, but capability erodes tomorrow

How should AI systems be designed for both performance and long-run capability?

Today's Talk

Should AI tell people exactly what to do or teach them how to think?

Precise or Broad? Designing Algorithmic Advice for Learning in Sequential Decision Making

Philippe Blaettchen
Lee Kong Chian School of Business, Singapore Management University, pblaettchen@smu.edu.sg
Wichinpong Park Sinchaisri
Haas School of Business, University of California, Berkeley, parksinchaisri@berkeley.edu


Outline: Stylized Model → Hypotheses/Predictions

Experimental Design: EV Driving Game

Study 1: Basic Pattern

Study 2: Mechanisms + Boundary Conditions

Mechanism: Inverse RL

Next Step: Compliance-Aware RL

Preview: Precise advice wins today. Broad advice wins tomorrow. Which matters more depends on your environment.

Managing Human Agents with Compliance-Aware Reinforcement Learning

BRYCE MCLAUGHLIN, University of Pennsylvania, USA
WICHINPONG PARK SINCHAISRI, Haas School of Business, University of California, Berkeley, USA


Advice Precision is Everywhere

	Precise	Broad
ZARA	"Order up to S"	"As underage cost rises, raise the target fractile"
Uber Grab	"Go to Zone 4 now"	"Stay close to the high-demand corridor"
Alibaba amazon	"Pack items in sequence A-B-C"	"Place heavy, flat items first"
Singapore General Hospital SingHealth	"Escalate this patient now"	"Prioritize cases with high risk and high delay cost"

Stylized Model

Environment changes by $\delta \in [0, 1]$

E1: advice available → E2: advice removed

- Designer chooses advice type **(precise / broad)**
- User chooses effort $e_1 \in [0, 1]$ with cost $c(e_1) = \frac{k}{2}e_1^2$ with $k > 0$.
- P(choosing the best action)

$$\pi_a^1(e_1) = \alpha_a + \beta_a e_1$$

- Precise is easier to follow

$$0 \leq \alpha_b < \alpha_p < 1$$

- Broad better converts effort into understanding

$$0 \leq \beta_p < \beta_b < 1$$

- User chooses effort $e_2 \in [0, 1]$ with cost $c(e_2) = \frac{k}{2}e_2^2$
- P(choosing the best action)

$$\pi_a^2(e_1, e_2) = e_2\left[\lambda + (1-\delta)^2\alpha_a + (1-\delta)\beta_a e_1\right]$$

direct action-level carryover

higher-order transferable understanding

Objective: Choose advice precision to balance rewards in E1 and E2

(Some) Model Results

Precise improves immediate performance

Easier to implement directly

Higher compliance while advice is present

Stronger contemporaneous performance in E1

Broad can create a later learning advantage

Broad requires interpretation, inducing more effort in E1

When E2 is sufficiently different, higher-order understanding matters more

If task sequential rather than one-shot, learning advantage increases with horizon length (more exploration)

Broad's advantage is non-monotonic

Too little change: precise still better

Moderage change: broad can dominate

Too much change: neither advice transfers well, so precise's short-term benefit is more important

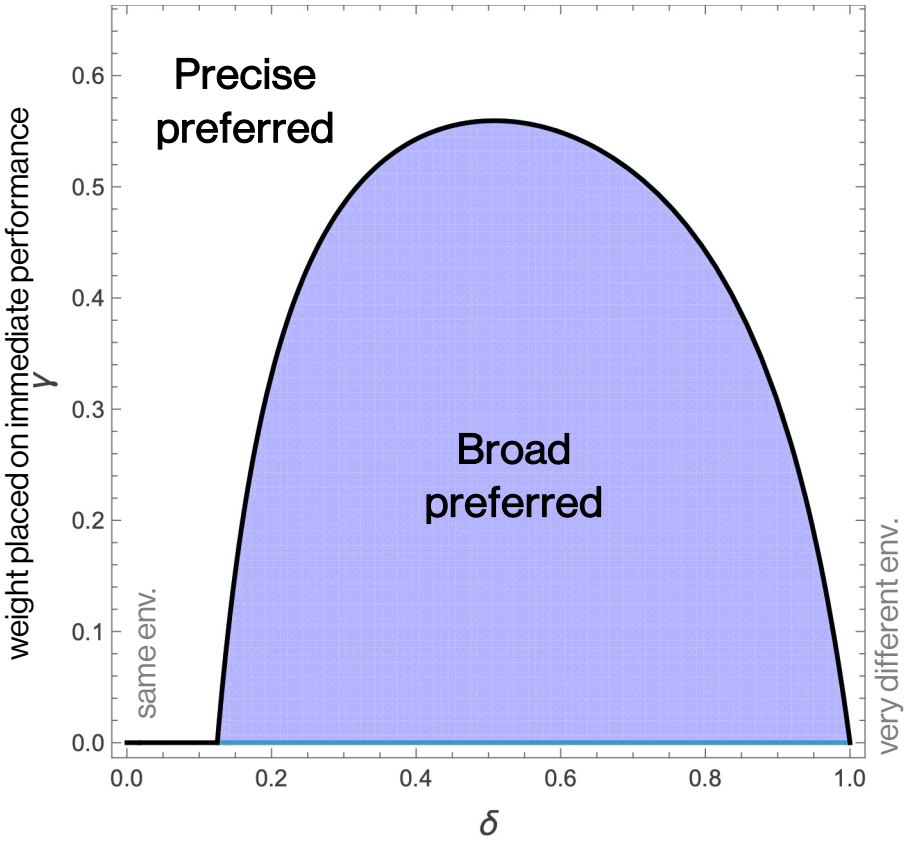
The Reward-Learning Frontier

Reward Gap: Δ_1
Immediate benefit of precise:
easier to implement,
stronger E1 performance

Learning Gap: $\Delta_2(\delta)$
Long-run benefit of broad:
more portable understanding
depends on environmental difference

weight placed on
immediate performance

$$J(p) - J(b) = \gamma \Delta_1 - (1-\gamma)\,\Delta_2(\delta)$$



PROPOSITION 1 **(Optimal Advice)**. *Define*

$$\gamma^\star(\delta) = \begin{cases} 0, & \text{if } \delta \in [0, \delta_0], \\ \frac{\Delta_2(\delta)}{\Delta_1 + \Delta_2(\delta)}, & \text{if } \delta \in (\delta_0, 1]. \end{cases}$$

Then the designer optimally chooses b if and only if $\gamma < \gamma^\star(\delta)$.

pre-registered at https://aspredicted.org/7LP_J9G

Hypotheses/Predictions

E1: advice available → E2: advice removed

H1a – advice compliance:
Compliance is lower under broad than under precise

H1b – reward gap:
Performance is higher under precise than under broad

H2 – exploration: Users with broad advice visit more different states

H3a – post-advice strategy:
Without advice, strategies are closer to the optimal one after getting broad vs precise advice in E1
⟺ E1 and E2 are different

H3b – learning gap:
Performance is higher under broad than precise ⟺ E1 and E2 are substantially different.

Experimental setting: Sequential decision-making task that involves uncertainty and future changes in the environment

Today's Talk

Precise or Broad? Designing Algorithmic Advice for Learning in Sequential Decision Making

Philippe Blaettchen
Lee Kong Chian School of Business, Singapore Management University, pblaettchen@smu.edu.sg
Wichinpong Park Sinchaisri
Haas School of Business, University of California, Berkeley, parksinchaisri@berkeley.edu


Outline: Stylized Model → Hypotheses/Predictions

Experimental Design: EV Driving Game

Study 1: Basic Pattern

Study 2: Mechanisms + Boundary Conditions

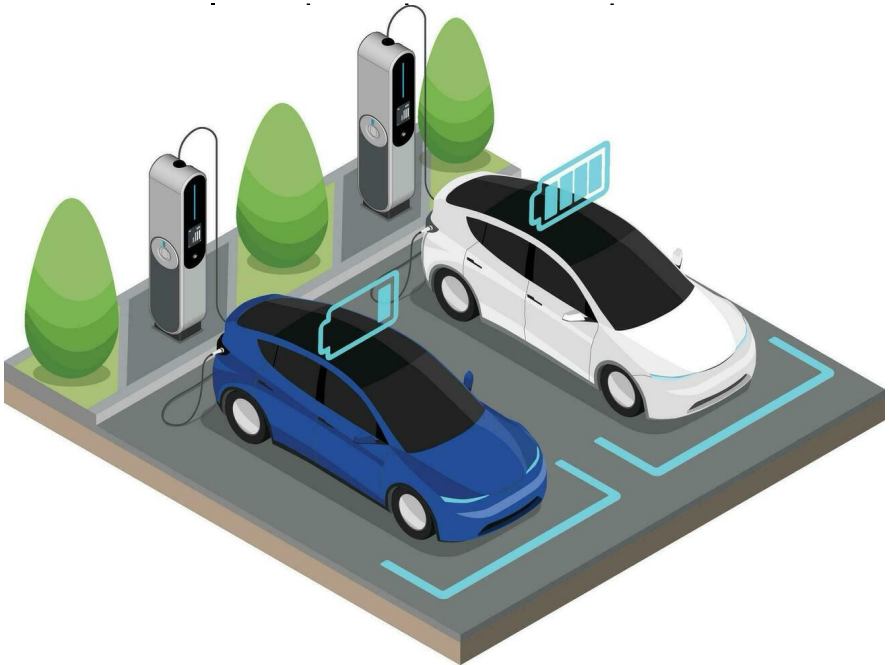
Mechanism: Inverse RL

Next Step: Compliance-Aware RL

Sequential Task

Why sequential tasks?

- current choices enable or constrain later choices
- small behavioral differences can compound over time



Why EV routing/charging?

- Sequentiality: charging now changes future charging needs
- Non-trivial optimal policy: requires genuine forward planning, not just local myopia
- We can solve the MDP exactly → measure every decision against optimal
- Continuous action space: participants choose any amount 0-100% → rich behavioral data
- Traffic uncertainty makes risk management essential

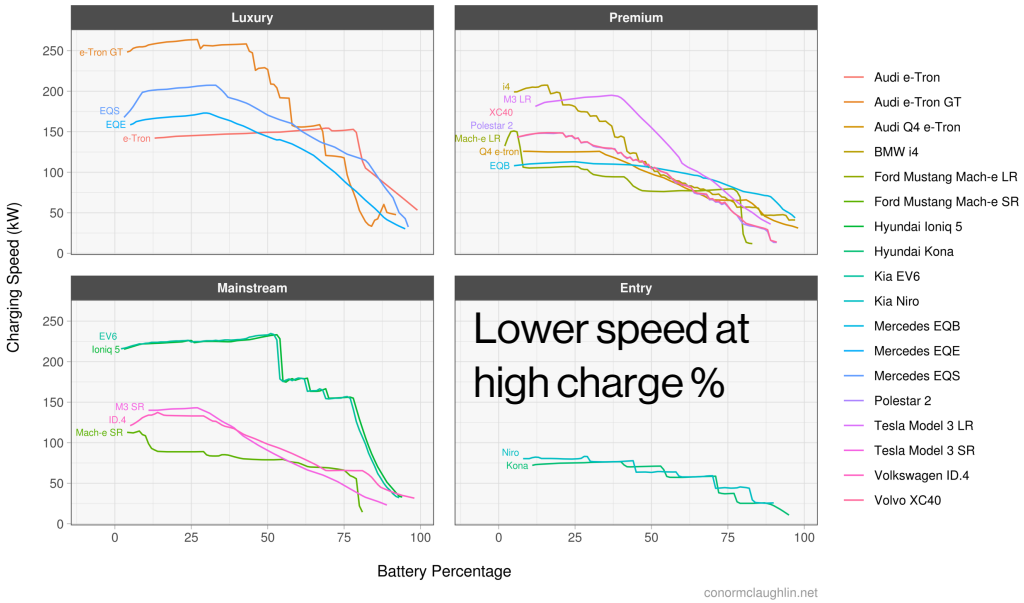
Why EV Charging?

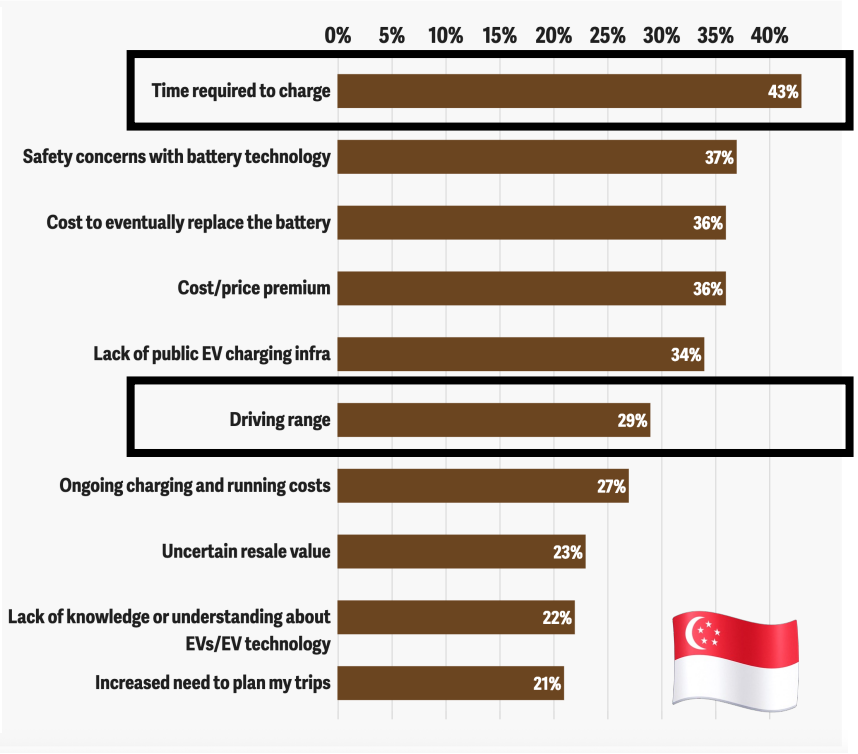
Deloitte.

2026 Global Automotive Consumer Study

Electric Vehicle (EV) Charge Curves

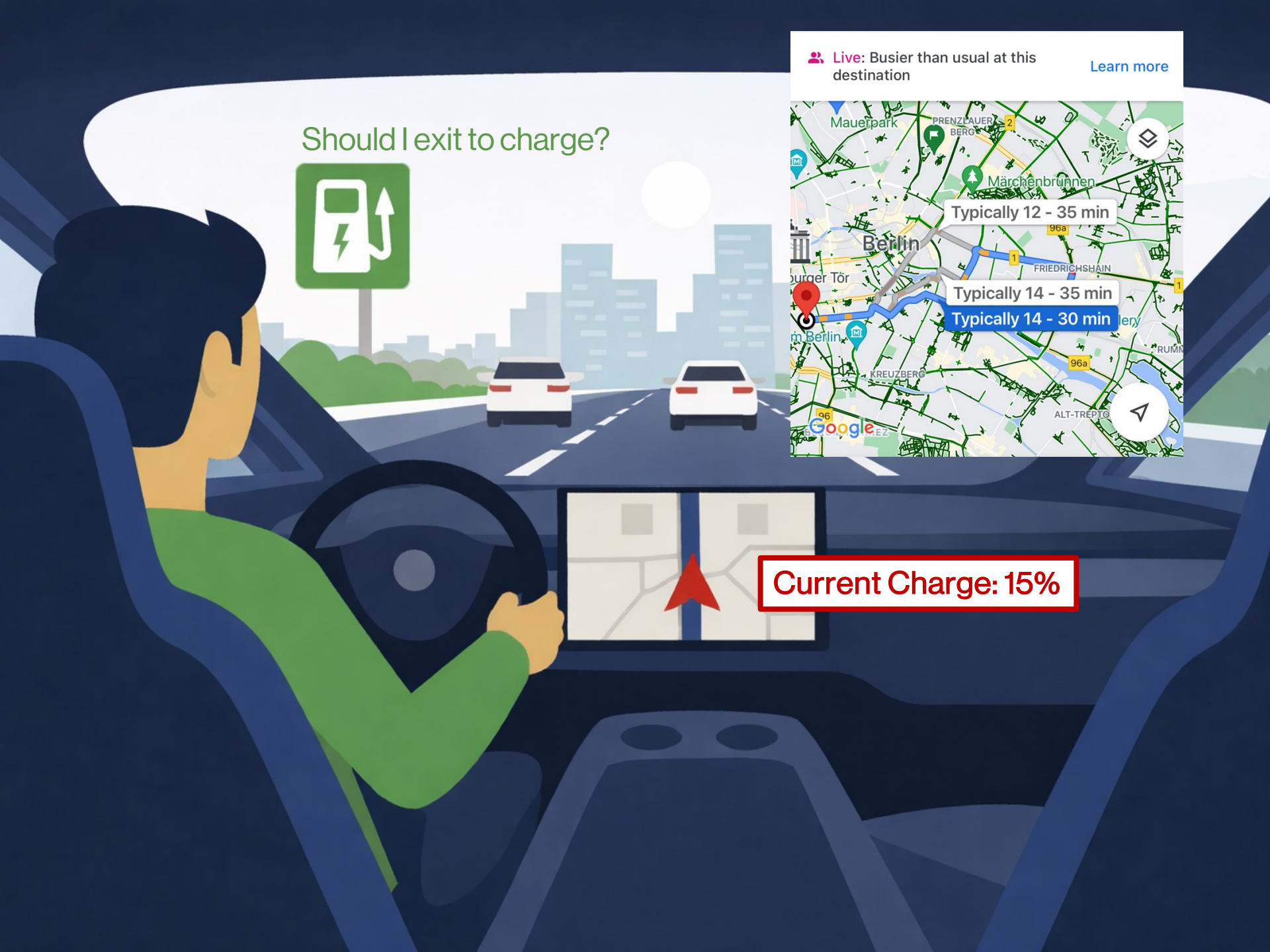
For EV charging, power delivery is non-uniform over the duration of the session. Rather, it generally follows a curve: maximum power deliver happens when the battery level is relatively low, and power delivery tapers off as the battery becomes increasingly full. However, there is lots of differentiation in the shape of these curves - certain battery architectures have high speeds, then a significant dropoff, while others look to achieve a more stable rate throughout.



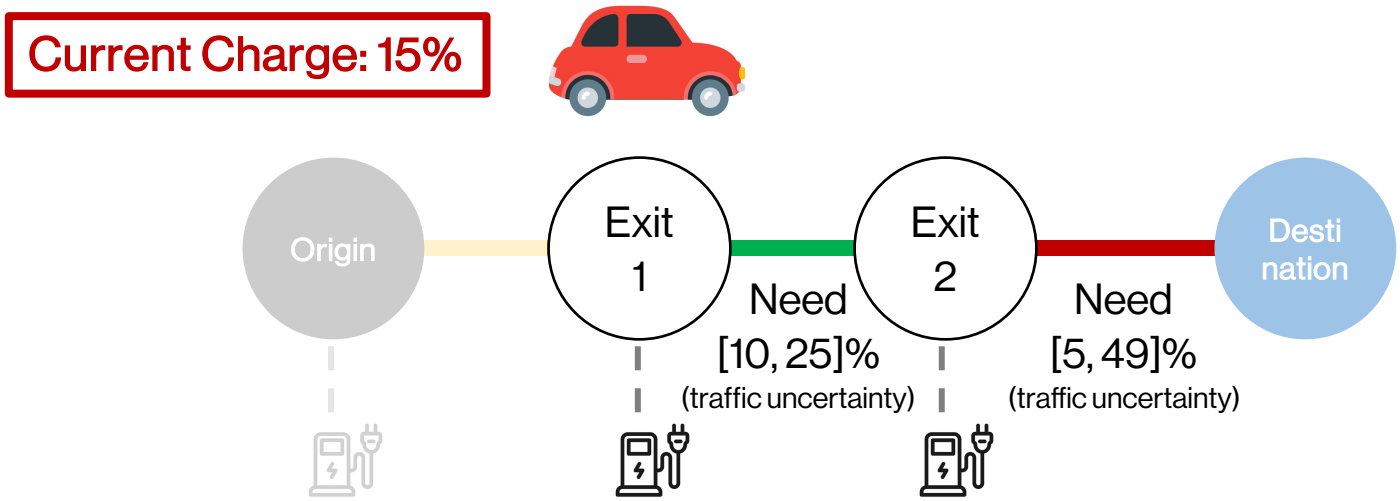


This is not a paper about EV adoption! It's about sequential decision-making, and EVs give us a setting where the stakes feel real and the uncertainty is relatable. No real physics! 🙏

https://www.deloitte.com/us/en/insights/industry/retail-distribution/global-automotive-consumer-study.html


Should I exit to charge?
Live: Busier than usual at this destination
Learn more
Typically 12 - 35 min
Typically 14 - 35 min
Typically 14 - 30 min
Current Charge: 15%

In the Driver's Seat



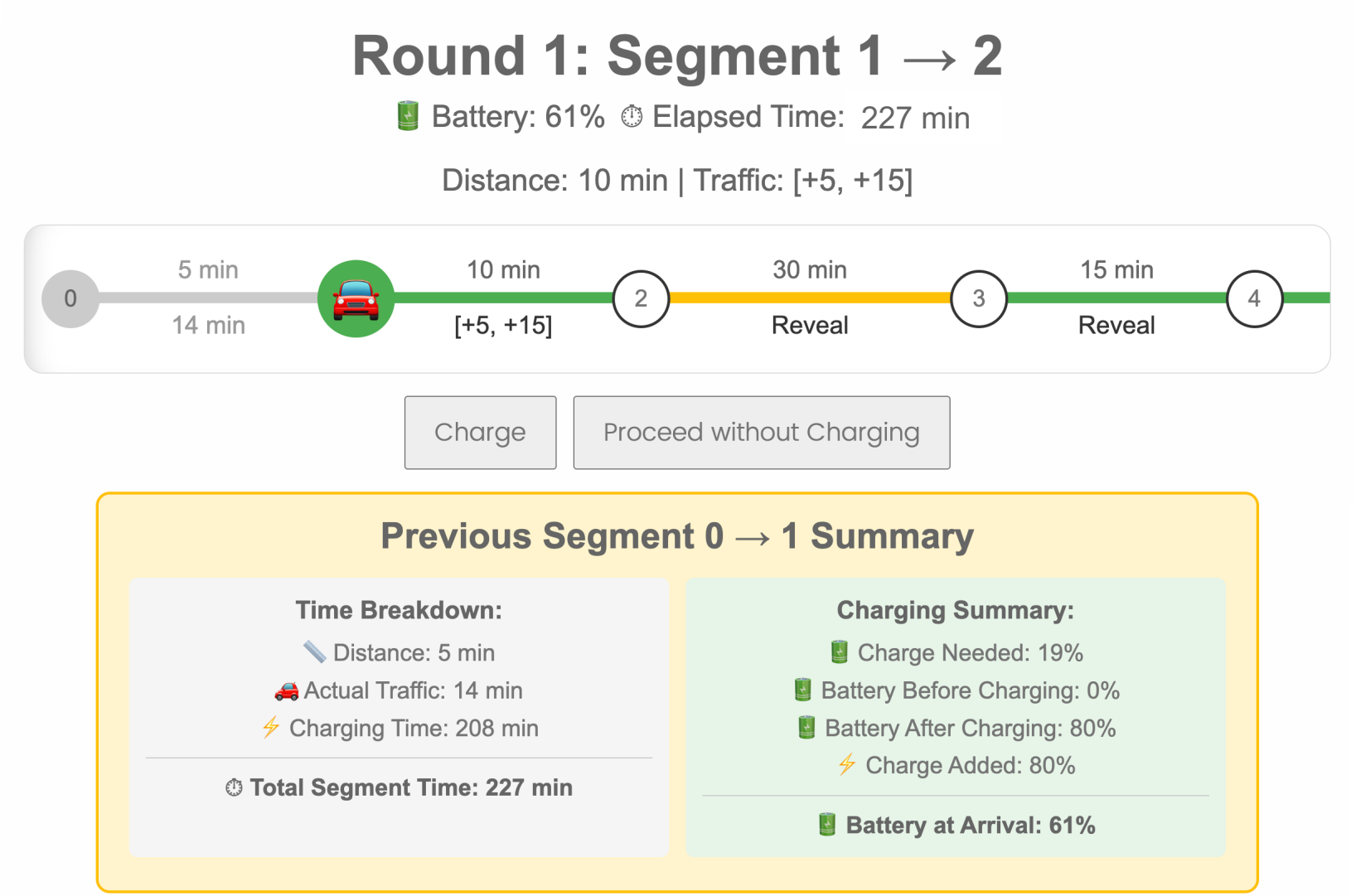
- For simplicity, 1% of battery = 1 minute of driving
- Facing uncertain traffic; **exit to charge or go ahead?**
- If you exit: fixed overhead (30 mins) + charging time

  Charging time is non-linear (convex)
- If you run out, large penalty (300 mins)

pre-registered at https://aspredicted.org/7LP_J9G

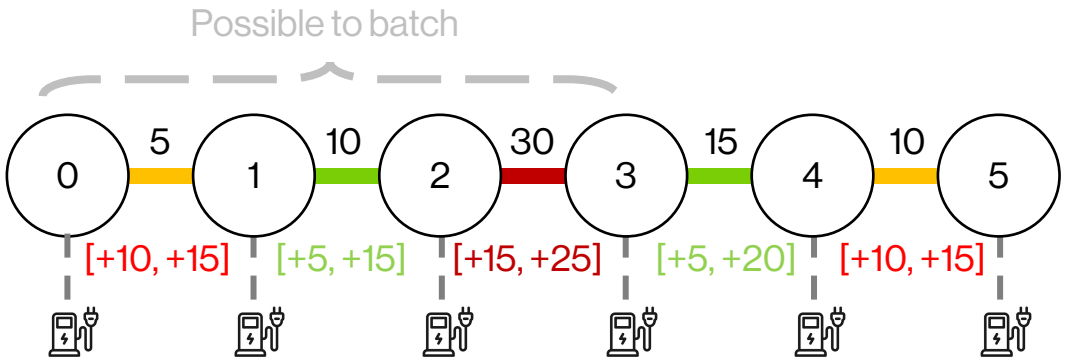
Study 1

Design: Task/Interface



Study 1

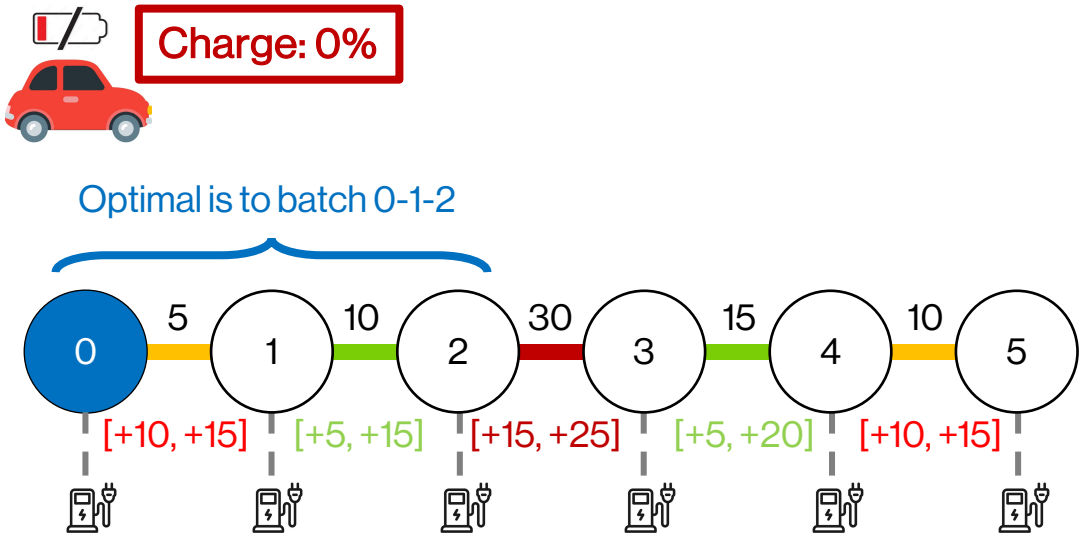
Design: Batching



pre-registered at https://aspredicted.org/7LP_J9G

Study 1

Design: Batching

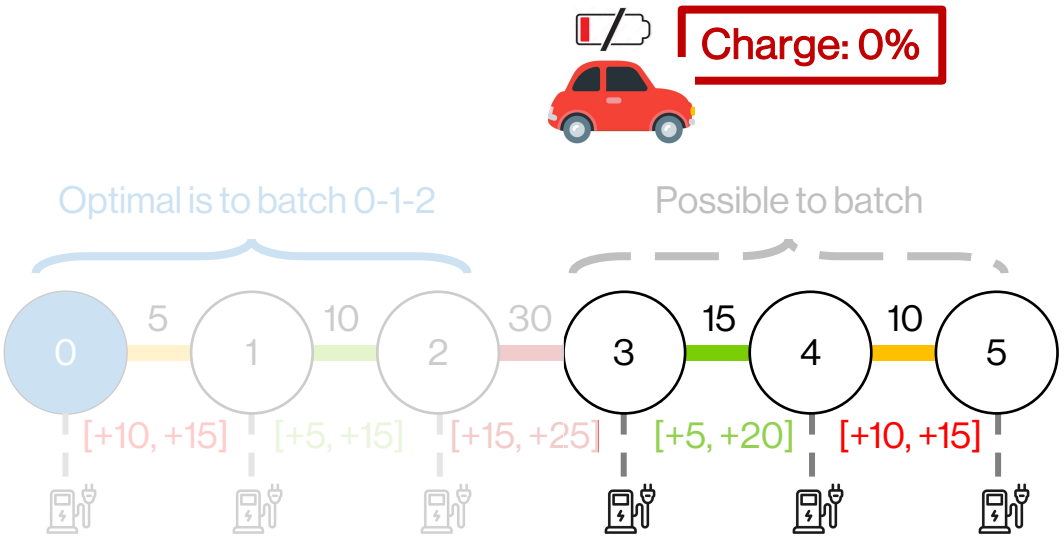


0

Optimal = **"batch"** required charges
for the next two stops (0 → 2)
rather than just 0 → 1 or
further batch 0 → 3.

Study 1

Design: Batching



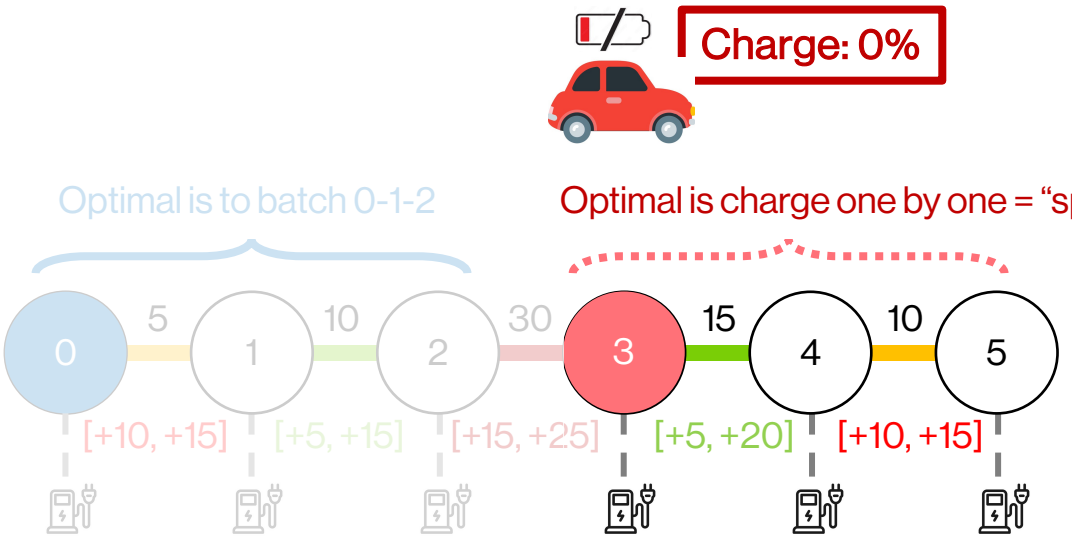
0

Optimal = "batch" required charges
for the next two stops (0 → 2)
rather than just 0 → 1 or
further batch 0 → 3.

pre-registered at https://aspredicted.org/7LP_J9G

Study 1

Design: Batching



0

Optimal = **"batch"** required charges for the next two stops (0 → 2)

rather than just 0 → 1 or further batch 0 → 3.

3

Optimal = **"split"** = only charge for the next stop (3 → 4)

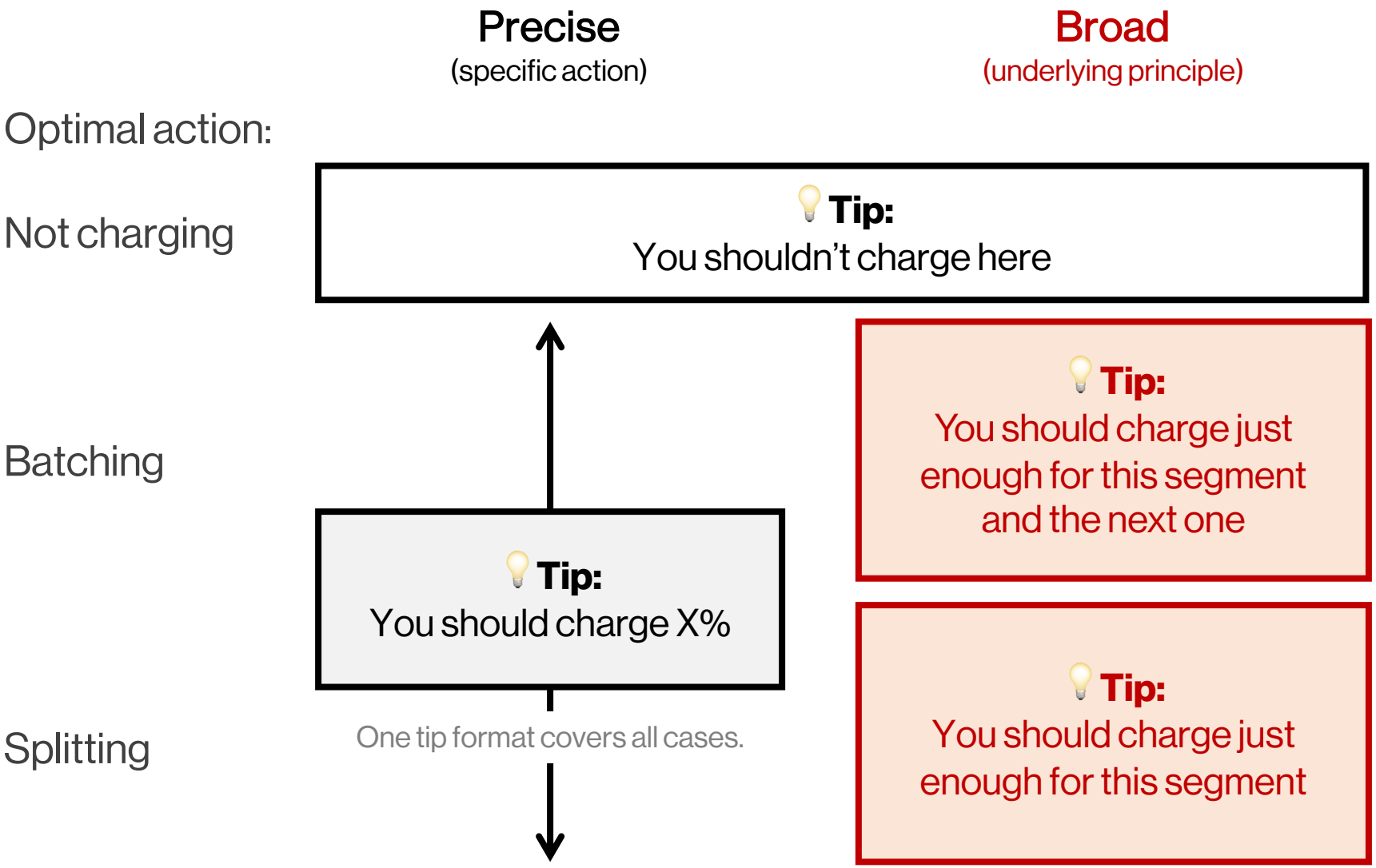
rather than batch 3 → 5.

The best action depends on traffic uncertainty, current charge, and where future charging opportunities lie.

pre-registered at https://aspredicted.org/7LP_J9G

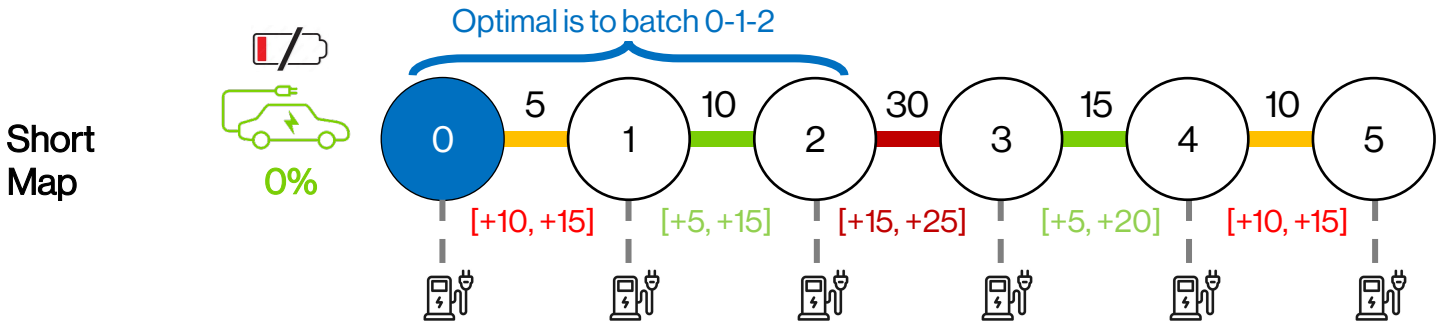
Study 1

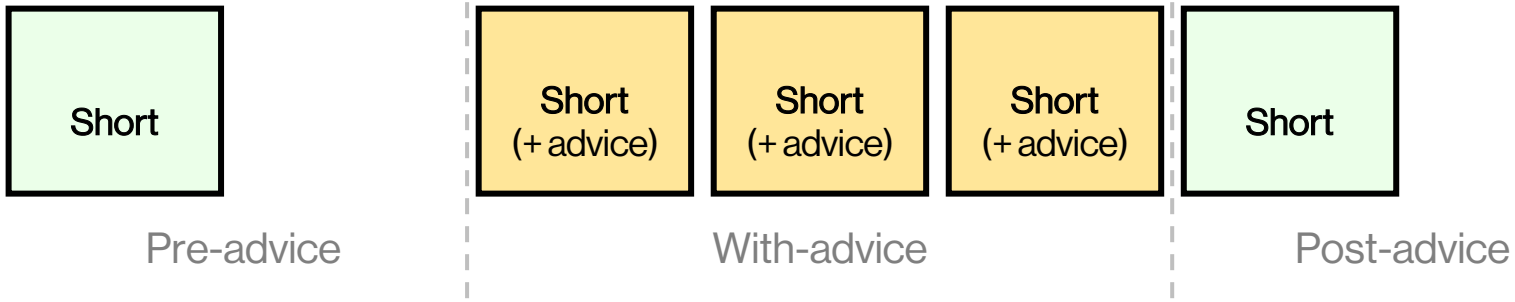
Precise vs Broad Advice



Study 1

Flow: Three-Phase Game

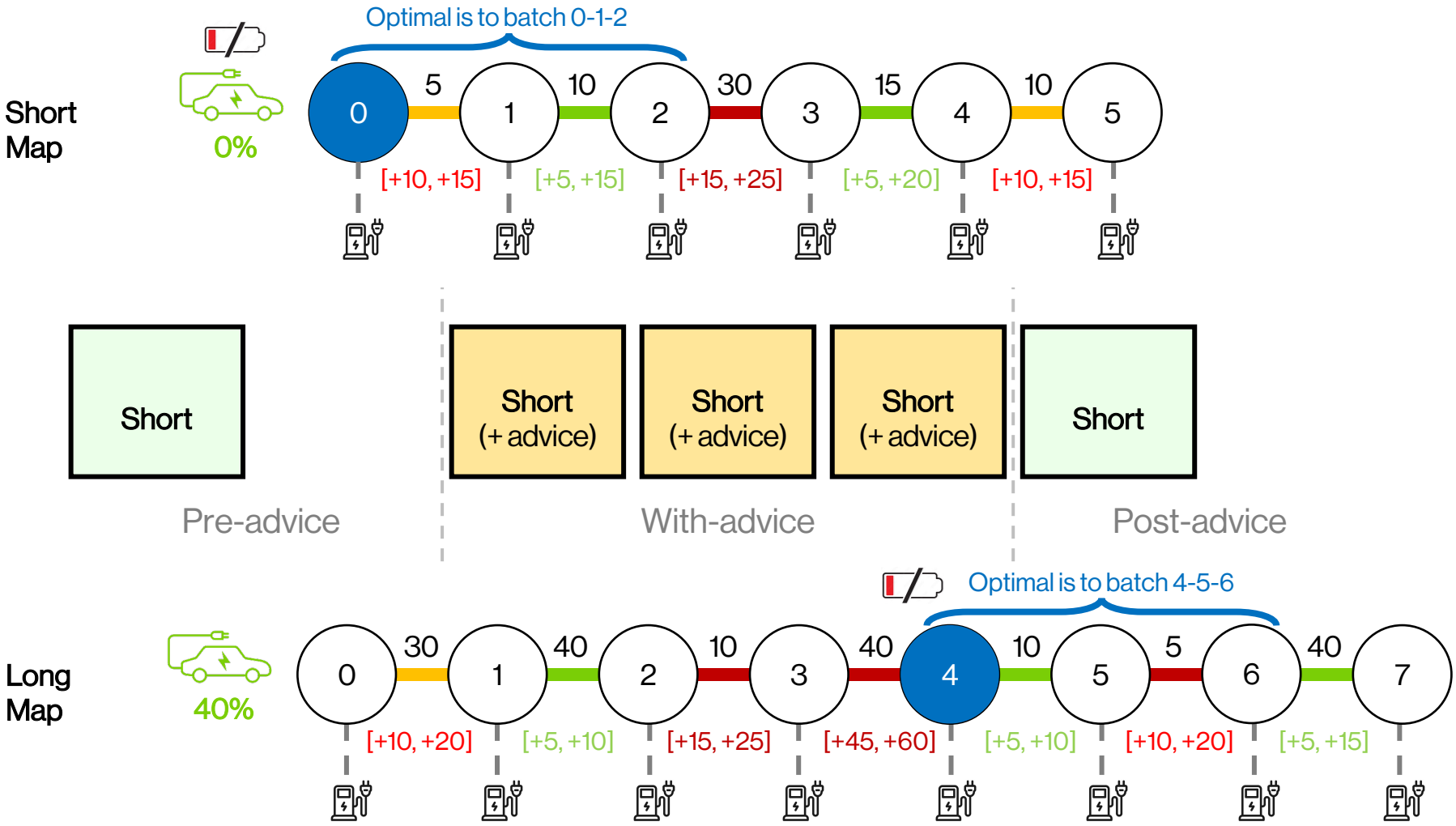




pre-registered at https://aspredicted.org/7LP_J9G

Study 1

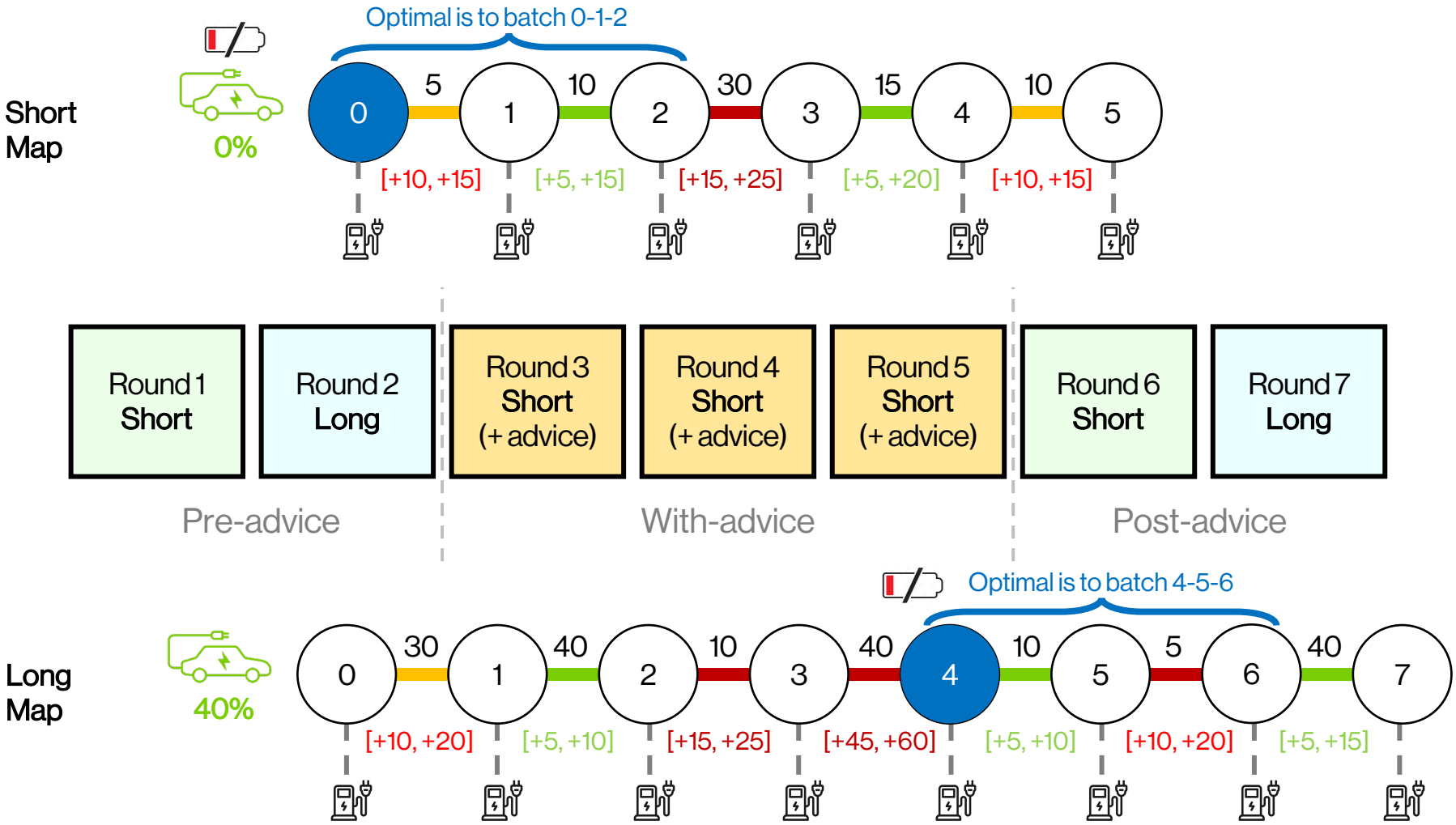
Flow: Three-Phase Game



pre-registered at https://aspredicted.org/7LP_J9G

Study 1

Flow: Three-Phase Game



pre-registered at https://aspredicted.org/7LP_J9G

Study 1

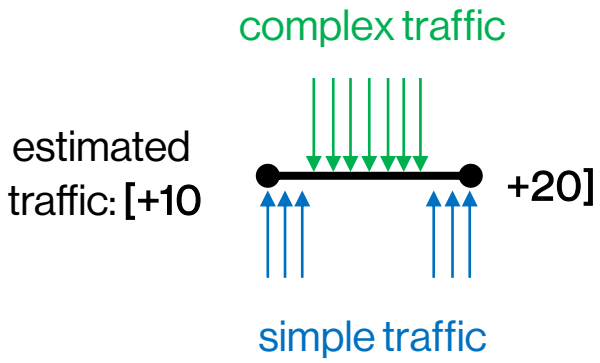
Treatment Conditions

- 2 x 2 factorial design
- Advice precision
  - Precise
  - Broad

Compliance + immediate performance in E1
+ post-advice (E2) performance

- Task difficulty
  - Simple traffic
  - Complex traffic

Complexity manipulation tests robustness:
do effects hold when the underlying task is harder?

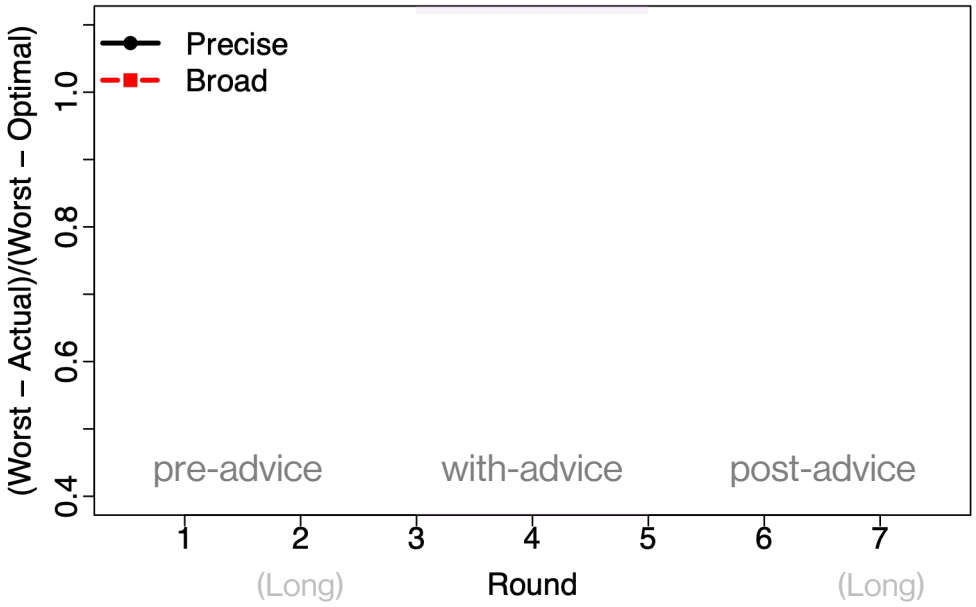


Amazon Mechanical Turk
N = 102, 3,978 decision points

pre-registered at https://aspredicted.org/7LP_J9G

Study 1

How We Measure Performance



Normalized Performance Score:
(Worst - Actual) / (Worst - Optimal)

Worst = completion time if made the worst possible decisions at every state
Optimal = the MDP optimal (shortest possible) completion time
Actual = completion time the participant actually got
Score of 1.0 = optimal play.
Score of 0 = worst possible play.

Higher = Better!

Amazon Mechanical Turk
N = 102, 3,978 decision points

pre-registered at https://aspredicted.org/7LP_J9G

Study 1

Result: Precise Works Instantly

In most organizations, the experiment would end here. The metrics look good. Precise advice is working.

While adv… the strongest contemporaneous gains.

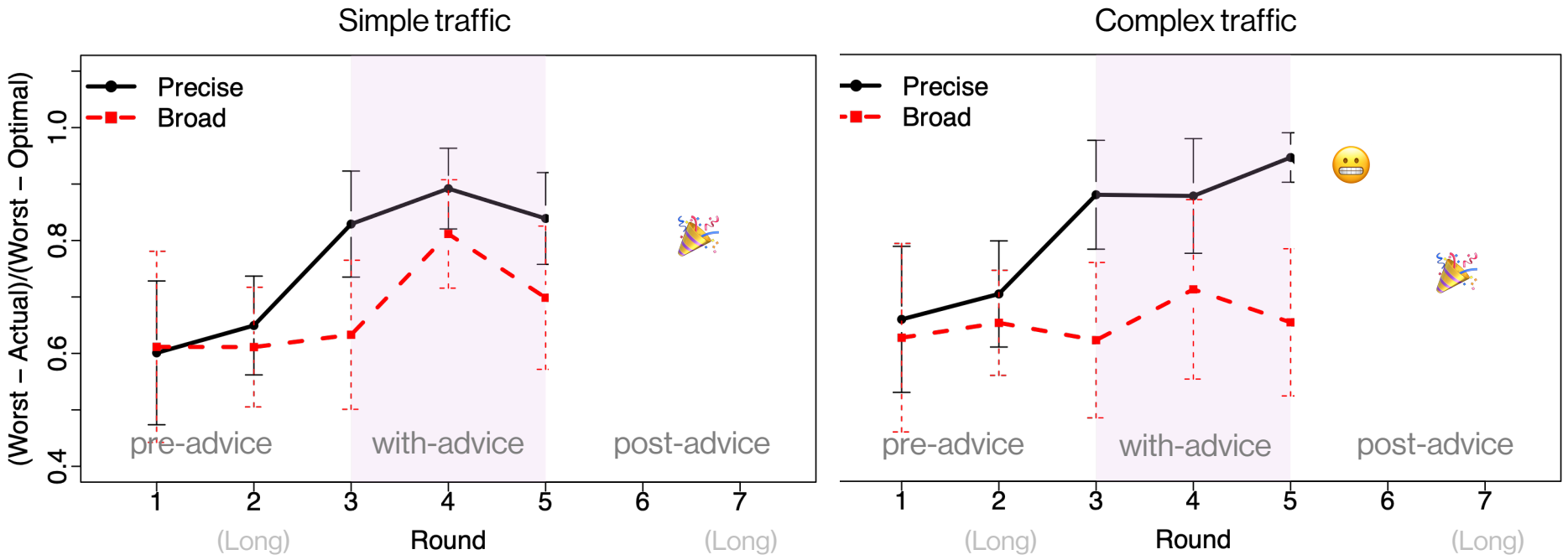
Amazon Mechanical Turk
N = 102, 3,978 decision points

…ment
compared to optimal
Higher = better

pre-registered at https://aspredicted.org/7LP_J9G

Study 1

After Advice is Removed...



While advice is available, precise advice yields the strongest contemporaneous gains.

Post-advice, the lines cross. Participants under broad outperform those under precise: most clearly under complex traffic and on the less familiar long map.

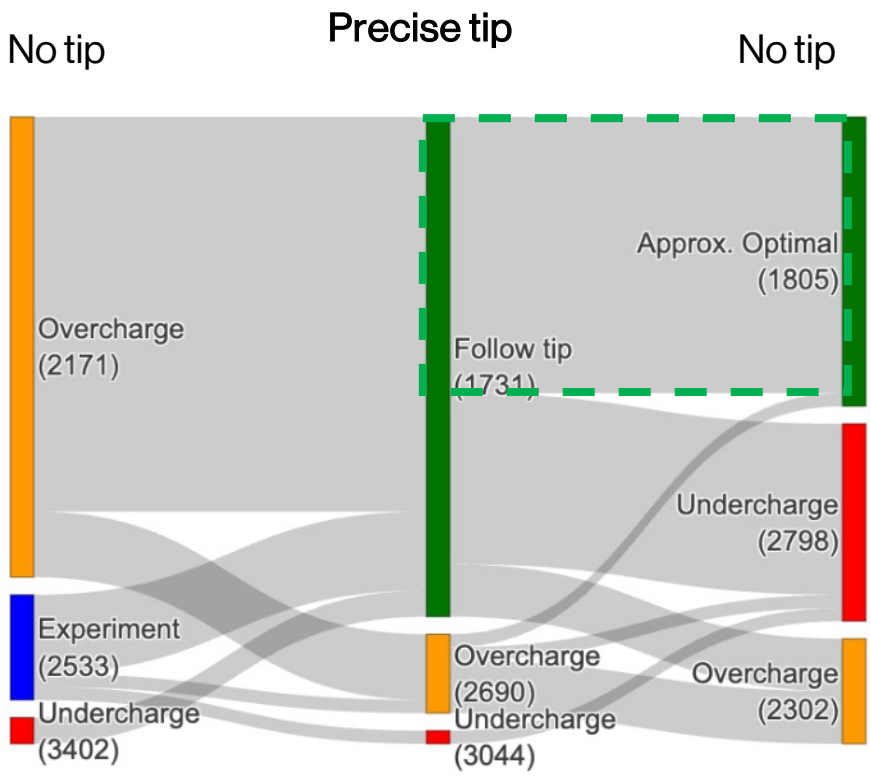
Y-axis = how much improvement compared to optimal
Higher = better

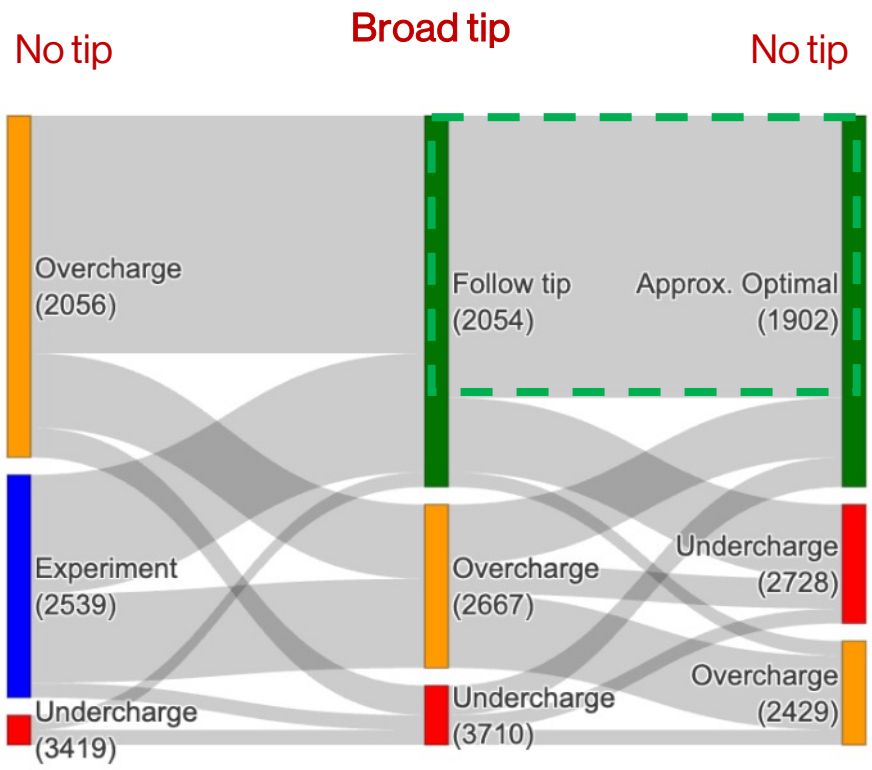
Amazon Mechanical Turk
N = 102, 3,978 decision points

pre-registered at https://aspredicted.org/7LP_J9G

Study 1

Why Might Broad Help Later?





Precise: **55%**
stay with optimal strategy afterwards

Broad: **76%**
stay with optimal strategy afterwards

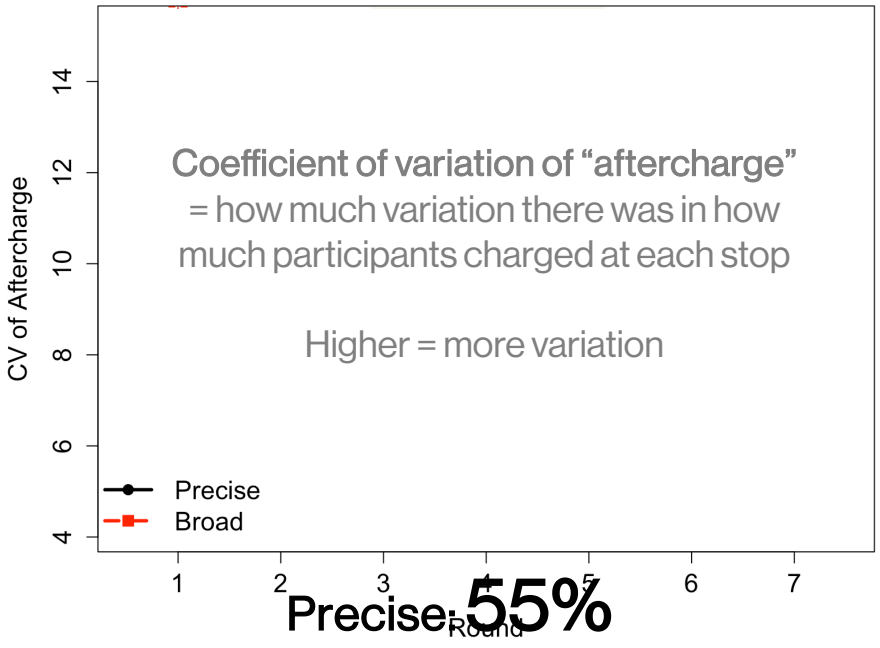
Amazon Mechanical Turk
N = 102, 3,978 decision points

pre-registered at https://aspredicted.org/7LP_J9G

Study 1

Why Broad = More Durable Strategy?

If broad advice promotes active learning: participants should explore more diverse strategies while receiving it (H2)



Precise: 55%
stay with optimal strategy afterwards

Broad: 76%
stay with optimal strategy afterwards

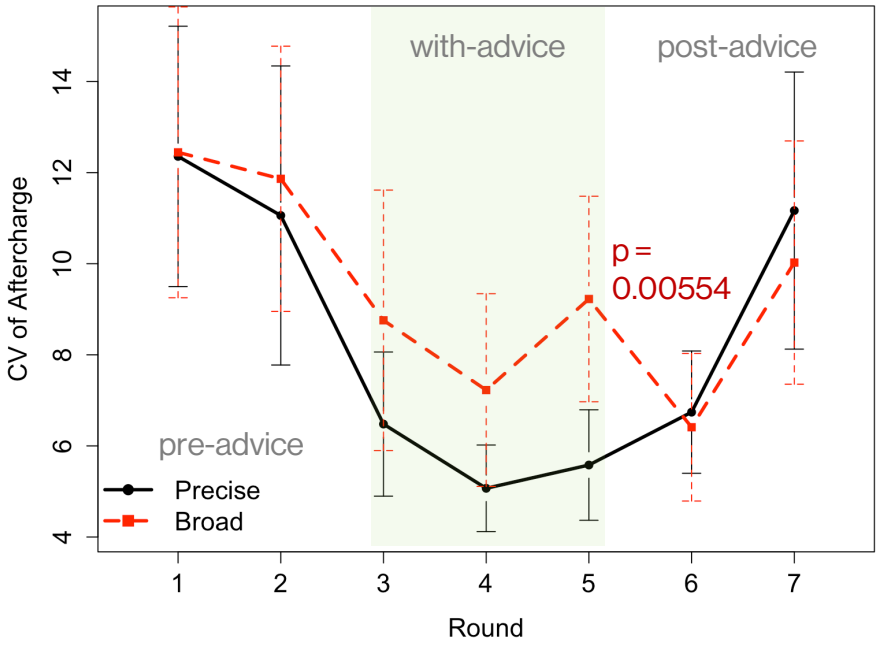
Amazon Mechanical Turk
N = 102, 3,978 decision points

pre-registered at https://aspredicted.org/7LP_J9G

Study 1

Why Broad = More Durable Strategy?

If broad advice promotes active learning: participants should explore more diverse strategies while receiving it (H2)



Broad requires interpretation

↓

Participant must reason about each decision independently

↓

Charging amounts stay varied (CV stays elevated through Round 5)

↓

More of the task's state space encountered

↓

Richer representation of trade-offs built

↓

Strategic knowledge persists post-advice

↓

Broad: 76%

stay with optimal strategy afterwards

Amazon Mechanical Turk
N = 102, 3,978 decision points

Beyond EV Charging

PNAS RESEARCH ARTICLE | ECONOMIC SCIENCES

Generative AI without guardrails can harm learning: Evidence from high school mathematics

Hamsa Bastani[a,b,1], Osbert Bastani[c,1], Alp Sungu[a,1,2], Haosen Ge[b], Özge Kabakcı[d], and Rei Mariman[e]

AI Meets the Classroom: When Do Large Language Models Harm Learning?

Matthias Lehmann
University of Cologne, matthias.lehmann@wiso.uni-koeln.de

Philipp B. Cornelius
Rotterdam School of Management, Erasmus University, cornelius@rsm.nl

Fabian J. Sting
University of Cologne, Rotterdam School of Management, Erasmus University, sting@wiso

Action vs. Attention Signals for Human-AI Collaboration: Evidence from Chess

Stefanos Poulidis
INSEAD, Decision Sciences, stefanos.poulidis@insead.edu

Haosen Ge
Wharton School, AI & Analytics Initiative, hge@wharton.upenn.edu

Hamsa Bastani
Wharton School, Operations Information and Decisions, hamsab@wharton.upenn.edu

Osbert Bastani
University of Pennsylvania, Computer and Information Science, obastani@seas.upenn.edu

Study 1: Pattern is clear.

✓ Precise = higher compliance + performance

✓ Broad keeps users cognitively active

✓ Broad produces more durable strategies

The performance learning gap in Round 7 is directional and consistent, but we can't yet say definitively:

At what level of environmental change does the broad advantage materialize, and where does it disappear?

Study 2 is designed to answer that precisely.

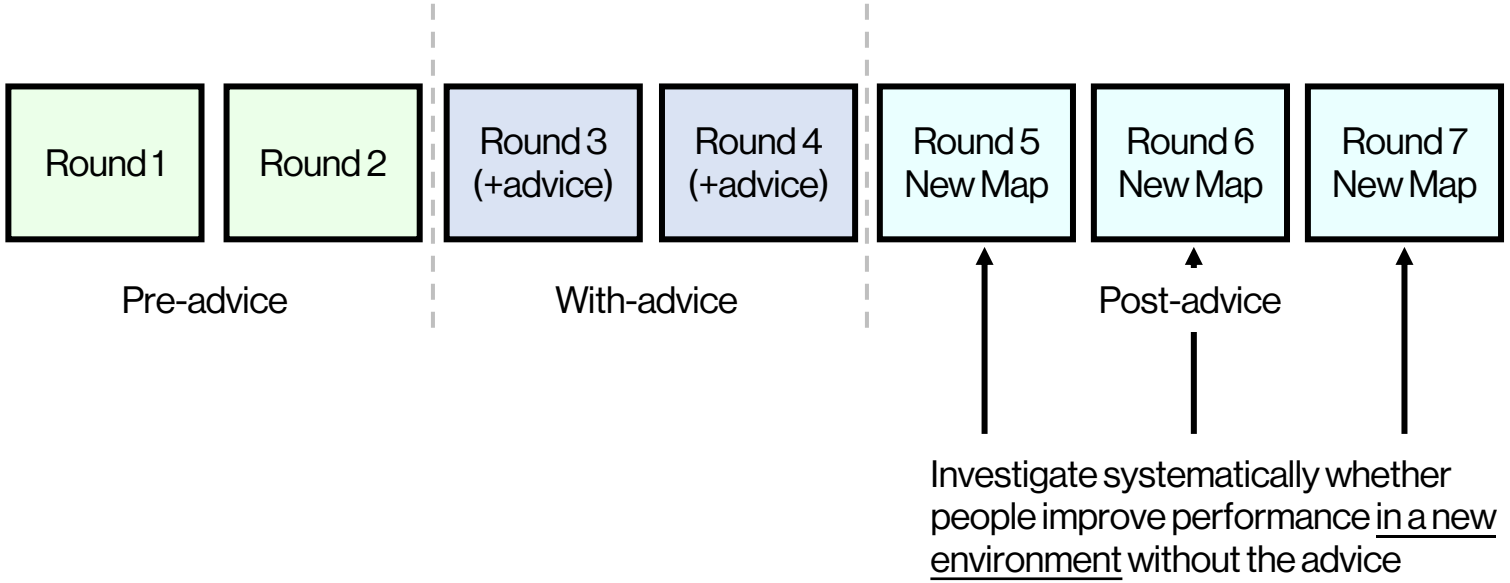
Where Does Transfer Break Down?

δ low (near transfer)		δ moderate (substantial)		δ high (too different)
Same map	→	Same structure, diff. params	→	Completely new map
Broad ≈ Precise		Broad > Precise		Broad ≈ Precise
Study 1 Round 6 Exact same map as the with-advice rounds		Study 1 Round 7 Experienced Long map in Round 2 before		**Study 2 will introduce never-before-seen maps**

pre-registered at https://aspredicted.org/23C_WBQ

Study 2

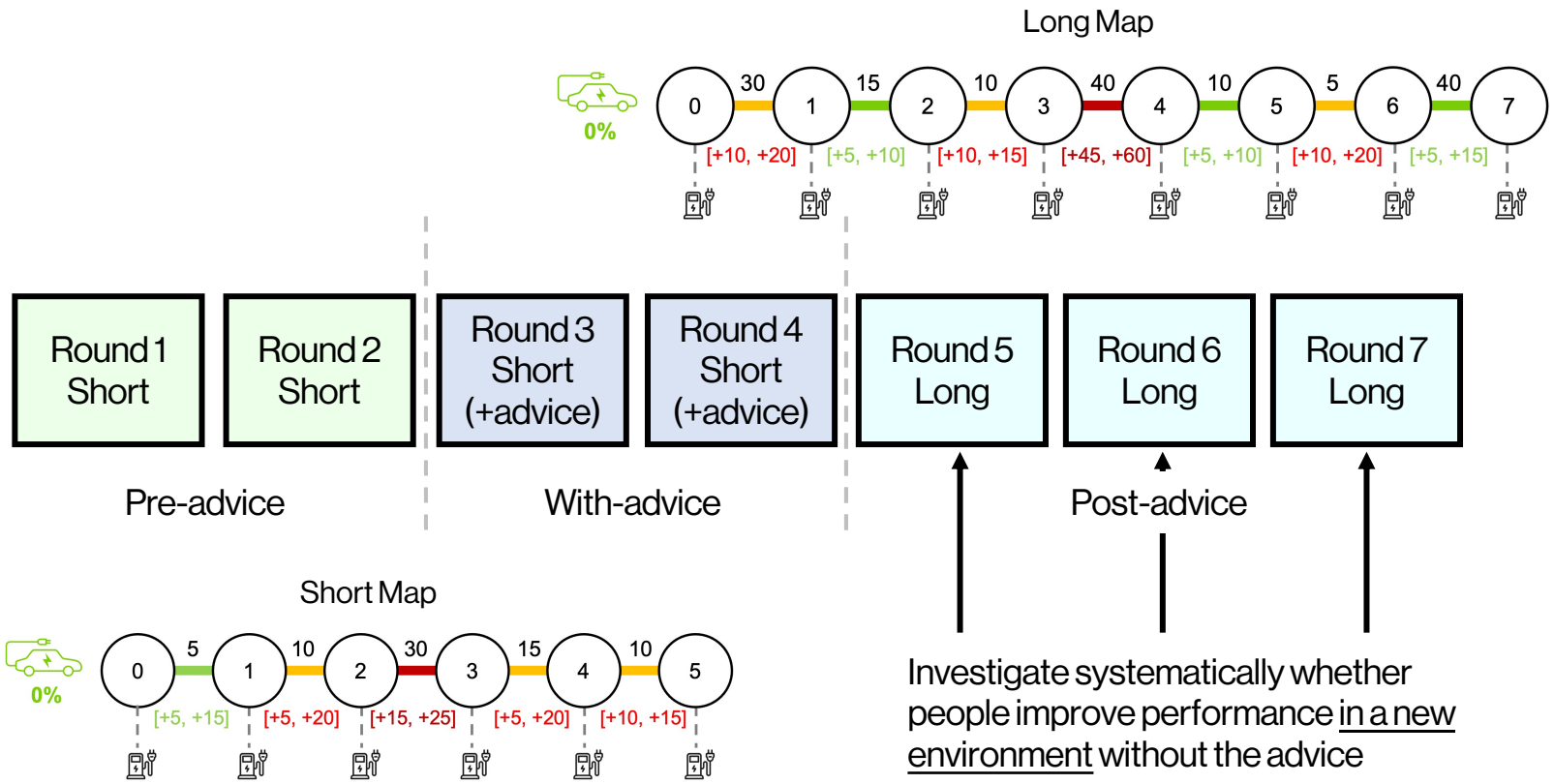
Familiar/Unseen Environments



pre-registered at https://aspredicted.org/23C_WBQ

Study 2

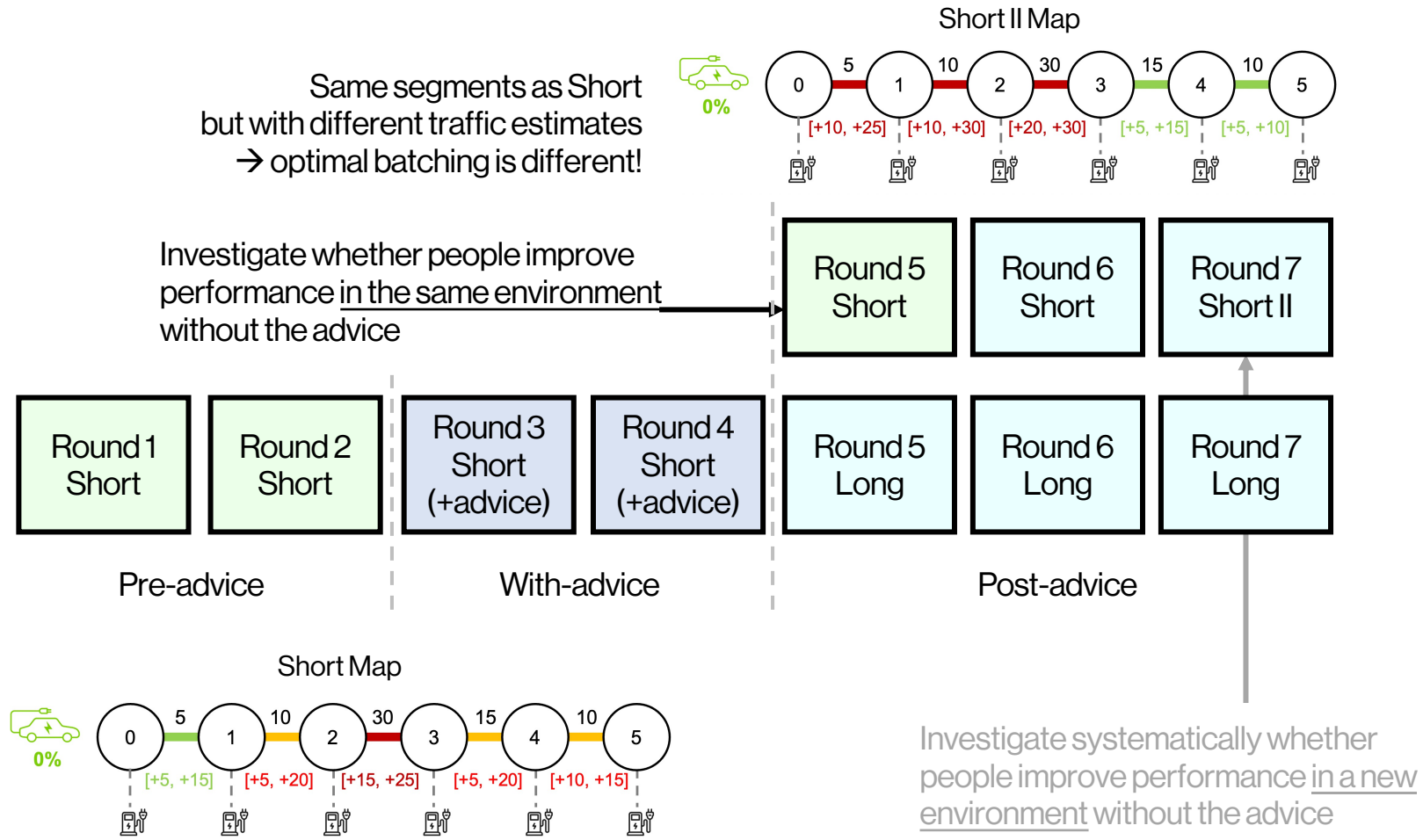
Familiar/Unseen Environments



pre-registered at https://aspredicted.org/23C_WBQ

Study 2

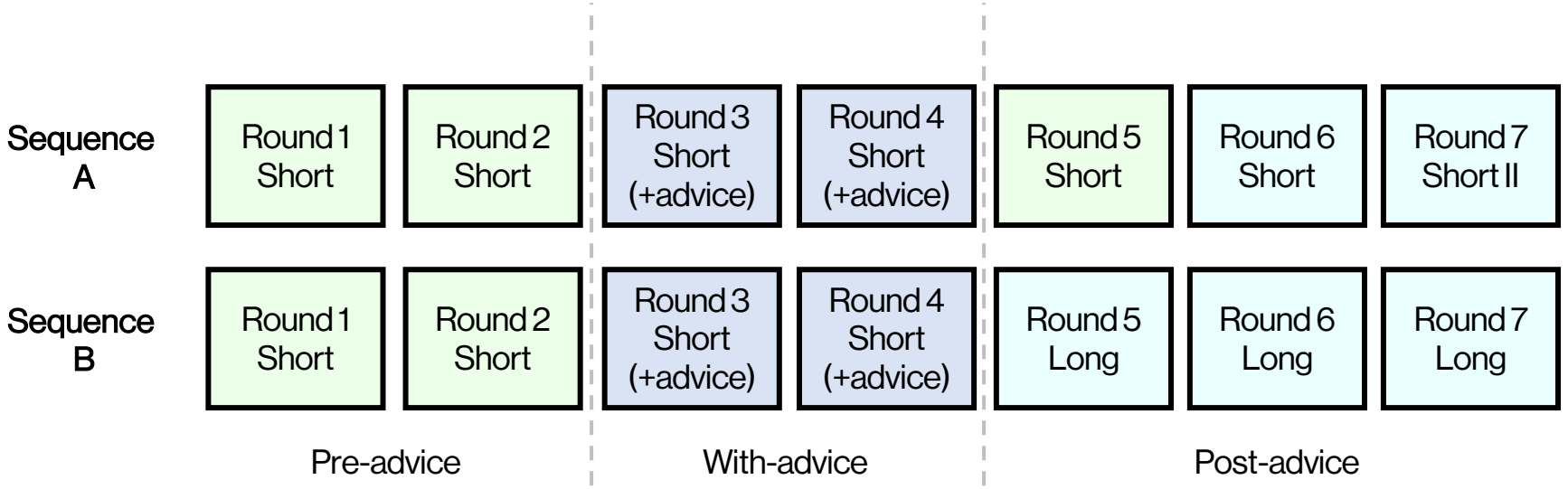
Familiar/Unseen Environments



pre-registered at https://aspredicted.org/23C_WBQ

Study 2

Familiar/Unseen Environments



Broad works because it forces users to translate a principle into an action. The compliance cost is also because of that translation burden.

#1: Reduce the translation burden:

Add specificity to broad advice. Does partial specificity preserve learning while recovering some compliance?

#2: Bypass the translation

Add an explanation to precise advice. Can precise + explanation match broad's learning effect?

Prior work is mixed (Bader et al. 2011; Buçinca et al. 2021; Ghai et al 2020): They can help calibration trust + engage users, but also add cognitive load + anchor judgment

pre-registered at https://aspredicted.org/23C_WBQ

Study 2

Making Broad More Specific

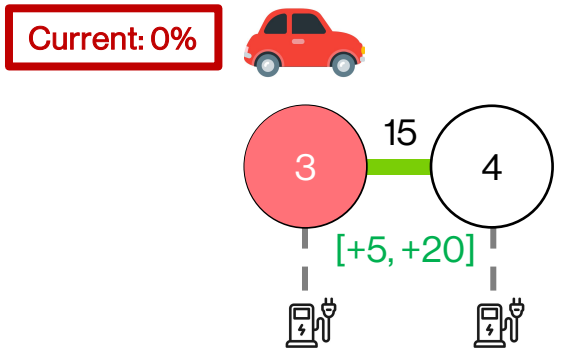
narrowing the action/
lowering cognitive load

	Precise	Specific Broad	Broad
Batching	💡 **Tip:** You should charge X%	💡 **Tip:** You should charge just enough for this segment and the next one, assuming worst case traffic	💡 **Tip:** You should charge just enough for this segment and the next one
Splitting		💡 **Tip:** You should charge just enough for this segment , assuming worst case traffic	💡 **Tip:** You should charge just enough for this segment

pre-registered at https://aspredicted.org/23C_WBQ

Study 2

Making Broad More Specific



	Precise	Specific Broad	Broad
Splitting	💡 **Tip:** You should charge **35%**	💡 **Tip:** You should charge just enough for this segment, assuming worst case traffic	💡 **Tip:** You should charge just enough for this segment

These two tips lead to the same charge %!

pre-registered at https://aspredicted.org/23C_WBQ

Study 2

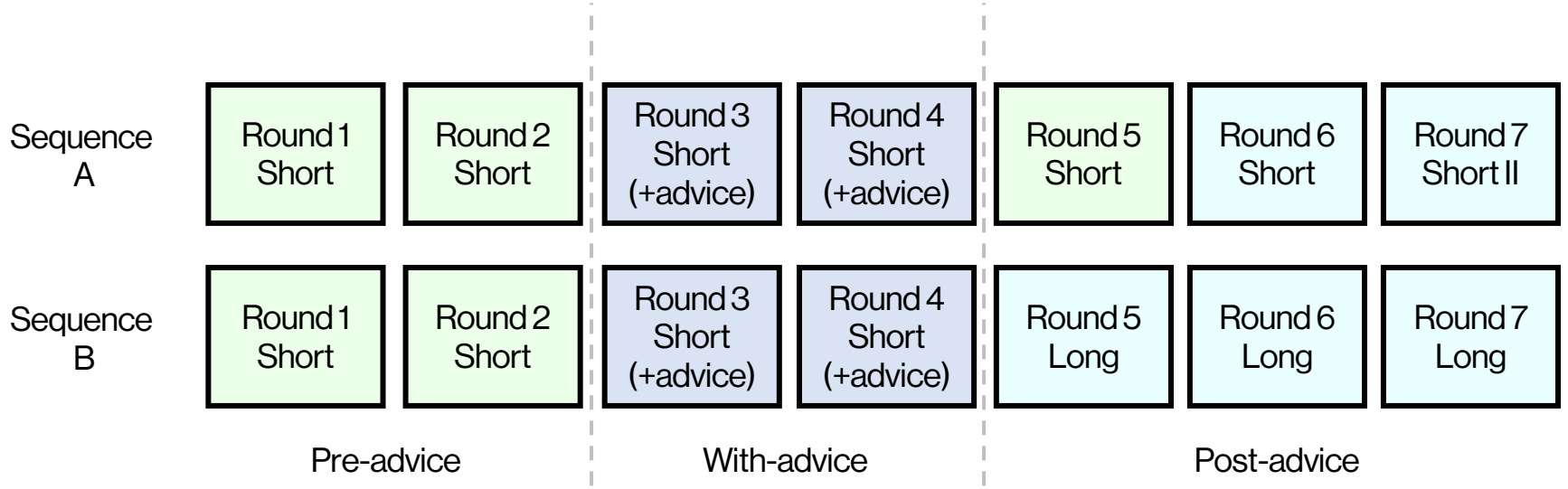
Adding a Rationale

	Precise	Broad
Batching	💡 **Tip:** You should charge X%. This minimizes total charging time by reducing the number of stops, since exiting incurs a 30-minute overhead.	💡 **Tip:** You should charge just enough for this segment and the next one. This minimizes total charging time by reducing the number of stops, since exiting incurs a 30-minute overhead.
Splitting	💡 **Tip:** You should charge X%. This minimizes total charging time, because charging is faster when the total charge is small.	💡 **Tip:** You should charge just enough for this segment. This minimizes total charging time, because charging is faster when the total charge is small.

pre-registered at https://aspredicted.org/23C_WBQ

Study 2

Treatment Conditions



2 map sequences x 3 advice precision x 2 rationale reveal

precise
specific broad
broad

Prolific: N = 400

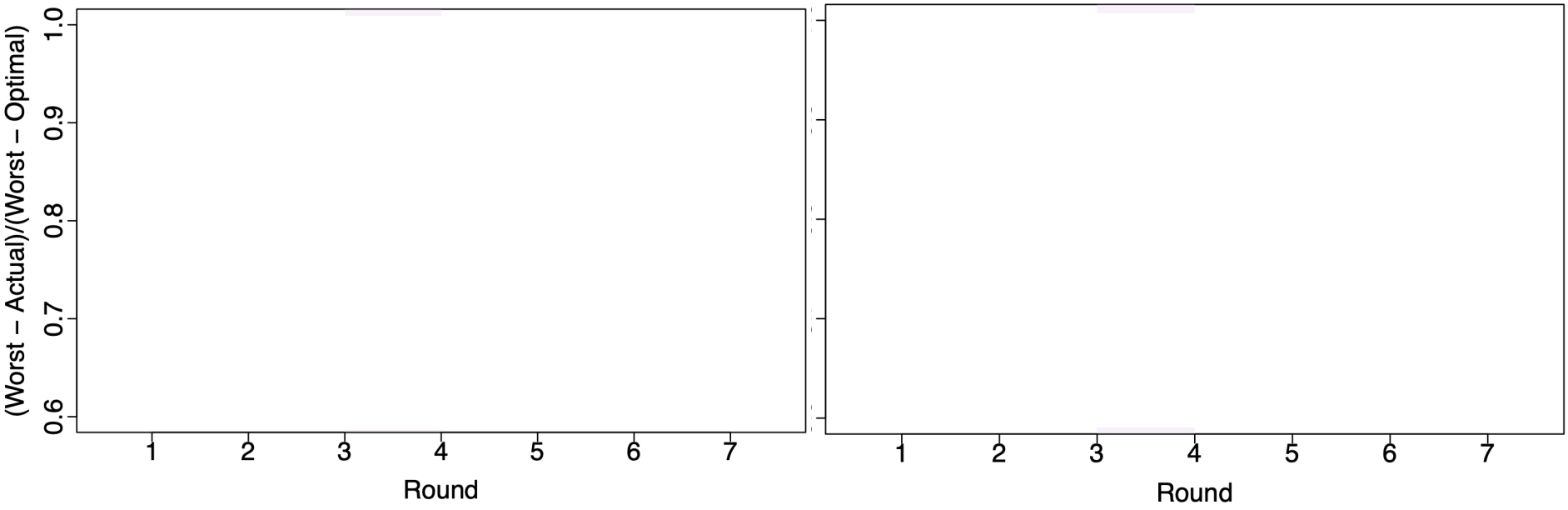
pre-registered at https://aspredicted.org/23C_WBQ

Study 2

Result: Without Rationale

😊 Familiar new map, no rationale

Unfamiliar new map, no rationale 🥺



The learning advantage of Broad is replicated!

...but we identify the limit to transfer learning (Broad does not help when environments are too different)

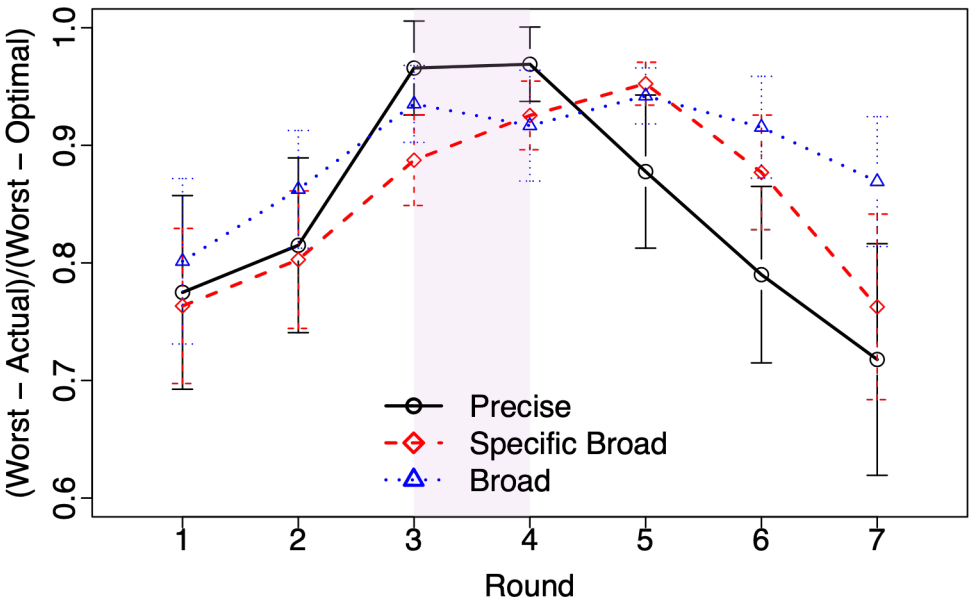
Prolific: N = 400

pre-registered at https://aspredicted.org/23C_WBQ

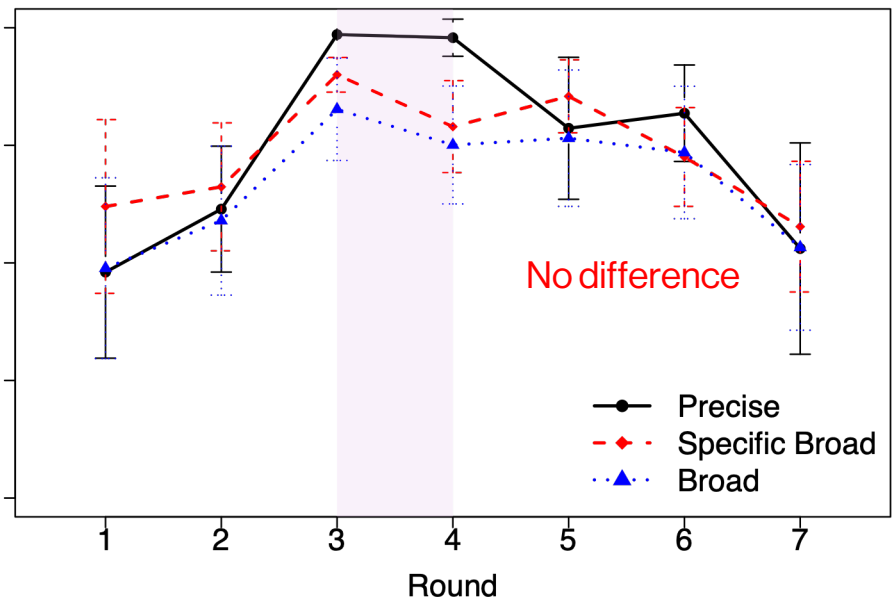
Study 2

Result: Adding Rationale...

😊 Familiar new map, no rationale



Familiar new map, + rationale 😯



Rationale eliminates the post-advice gap: Precise + Explanation ≈ Broad

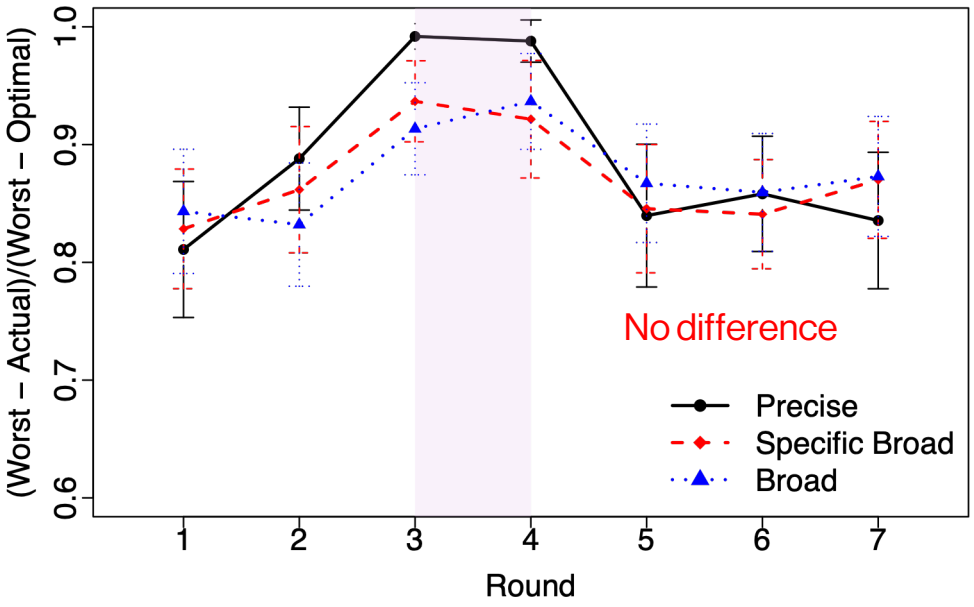
Prolific: N = 400

pre-registered at https://aspredicted.org/23C_WBQ

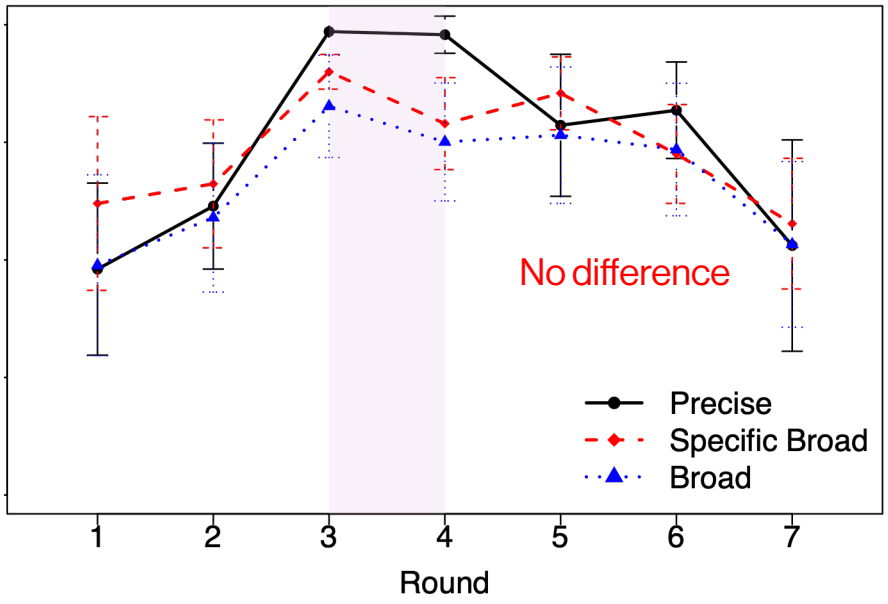
Study 2

Result: Adding Rationale...

😲 Unfamiliar new map, + rationale



Familiar new map, + rationale 😲



Rationale eliminates the post-advice gap: Precise + Explanation ≈ Broad

Why? Broad already communicates the underlying logic.
Adding that logic as an explanation to Precise closes the gap.

Prolific: N = 400

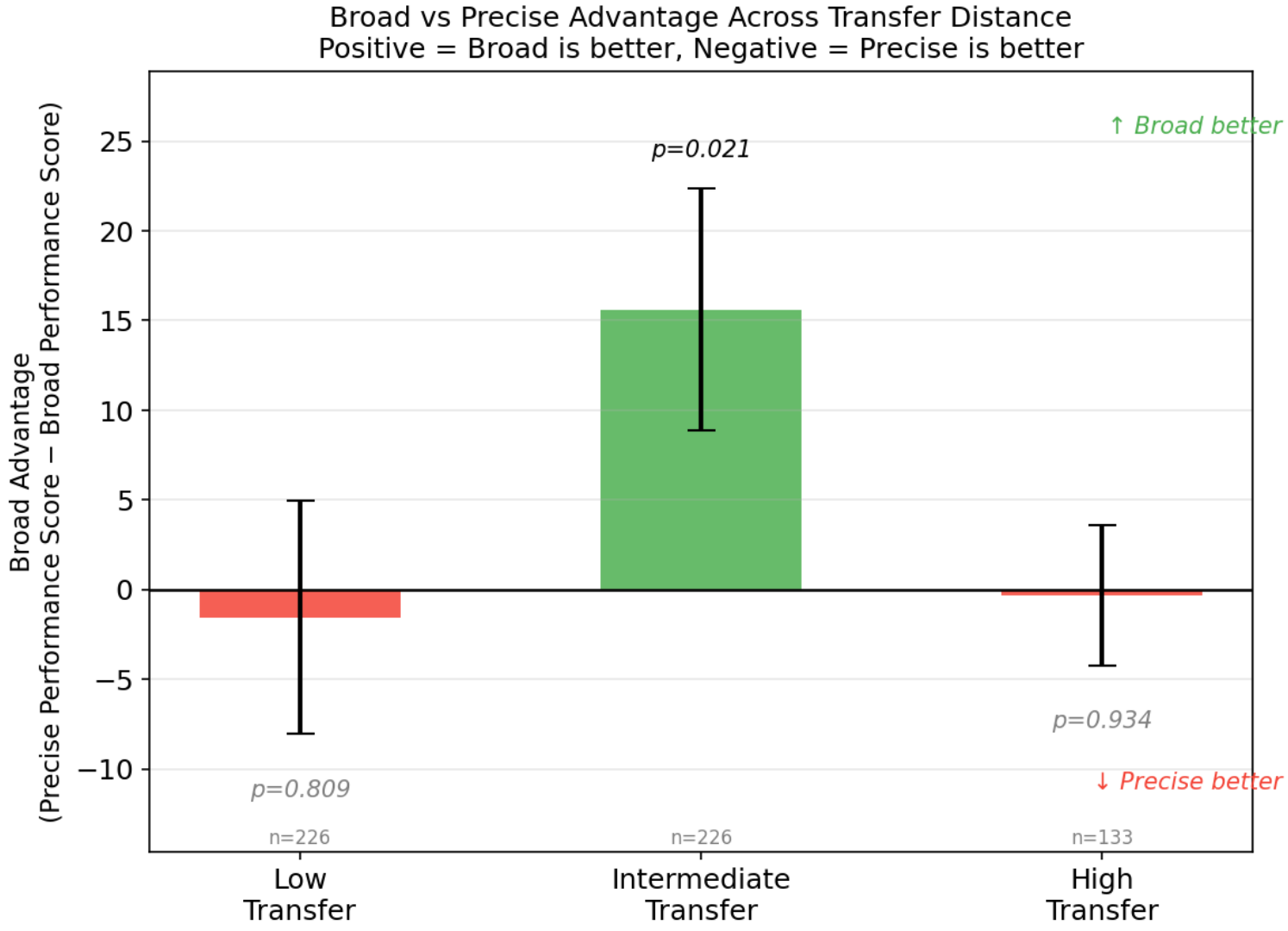
Study 1 Study 2

Varying the Distance

δ low (near transfer)		δ moderate (substantial)		δ high (too different)
Same map	→	Same structure, diff. params	→	Completely new map
Broad ≈ Precise		Broad > Precise		Broad ≈ Precise
Study 1 Round 6 Study 2 Rounds 5-6 Sequence A		Study 1 Round 7 Study 2 Round 7 Sequence A		Study 2 Rounds 5-7 Sequence B

Study 1 Study 2

Varying the Distance



Both Studies 1 + 2 gave us insights on conditions under which Broad vs Precise advice dominates.

But what actually happened?

Next: Quantifying human strategy + recovering their reward function via inverse reinforcement learning

What Are People Actually Optimizing?

Inverse RL infers reward function

- Assume participants optimize a personal reward function: a weighted mix of things they care about $r^h_{s_t}(a_t) = \sum_{j=1}^{k} \theta_j \phi_j(s_t, a_t)$
- We work backwards from their decisions to infer those weights θ.
- **Key challenge:** each participant has their own weights, and we only observe a few rounds / person

→ we use a Bayesian hierarchical model (SVI) that pools info across participants.

ϕ_t Time taken at current stop (including charging & emergency)

"Take the total amount of time needed to get to the next point after including worst traffic scenario..."

ϕ_s Simplicity (prefer 0%, 100%, following tip)

"I first just tried to make sure my charge was 100%. I then tried to make sure I had just enough charge for the next stop."

ϕ_r Risk exposure (keep a buffer)

"Play it safe!"

ϕ_b Batch (charge sufficiently for multiple stops)

"I would probably charge all the way on stop ... so you don't have to charge on the last stop"

scenario-specific shift (e.g., pre/with/post, precise vs broad)

individual's shift

$$\theta_i = \theta_0 + \Delta_s + \Delta_i$$

Advice Precision Changes Objectives

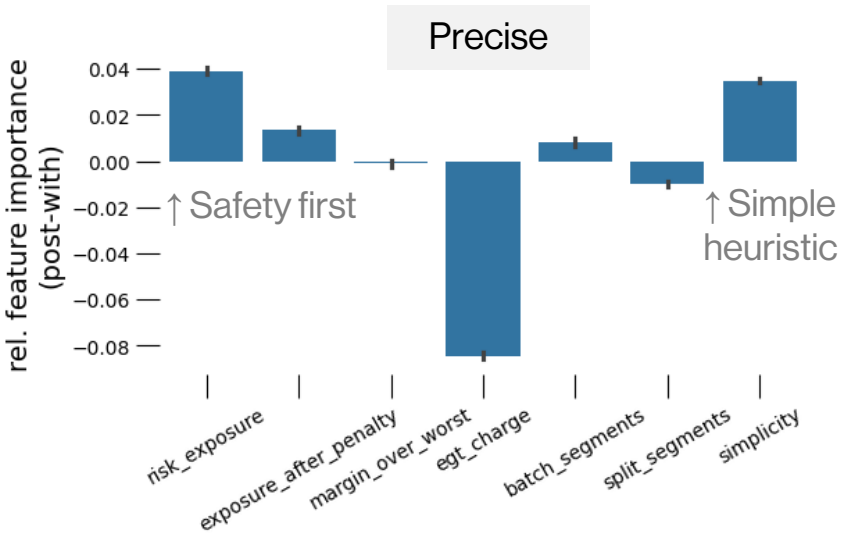
- Y-axis = change in how much participants weight each component, from **with-advice** to **post-advice** phase.

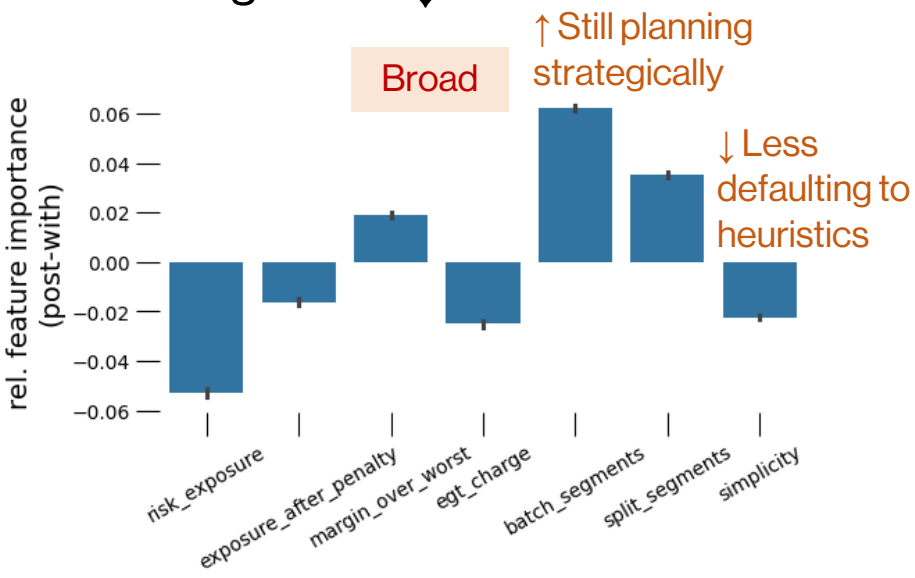
scenario-specific shift
(e.g., pre/with/post, precise vs broad)

individual's shift

$$\theta_i = \theta_0 + \Delta_s + \Delta_i$$

- Positive = weight ↑ after advice removed. Negative = ↓.

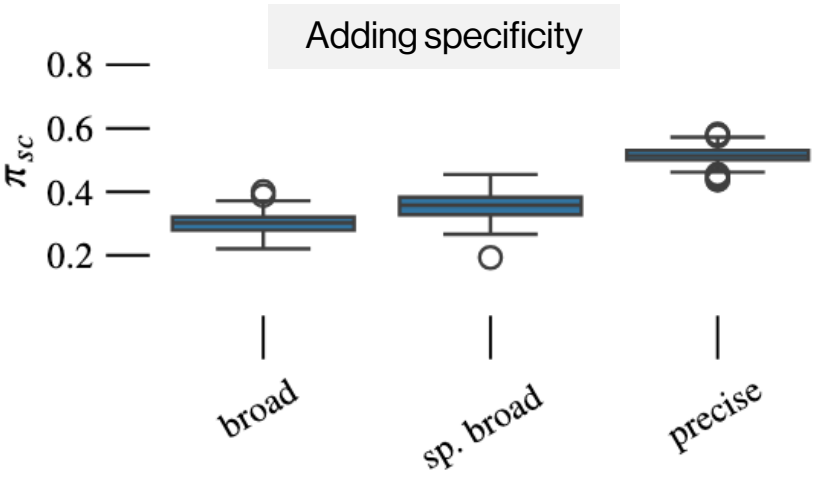




Precise: compliance was behavioral. When the tip left, so did the strategy.
Broad : compliance was cognitive. The strategic thinking stayed.

Compliance Confirms the Mechanism

We consider the average compliance probability for a given tip type, π_{sc}:



Adding specificity barely moves compliance. (It doesn't reduce the translation work.)

Adding an explanation brings broad up to precise levels, and precise up further. Same info, same compliance. Same info, same learning.

Takeaways/Implications


Optimizing for in-system performance can actively degrade long-run capability. The right question isn't "how well do workers perform with the advice?" It's "how well do they perform without it?"

Advice precision shapes what humans learn to optimize, not just what they do. Design question: do you want workers who can execute the algorithm's output, or workers who can reason like the AI?

It's the cognitive work of translation that produces learning, not the strategic information content of the advice. Specific broad gives more info, narrower action, but broad is still better. Learning takes doing the reasoning, not receiving it.

We've shown that (different types of) **real-time** AI advice impact willingness to accept **(compliance)** + behavior once removed **(learning).**


Humans resist counterintuitive AI advice. **A single tip** achieved better performance than no tips, **despite low compliance.**



performance
optimizing the
advice given?

with E
(Wharto

informs
https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–23
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Improving Human Sequential Decision Making with Reinforcement Learning

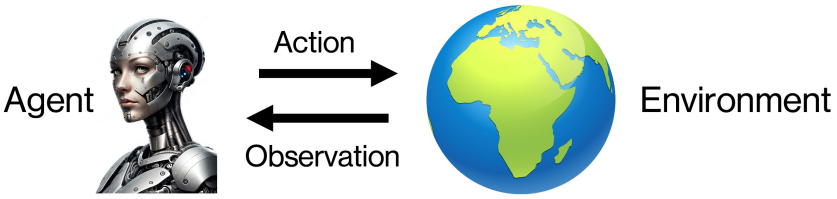
Hamsa Bastani,[a] Osbert Bastani,[b] Wichinpong Park Sinchaisri[c,*]

[a] Operations, Information and Decisions, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104; [b] Computer and Information Science, University of Pennsylvania, Philadelphia, Pennsylvania 19104; [c] Haas School of Business, University of California, Berkeley, Berkeley, California 94720

*Corresponding author

Contact: hamsab@wharton.upenn.edu, https://orcid.org/0000-0002-8793-4732 (HB); obastani@seas.upenn.edu, https://orcid.org/0000-0001-9990-7566 (OB); parksinchaisri@berkeley.edu, https://orcid.org/0000-0001-9351-0541 (WPS)

RL Reinforcement Learning

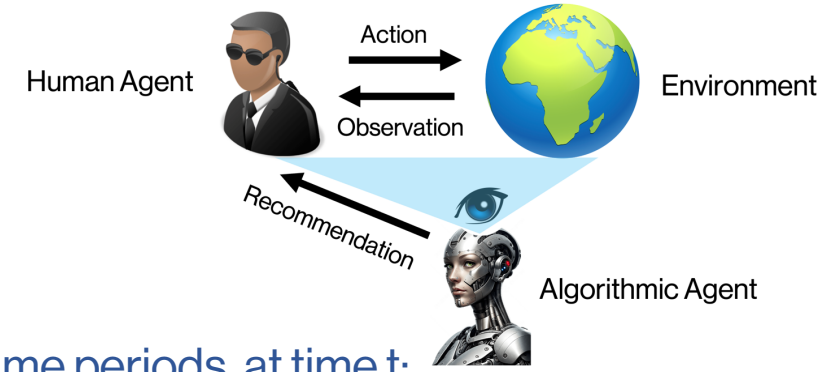


T time periods, at time t:

1. Agent chooses action A_t
2. Environment returns observation O_t
3. Learn from resulting reward

Action → Outcome

Compliance-Aware RL CA-RL



T time periods, at time t:

1. Algorithmic agent **might** recommend action $A_t^\dagger$
2. Human agent chooses action A_t
3. Environment returns observation O_t

Baseline policy: What the human would do on their own.

Compliance ψ: Probability $A_t^\dagger$ accepted = How much the recommendation shifts behavior.

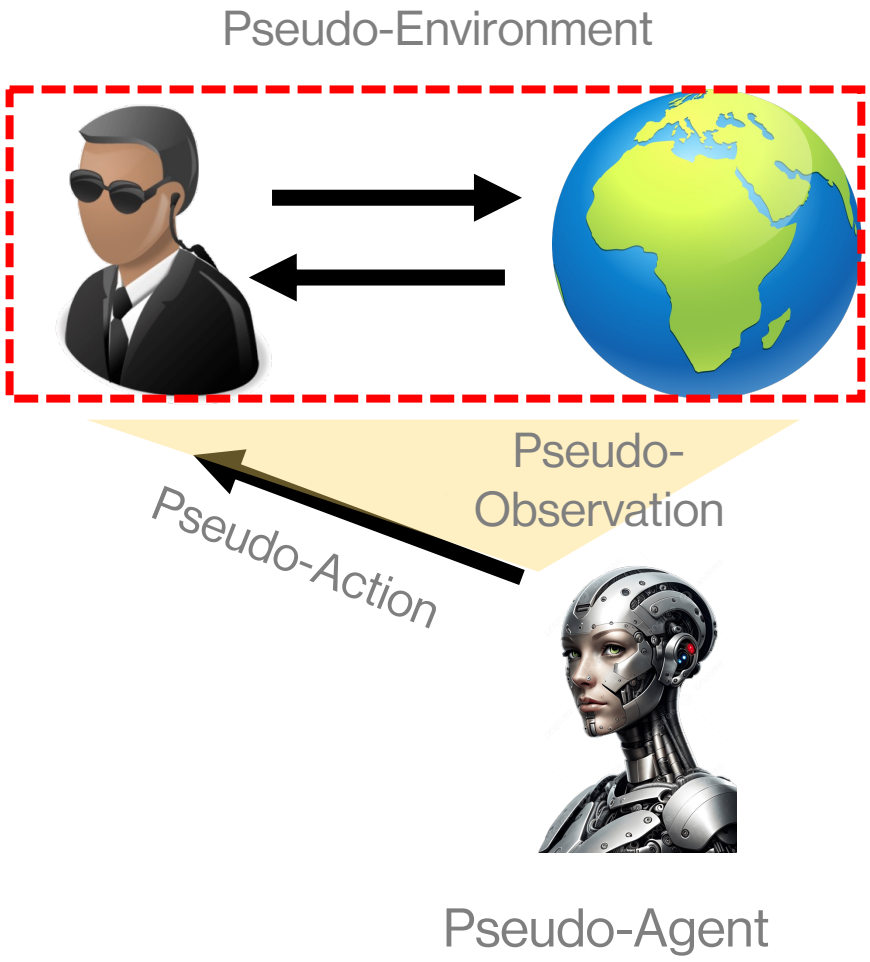
Recommendation ≠ action

CA-RL = RL w/ Useful Structure

CA-RL structures RL that considers the human as part of the environment

Exclusion: Recommendations affect outcomes only through the human action

Monotonicity: Recommending action A makes A more likely → Shink implementable policy space → Speed up learning



Always Recommending Optimal Action May Be Suboptimal

If the algorithm had direct control, it would choose the optimal action. But with humans in the loop, the best action need not be the best recommendation.

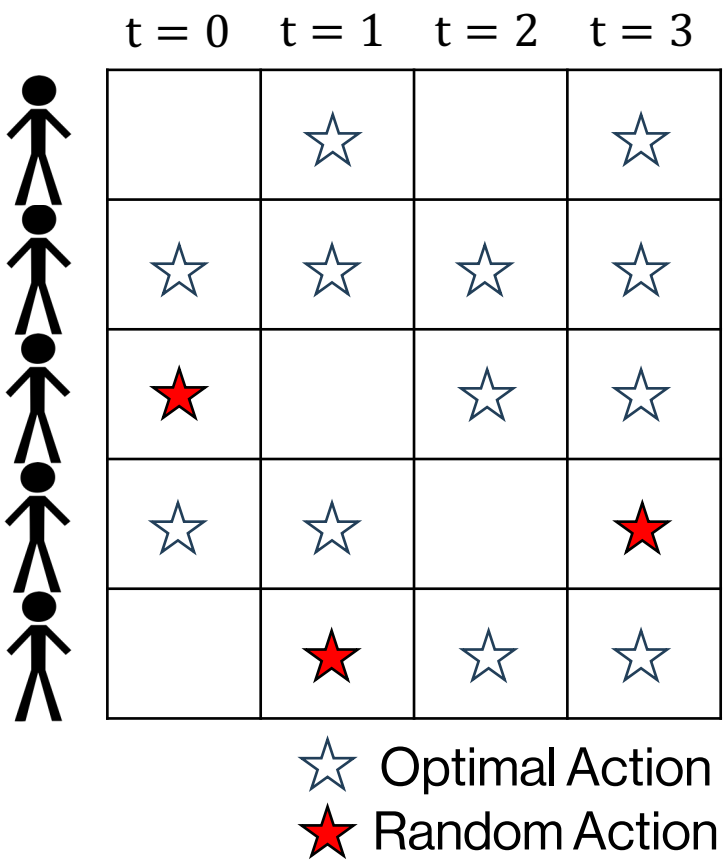
Too much advice can reduce compliance: advice fatigue, weak or noisy signal, misaligned with human intuition

Too much advice can reduce learning: weaker baseline policy later, less capability when advice is absent

Exploring Human Behavior with MRTs (Microrandomized Trials)

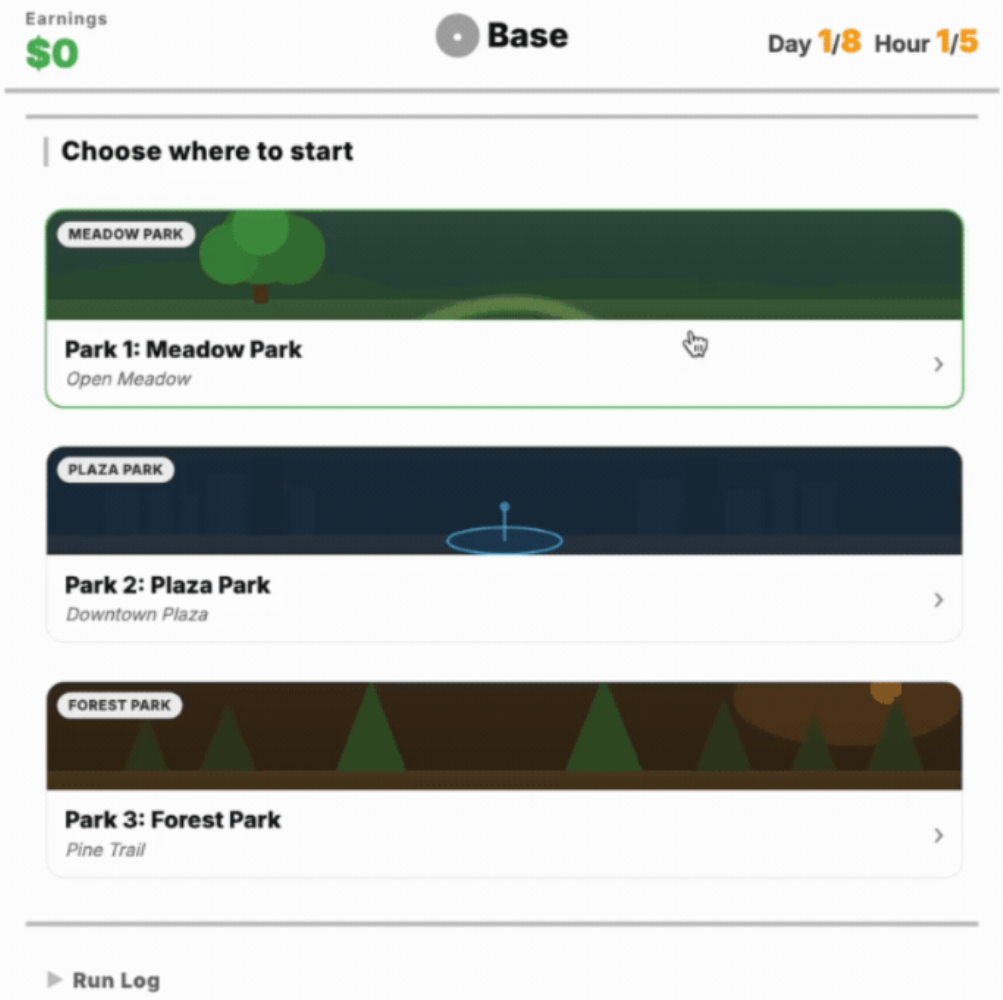
MRT

MRTs randomize at unit-time level



Food Truck Game

Design: Task + Interface



Learn which park to park your food truck to maximize earnings (learning customer demand + competition)

Advice conditions:

Low – 10%

Medium – 50%

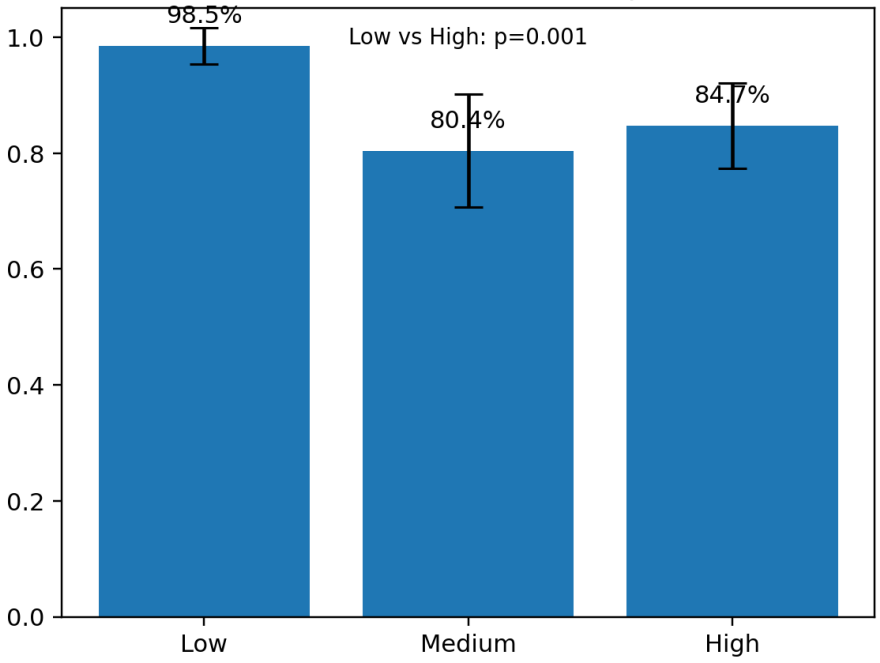
High – 90%

Undergrads: N = 60

Food Truck Game

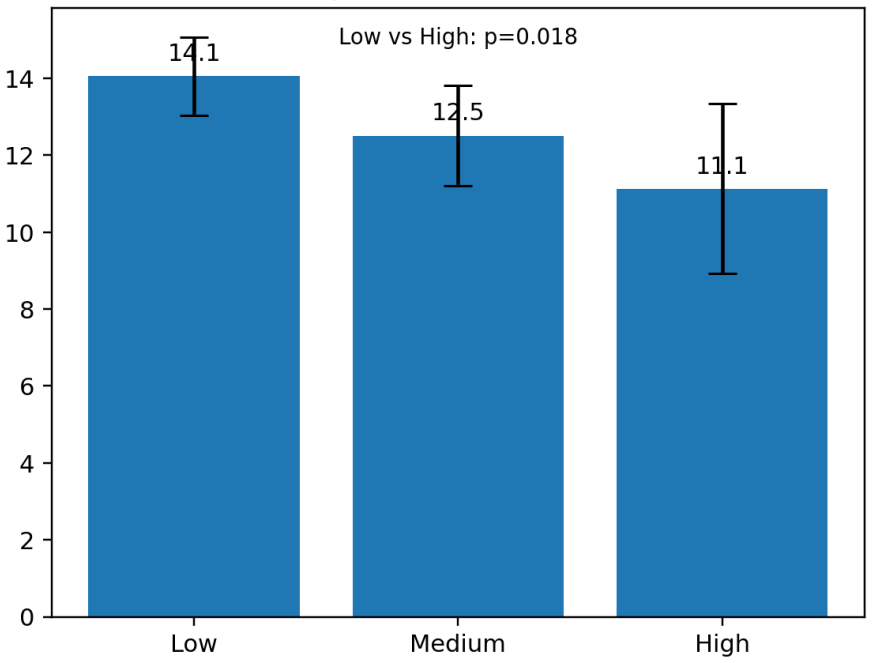
Pilot Results

Compliance when advice is available



Low advice has much higher immediate visible compliance than High: 98.5% vs 84.7%; p=0.001.

Post-advice performance



Low advice also has better post-treatment hidden-round reward than High: 14.1 vs 11.1; p=0.018.

The best immediate recommendation policy may not be the best long-run recommendation policy.

Summary + Thank You!

Implications: The most effective AI advice design depends on context: volatility, familiarity, and users' capacity to generalize. Shift reward function not just behavior!

Goal: How to Design/Deliver AI Tips to Support Long-Term Learning?

AI Tips

Broad tips promote strategic exploration and long-term learning, but only when users can infer the rationale themselves

"Driving Game"

Precise tips improve short-run efficiency, but without explanation, they can limit learning and adaptability.

New environment (no tips)

How Humans Make Sequential Decisions?

How Do Tips Shape Learning & Behavior Over Time?

Can Humans Adapt When Environment Changes?

Feedback (+ tips) very welcome!

Philippe Blaettchen
Singapore Management University

Park Sinchaisri
Berkeley Haas
parksinchaisri@berkeley.edu