UNSW AUSTRALIA

Never Stand Still

Learning the structure of an explore-exploit dilemma

Dan Navarro

A puzzle

Human RL needs to infer *which model* to apply in which context, solve problems with *large state spaces*, using *limited computational resources* and with *minimal training* data in any one context. How?

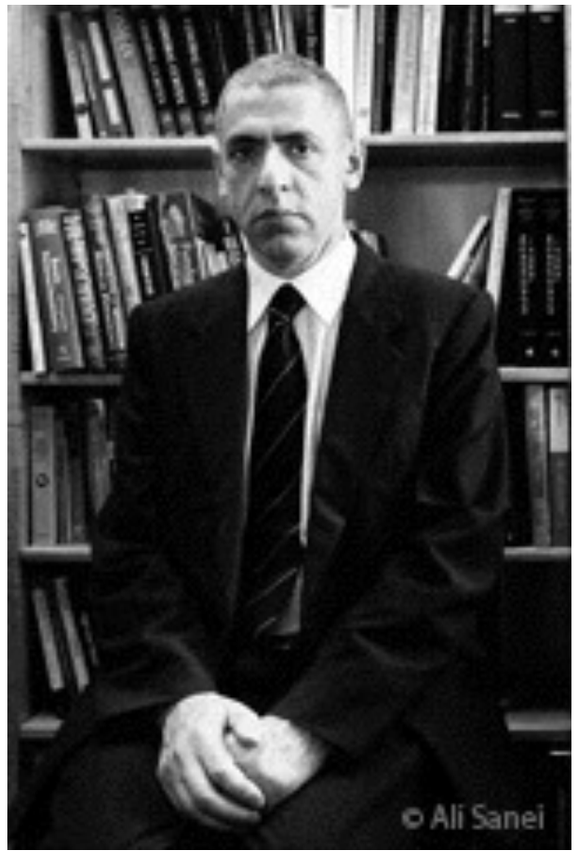
My decision making task this morning

Go to a kids party?

My decision making task this morning

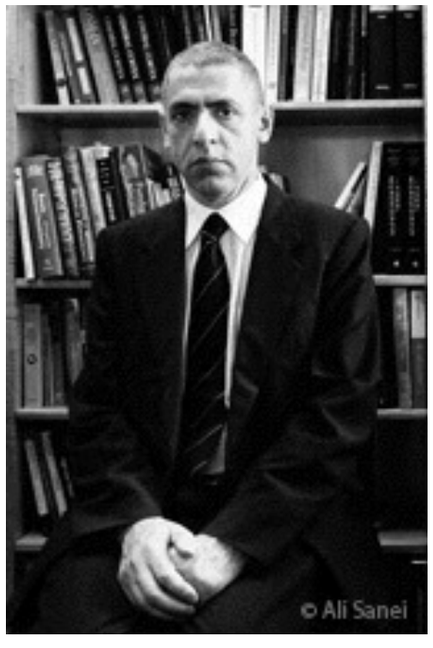
Go to a kids party?

Attend a decision neuroscience talk?



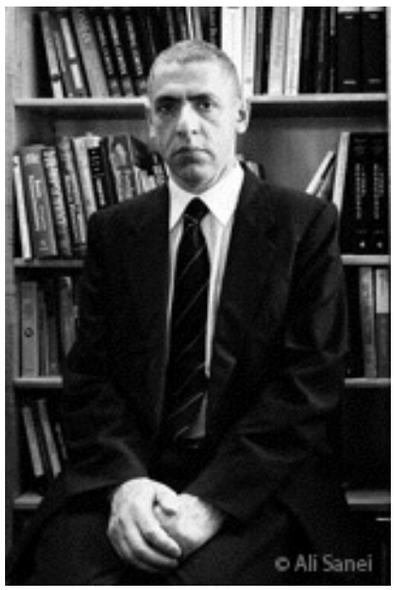
Choices vary in many respects

Immediately rewarding, not intellectually taxing, emotional competence required…



Probably long term rewarding, high cognitive load, not emotionally difficult…

My direct experience is non-existent



This choice problem doesn't actually arise very often

So I'm necessarily constructing a model on the fly of what might happen based on partially-relevant data

(… decision making requires inductive generalisation)

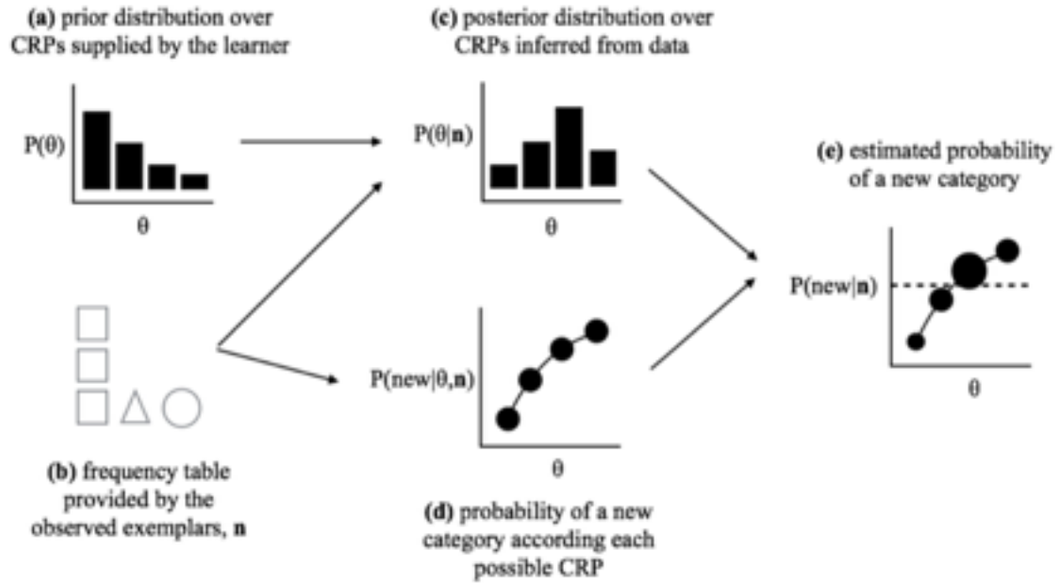
How do people acquire new knowledge?

(categorisation & reasoning)

How do people acquire new knowledge?

(categorisation & reasoning)

What kind of prior biases shape the acquisition of new knowledge?



Pas Foo ???

Which category does this belong to?

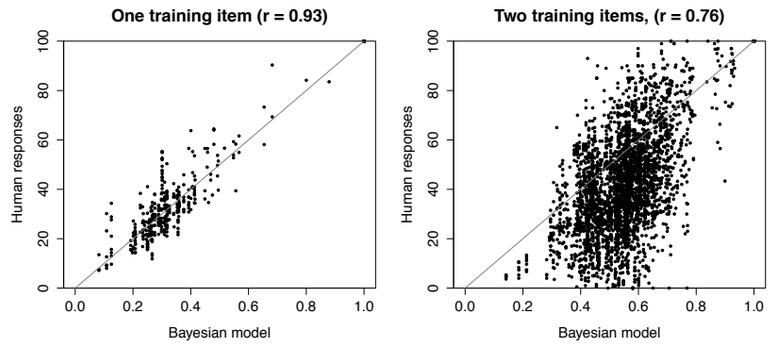
Pas Foo New

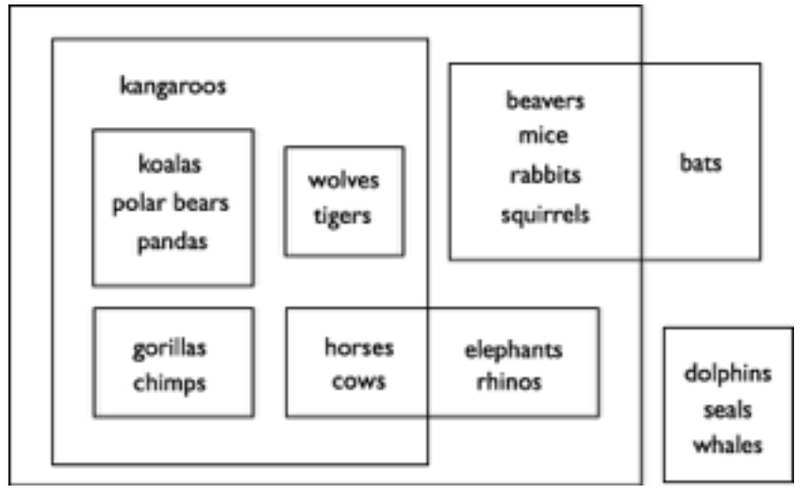
Navarro & Kemp (under revision). *Psych. Review*

How do people acquire new knowledge?

(categorisation & reasoning)

What old knowledge do people use to guide inferences?



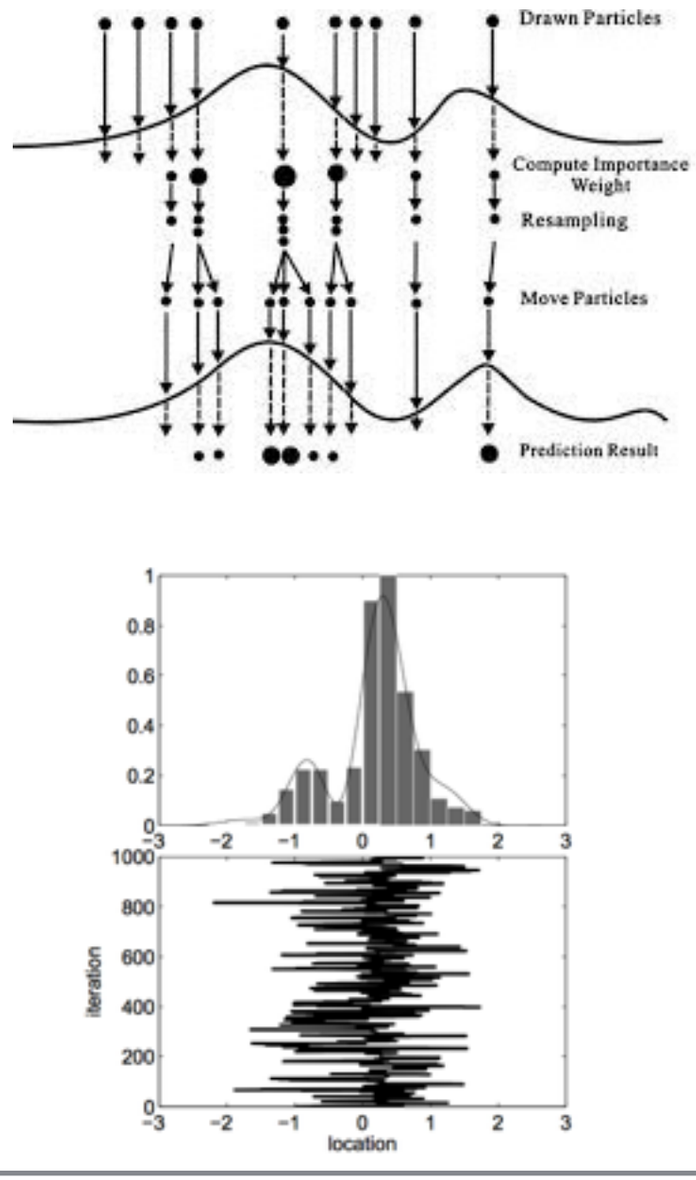


Tauber, Navarro, Perfors & Steyvers (in press). *Psych. Review*

How do people acquire new knowledge?

(categorisation & reasoning)

What computational strategies do people use to simplify complex problems?



Sanborn, Griffiths & Navarro (2010). *Psych. Review*

How do people acquire new knowledge?

(categorisation & reasoning)

How do we make choices in an uncertain world?

(judgment & decision making)

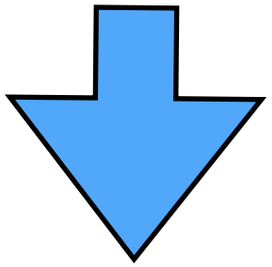
Sequential decision tasks under uncertainty

How do people acquire new knowledge?

(categorisation & reasoning)

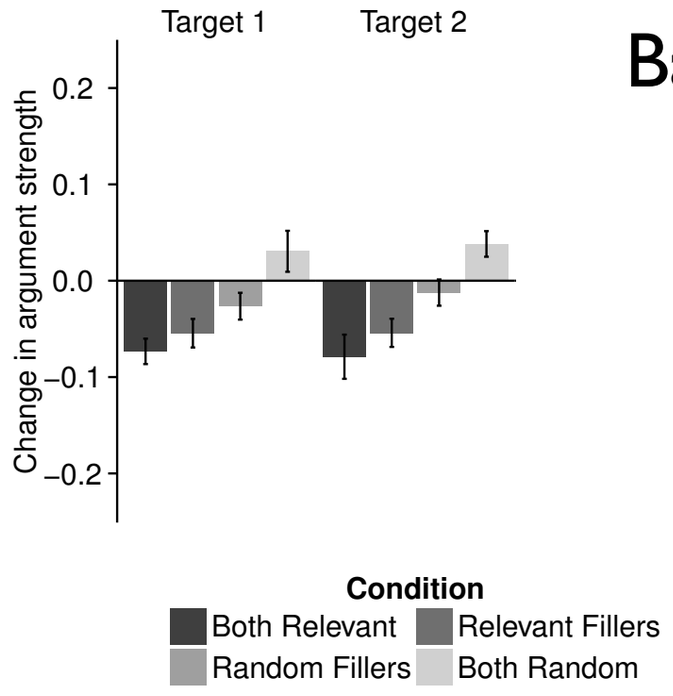
How do we make choices in an uncertain world?

(judgment & decision making)

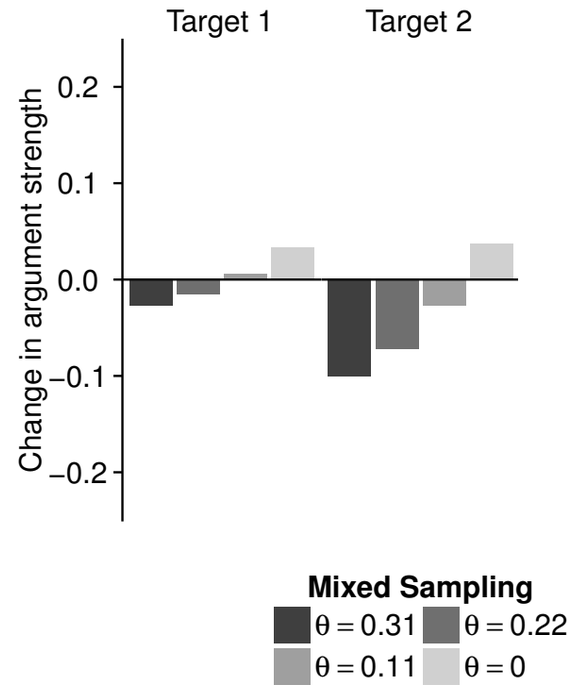


Learner's theory of the data generating mechanism induces qualitative shifts in reasoning

Human:



Bayes:

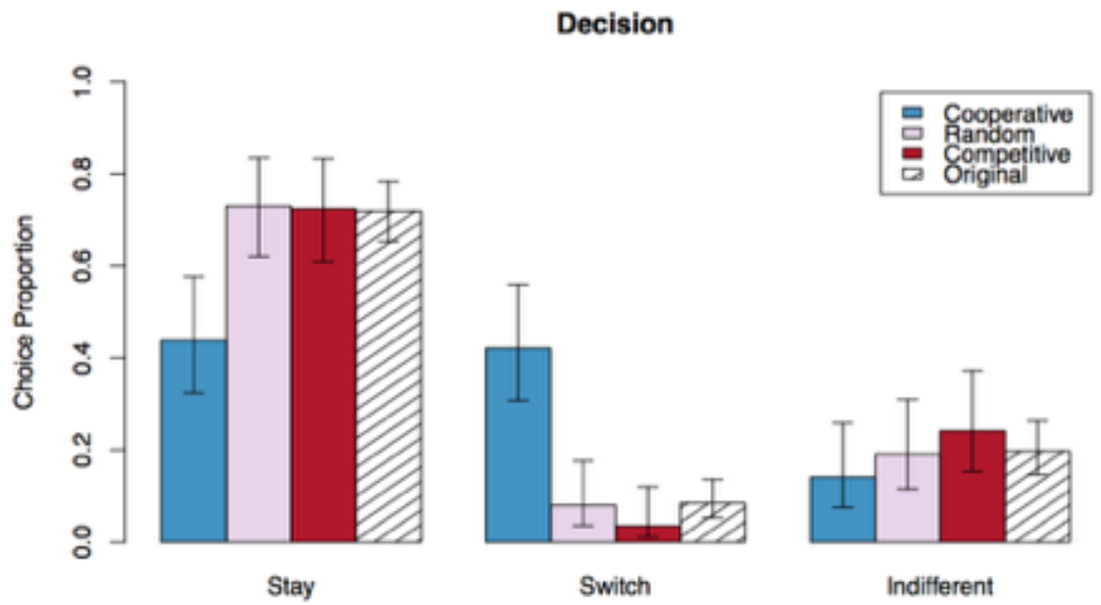


The evidentiary value of the same new fact points in opposite directions depending on <u>how it was selected</u>

<u>*Learner's theory*</u> of the data generating mechanism induces qualitative shifts in reasoning

Ransom, Perfors & Navarro (2016). *Cognitive Science*

Voorspoels, Navarro, Perfors, Ransom & Storms (2015). *Cognitive Psychology*



Human behaviour in an (extended) Monty Hall problem depends on <u>social intent of the host</u>

<u>*Learner's theory*</u> of the data generating mechanism induces qualitative shifts in reasoning

Perfors, Navarro, Donkin & Benders (under review). *Cognition*

Puzzle, reframed:

Where does the theory of (model for) the decision problem come from?

2 Chapter 1. Probability Models

servations that are mutually independent and identically distributed (IID), or X might be some general quantity. The set of possible values for X is the *sample space* and is often denoted as $\mathcal{X}$. The members P_θ of the parametric family will be distributions over this space $\mathcal{X}$. If X is continuous or discrete, then densities or probability mass functions[1] exist. We will denote the density or mass function for P_θ by $f_{X|\Theta}(\cdot|\theta)$. For example, if X is a single random variable with continuous distribution, then

$$P_\theta(a < X \leq b) = \int_a^b f_{X|\Theta}(x|\theta)dx.$$

If $X = (X_1, \ldots, X_n)$, where the X_i are IID each with density (or mass function) $f_{X_i|\Theta}(\cdot|\theta)$ when $\Theta = \theta$, then

$$f_{X|\Theta}(x|\theta) = \prod_{i=1}^{n} f_{X_i|\Theta}(x_i|\theta), \tag{1.1}$$

where $x = (x_1, \ldots, x_n)$. After observing the data $X_1 = x_1, \ldots, X_n = x_n$, the function in (1.1), as a function of θ for fixed x, is called the *likelihood function*, denoted by $L(\theta)$. Section 1.3 is devoted to a motivation of the above structure based on the concept of *exchangeability* and DeFinetti's representation theorem 1.49. Exchangeability is discussed in detail in Section 1.2, and DeFinetti's theorem is the subject of Section 1.4.

1.1.2 Classical Statistics

Classical inferential techniques include tests of hypotheses, unbiased estimates, maximum likelihood estimates, confidence intervals and many other things. These will be covered in great detail in the text, but we remind the reader of a few of them here. Suppose that we are interested in whether or not the parameter lies in one portion Ω_H of the parameter space. We could then set up a *hypothesis* $H : \Theta \in \Omega_H$ with the corresponding *alternative* $A : \Theta \notin \Omega_H$. The simplest sort of *test* of this hypothesis would be to choose a subset $R \subseteq \mathcal{X}$, and then reject H if $x \in R$ is observed. The set R would be called the *rejection region* for the test. If $x \notin R$, we would say that we do not reject H. Tests are compared based on their power functions. The *power function* of a test with rejection region R is $\beta(\theta) = P_\theta(X \in R)$. The *size* of a test is $\sup_{\theta \in \Omega_H} \beta(\theta)$. Chapter 4 covers hypothesis testing in depth.

Example 1.2. Suppose that $X = (X_1, \ldots, X_n)$ and the X_i are IID with $N(\theta, 1)$ distribution under P_θ. The usual size α test of $H : \Theta = \theta_0$ versus $A : \Theta \neq \theta_0$ is

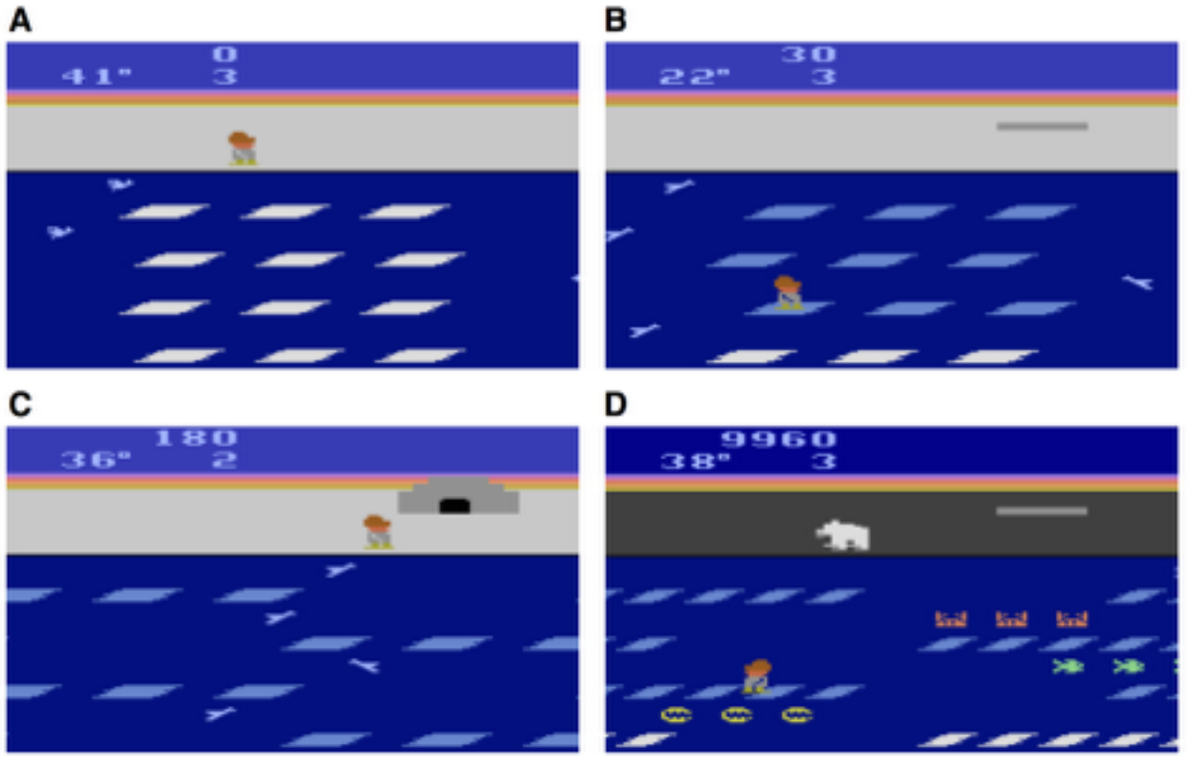
[1] Using the theory of measures (see Appendix A) we will be able to dispense with the distinction between densities and probability mass functions. They will both be special cases of a more general type of "density."

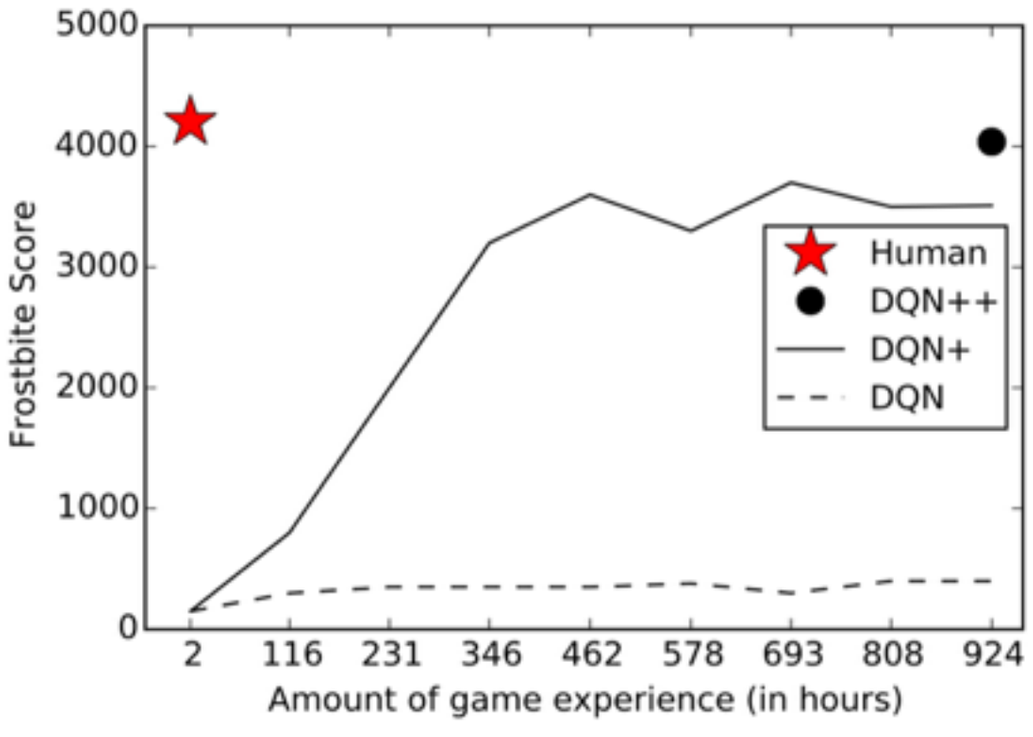
Building Machines That Learn and Think Like People

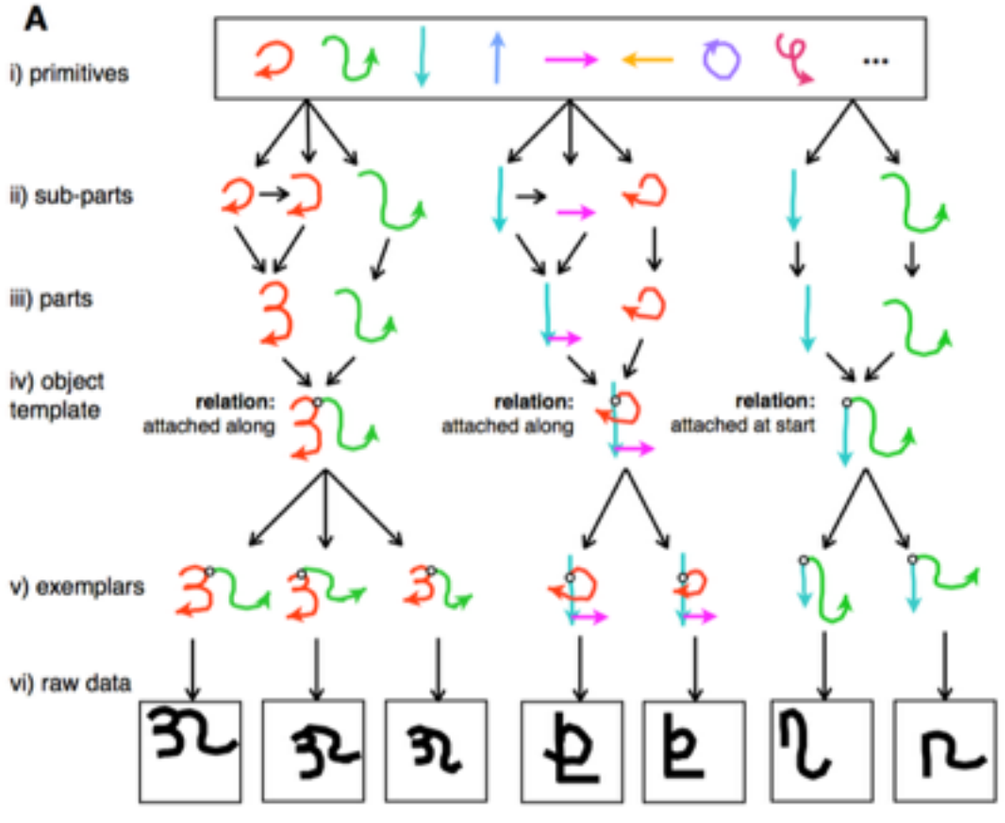
Brenden M. Lake,[1] Tomer D. Ullman,[2,4] Joshua B. Tenenbaum,[2,4] and Samuel J. Gershman[3,4]
[1]Center for Data Science, New York University
[2]Department of Brain and Cognitive Sciences, MIT
[3]Department of Psychology and Center for Brain Science, Harvard University
[4]Center for Brains Minds and Machines

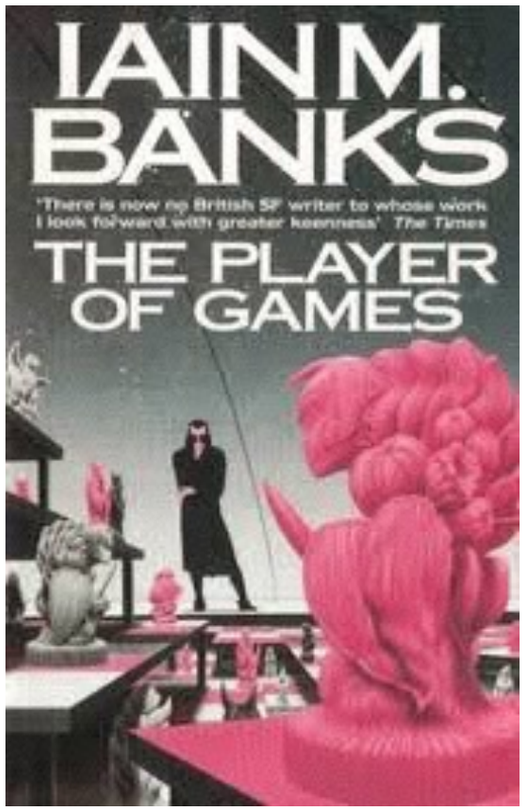
Abstract

Recent progress in artificial intelligence (AI) has renewed interest in building systems that learn and think like people. Many advances have come from using deep neural networks trained end-to-end in tasks such as object recognition, video games, and board games, achieving performance that equals or even beats humans in some respects. Despite their biological inspiration and performance achievements, these systems differ from human intelligence in crucial ways. We review progress in cognitive science suggesting that truly human-like learning and thinking machines will have to reach beyond current engineering trends in both what they learn, and how they learn it. Specifically, we argue that these machines should (a) build causal models of the world that support explanation and understanding, rather than merely solving pattern recognition problems; (b) ground learning in intuitive theories of physics and psychology, to support and enrich the knowledge that is learned; and (c) harness compositionality and learning-to-learn to rapidly acquire and generalize knowledge to new tasks and situations. We suggest concrete challenges and promising routes towards these goals that can combine the strengths of recent neural network advances with more structured cognitive models.



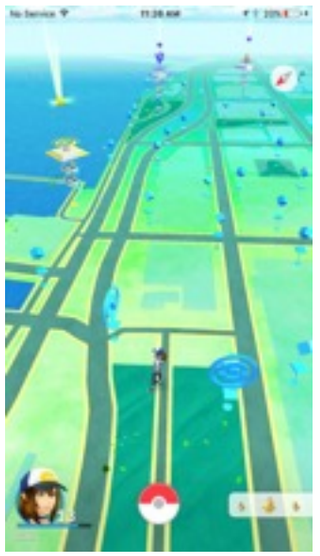


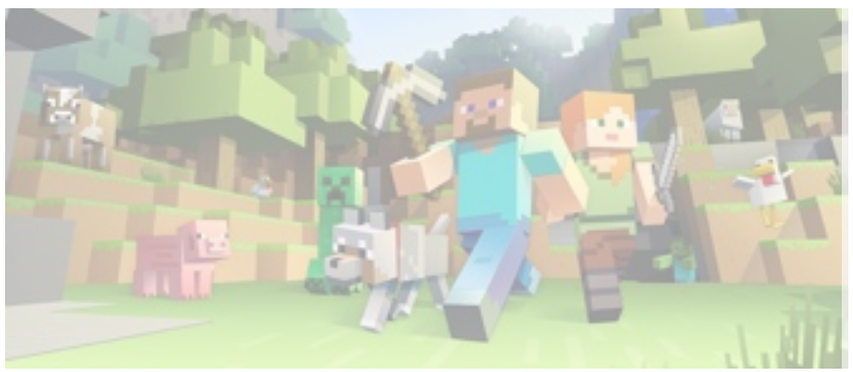


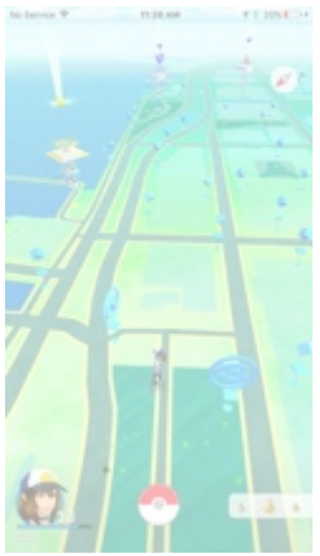


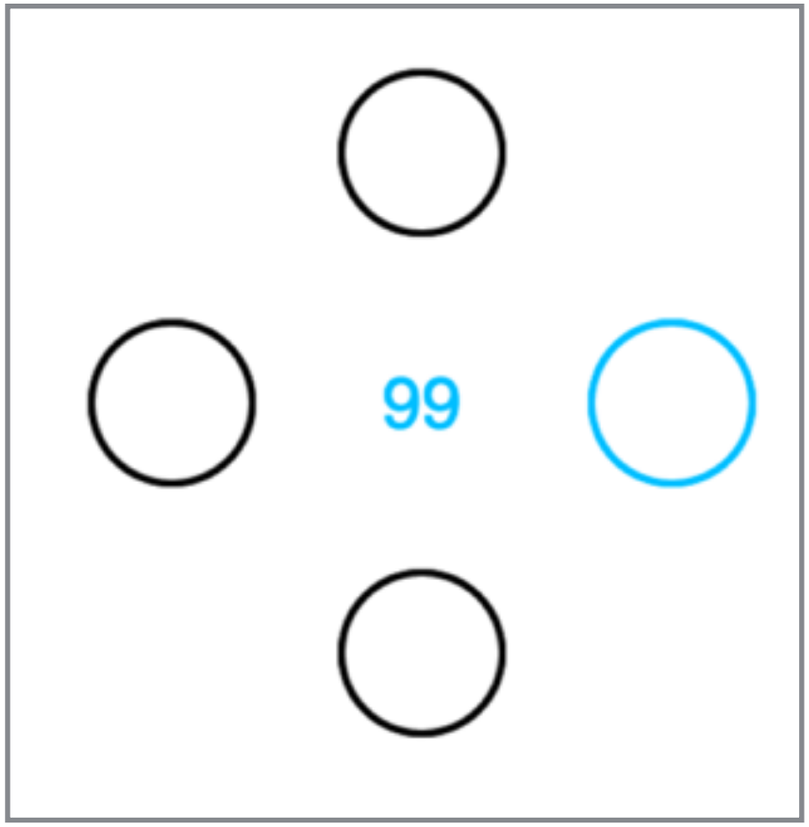
"You've spent all your life learning games; there can't be a rule, move, concept or idea in [*super complicated game*] you haven't encountered ten times before in other games; it just brought them all together."

The players of games





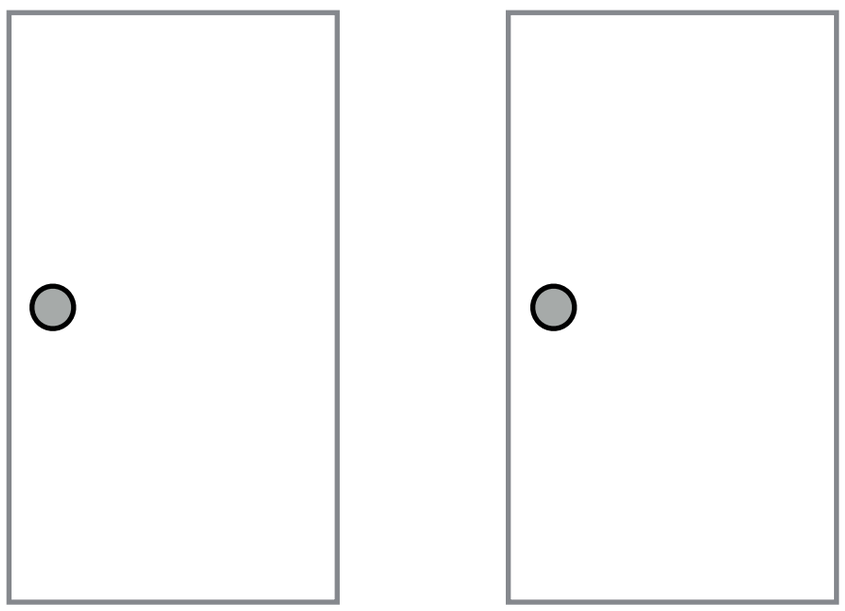



99

Trading-off information and reward

The tiger problem

(as per every text on POMDPs)



Actions?

Make an observation: Listen at the door for the sounds of a tiger

Take a bet: Open a door and see what's behind it

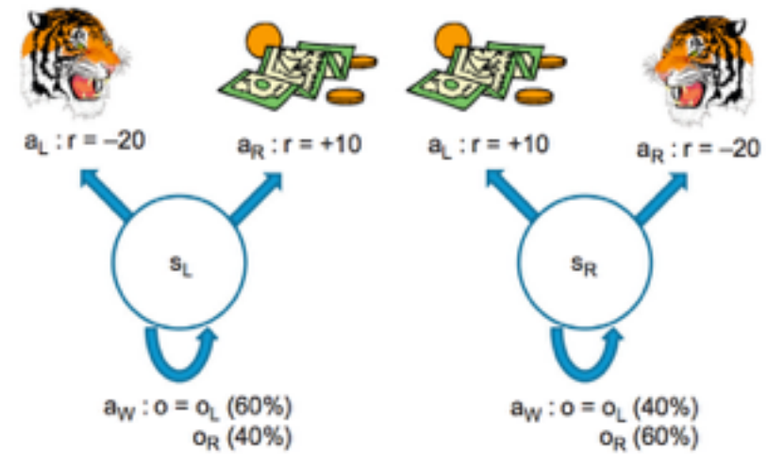


FIGURE 16.3 The tiger POMDP. The subject does not know if she is in state s_L, where the left door a_L is dangerous, or state s_R, where the right door a_R is dangerous. Only by waiting (a_W) and accumulating evidence about which state obtains is it safe to choose a door.

Information versus reward problems for online workers…

Do some work: Tag some images, and eventually get a reward when the requester pays. *If* they pay. No immediate knowledge… but possibly a reward

Do your research: Check out Turkopticon (etc.), read the reviews for the requester. Maybe check out Turk and see if there are any better jobs on offer? No immediate reward… only information

Information versus reward problems for resource companies…

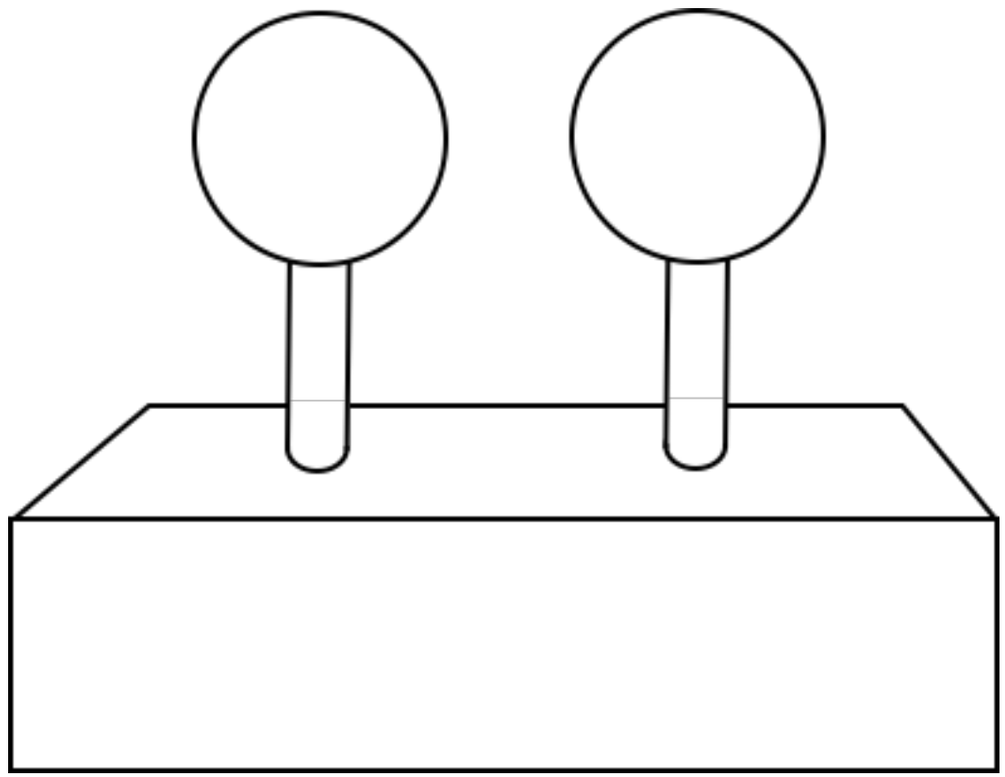
Invest in exploitation: Dig some mines, sink some wells, build some factories. Doesn't teach us much about the world (initially), but it's how the company makes money

Invest in exploration: Send out geologists, hire JDM researchers to teach geologists how to do statistics, etc. Doesn't sell any barrels of oil, but identifies potentially profitable actions

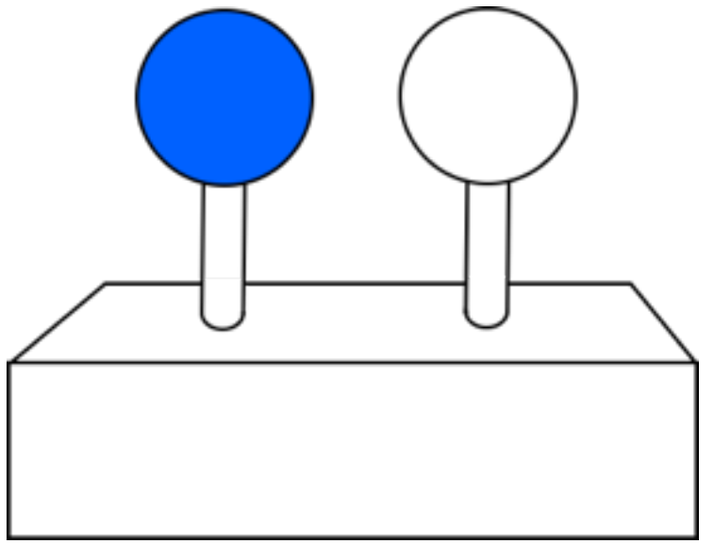
A simple experimental task
(adapted from Tversky & Edwards 1966)

Navarro, Newell & Schulze (2016). *Cognitive Psychology*

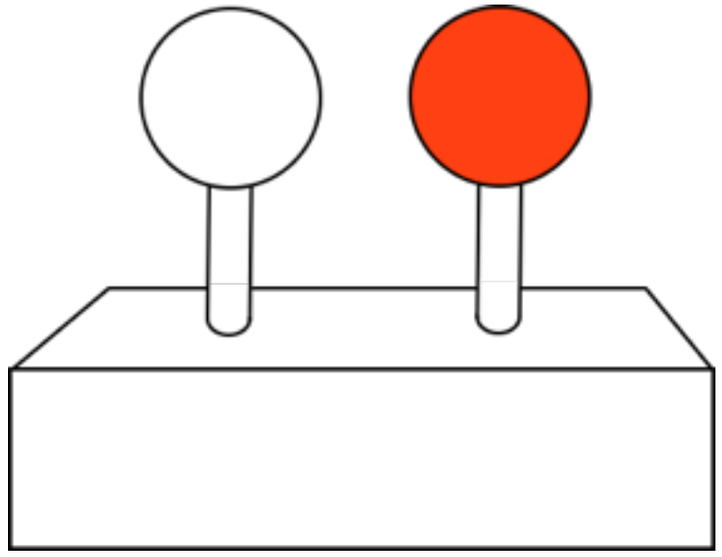
The observe or bet task



This is a “blox” machine

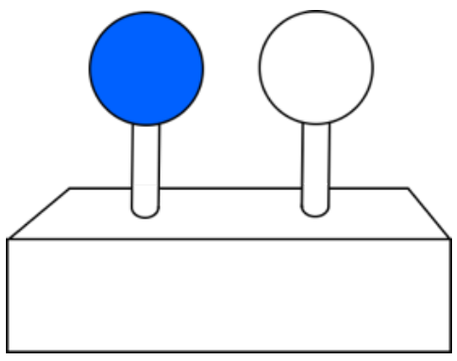
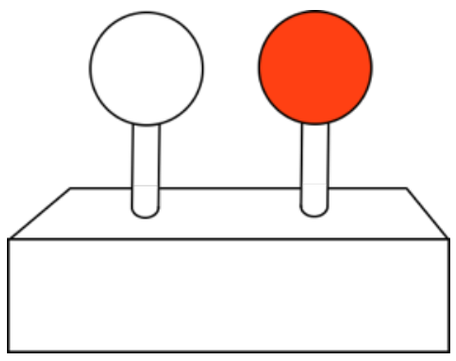
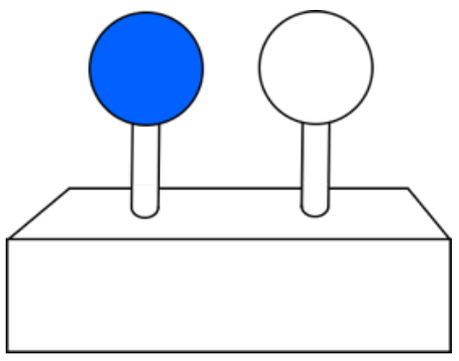
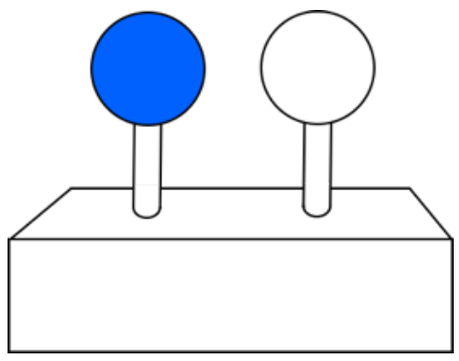
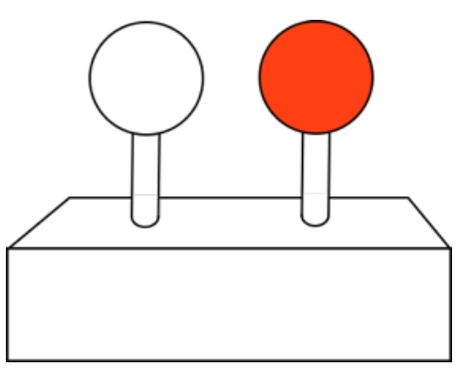
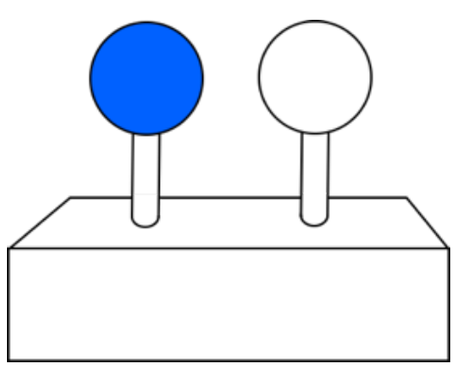
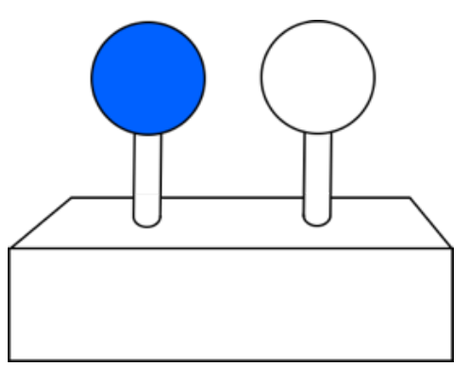
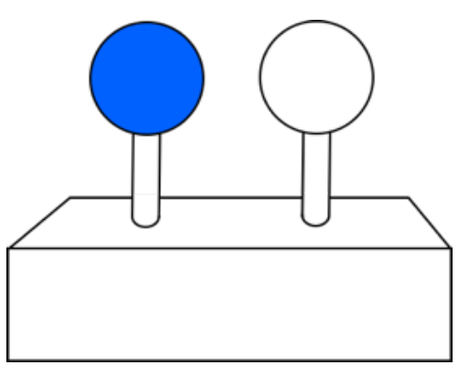


It has a blue light

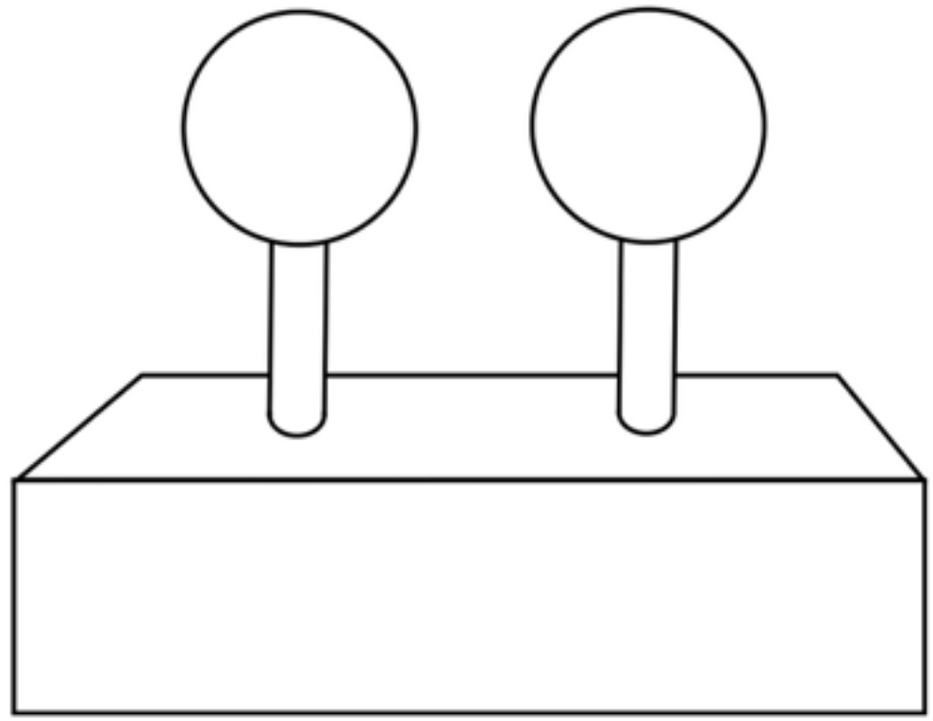


and a red light

These lights flash
intermittently.

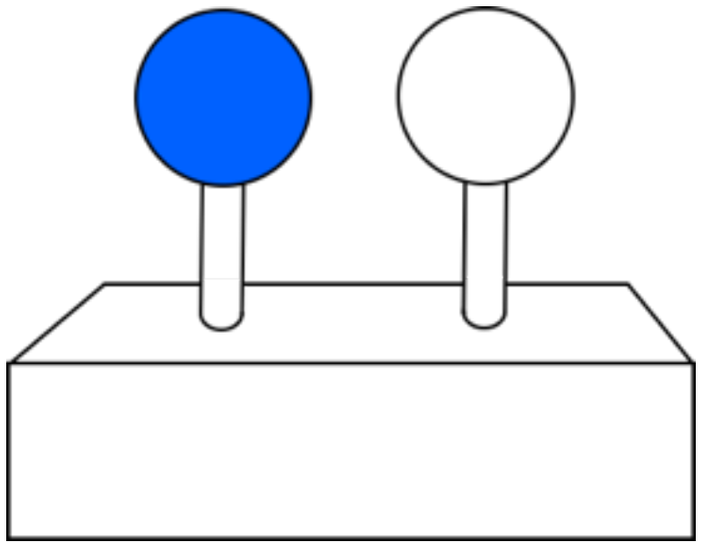
One light tends to come on
more often than the other.
You don't know which



Observe Guess Blue Guess Red

At every point in time, you can make an observation or bet on which outcome will occur…

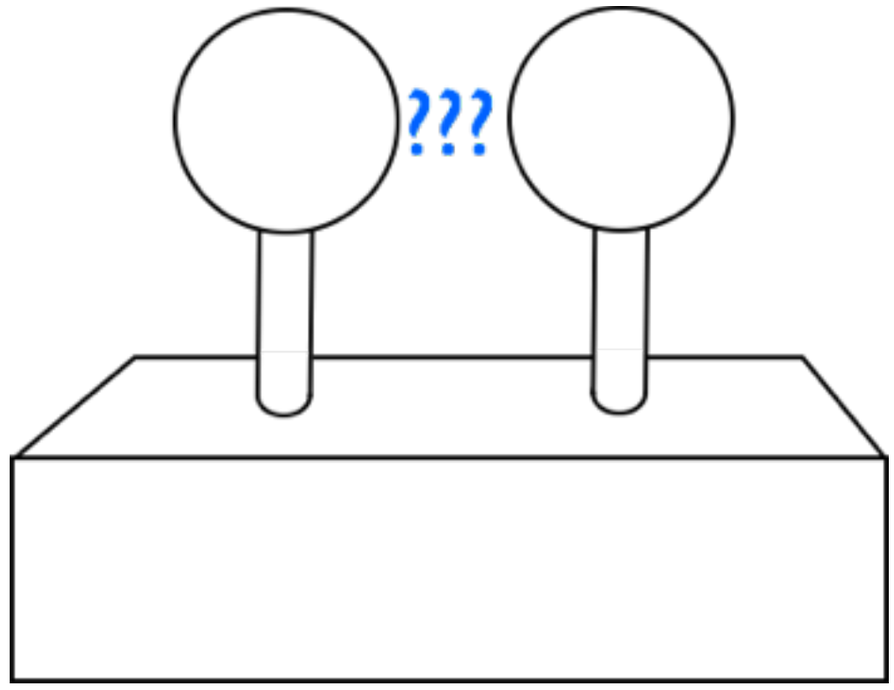
If you OBSERVE, you get to see
which light turns on



… but you receive no reward

(information only)

If you **BET** (on blue) you receive a point (+1) if you're correct, and lose if you're wrong (-1).



But the outcome is hidden from you until the end to the task, so you can't learn from this trial

(delayed reward only)

The task

- Win as many points in a 50 trial "game"
- Play a series of 5 games

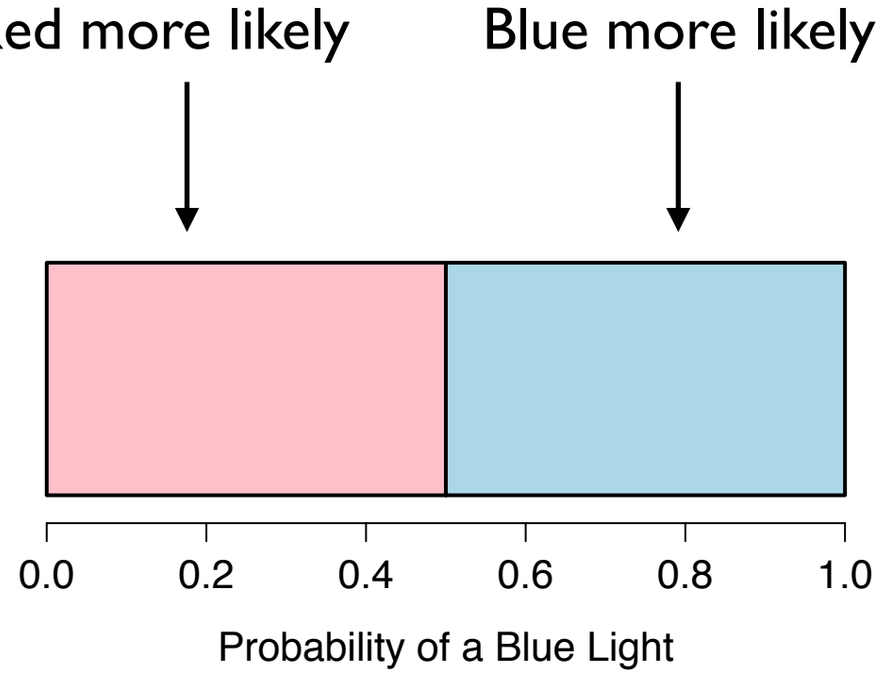
- Two kinds of environment
  - Static: Outcome probabilities are *fixed*
  - Dynamic: Outcome probabilities undergo *discrete* changes

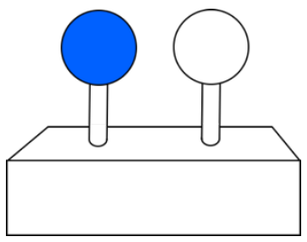
How does a rational agent allocate behaviour in this task?

A simple Bayesian analysis of the beliefs the agent holds

Prior beliefs about the probability that the light will be blue

$$P(\theta)$$

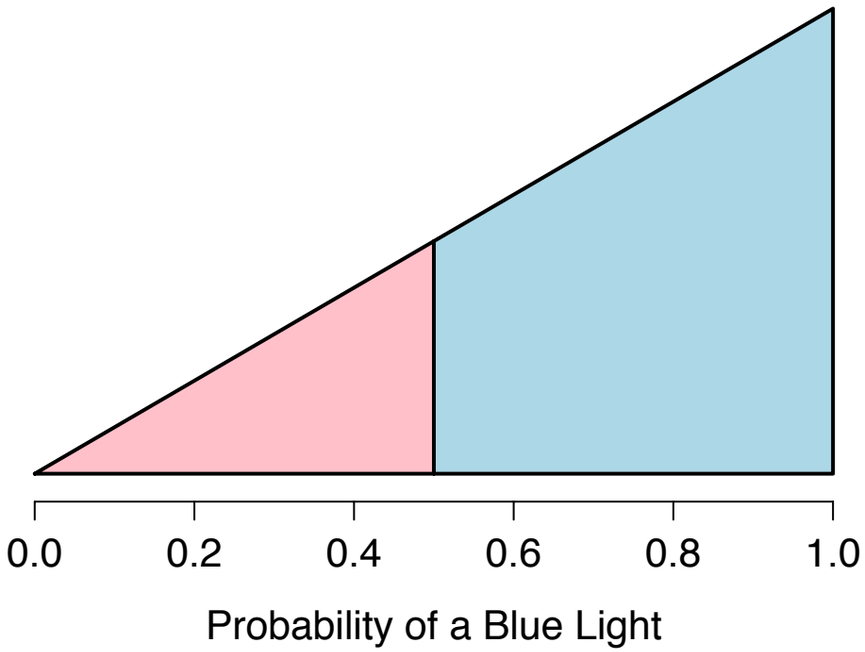


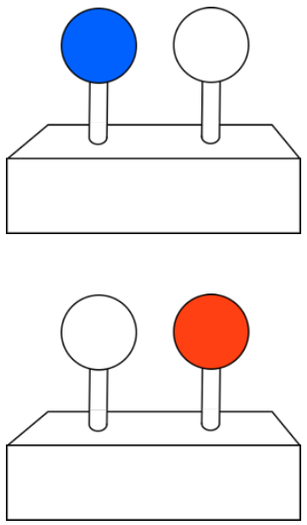


Posterior beliefs given a single
OBSERVE action on trial 1

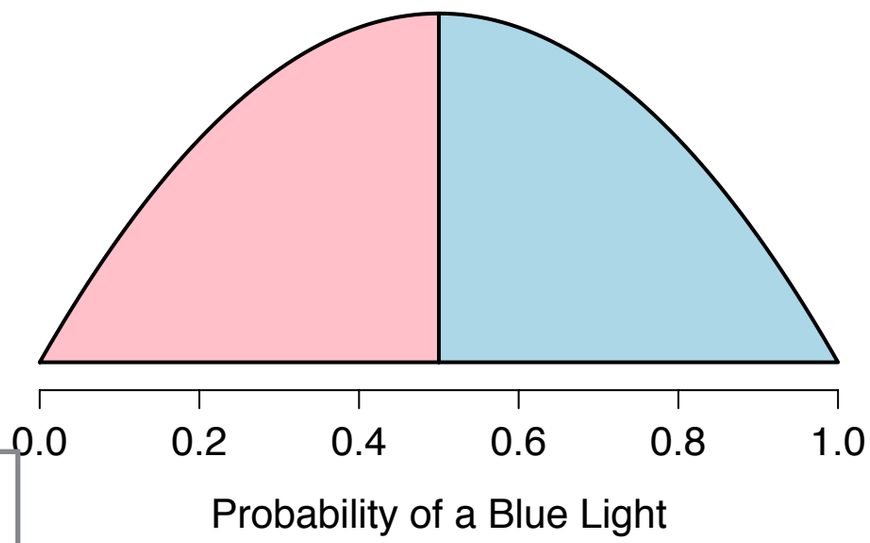
$$P(\theta|x) \propto P(x|\theta)P(\theta)$$

Confidence in Blue = 75%



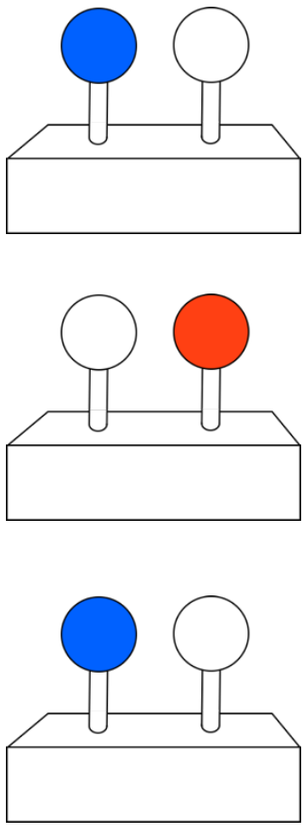


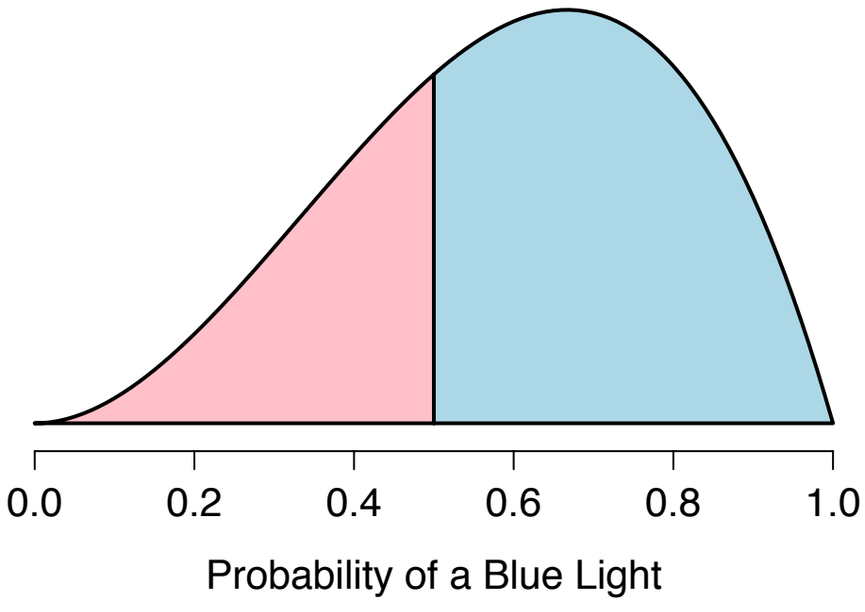
Confidence in Blue = 50%

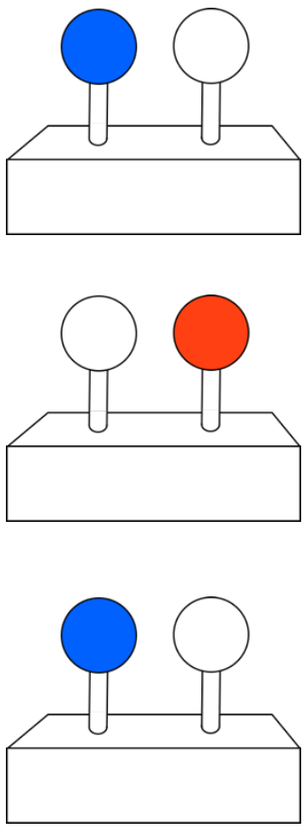


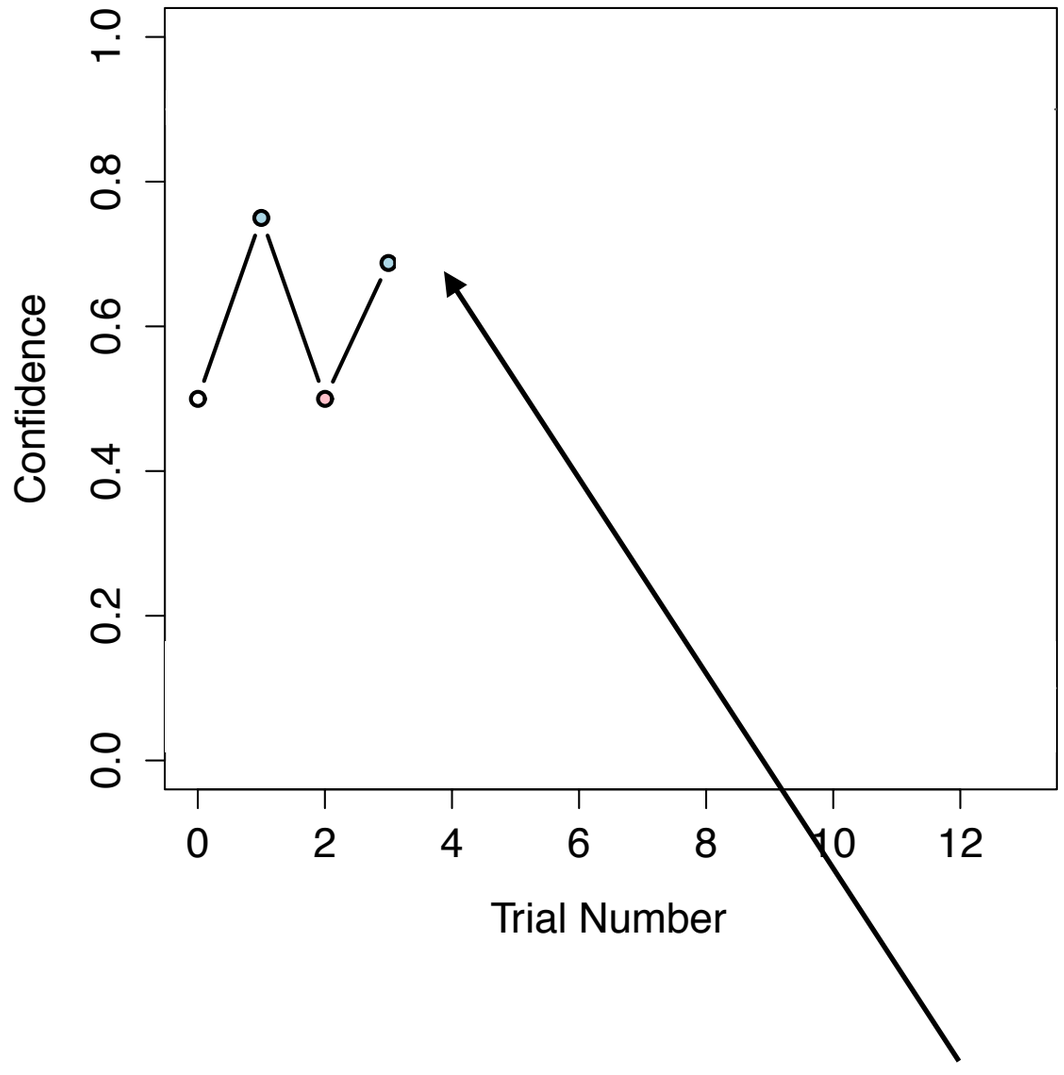
Beliefs updated sequentially: today's posterior is tomorrow's prior

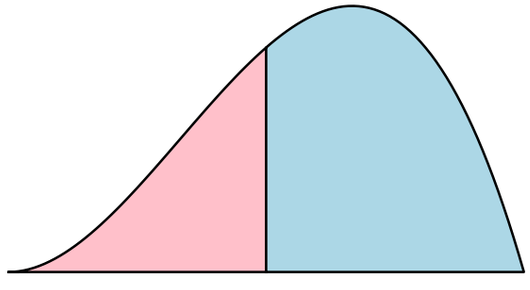
$$P(\theta|\mathbf{x}_t) \propto P(x_t|\theta)P(\theta|\mathbf{x}_{t-1})$$



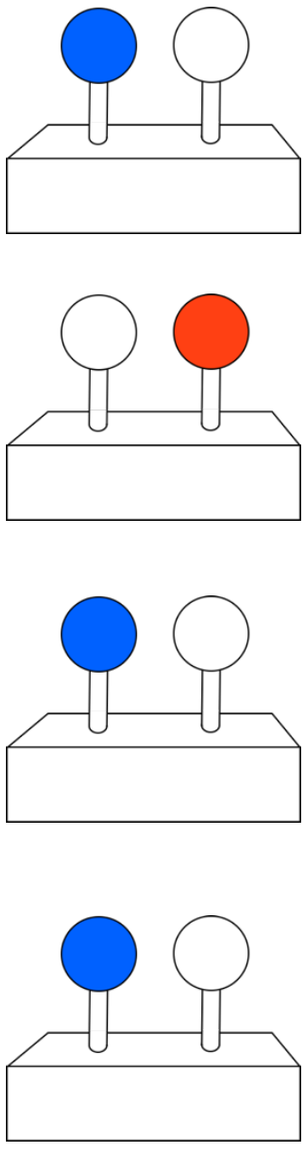

Confidence in Blue = 69%
0.0
0.2
0.4
0.6
0.8
1.0
Probability of a Blue Light

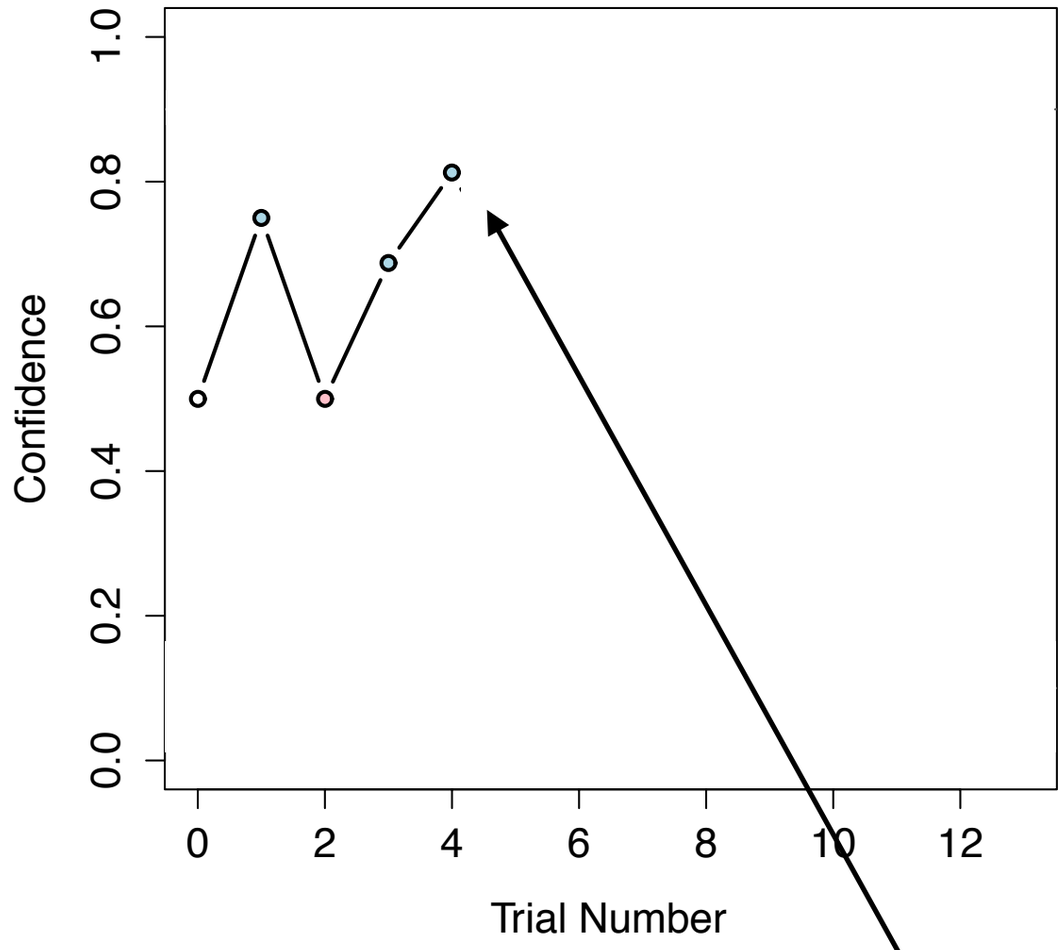


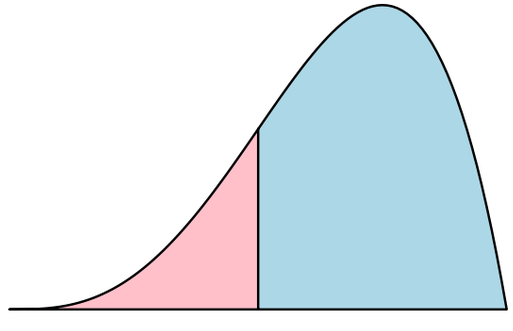


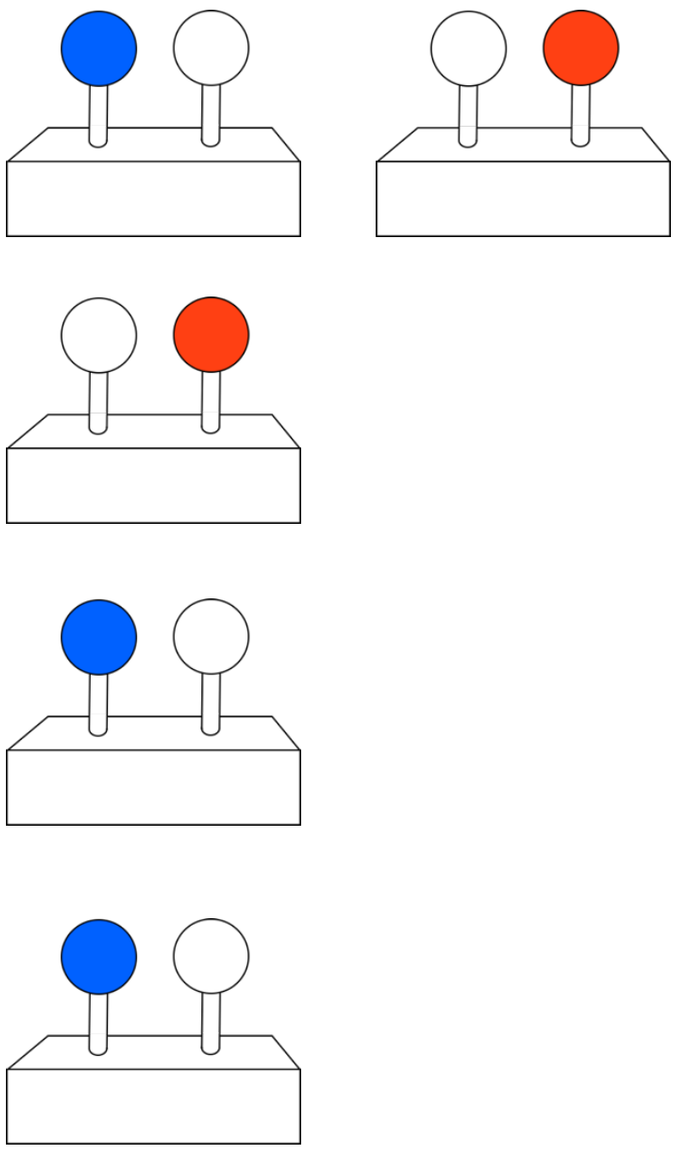


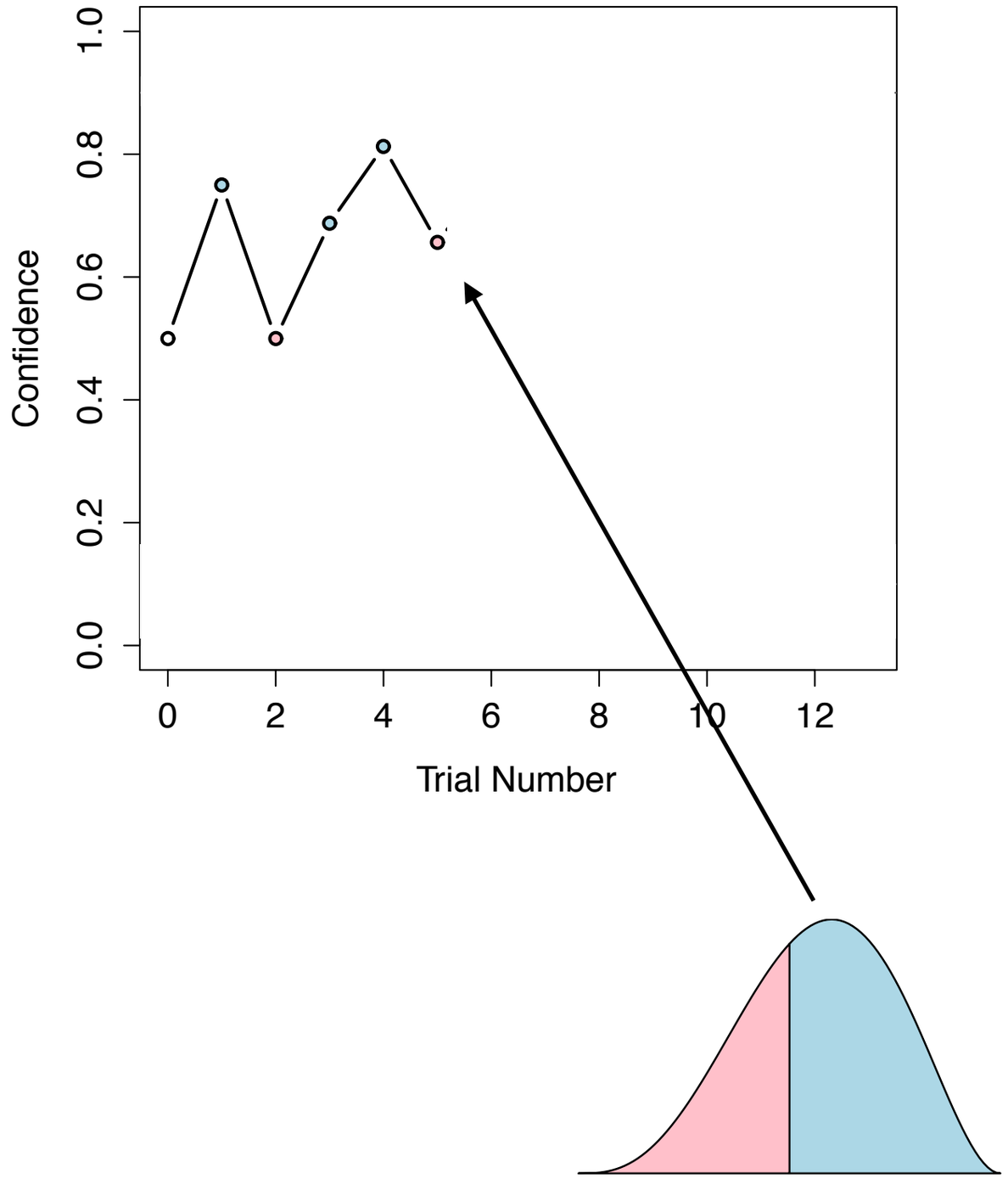
Plot the learner's confidence over time, as more observations are requested

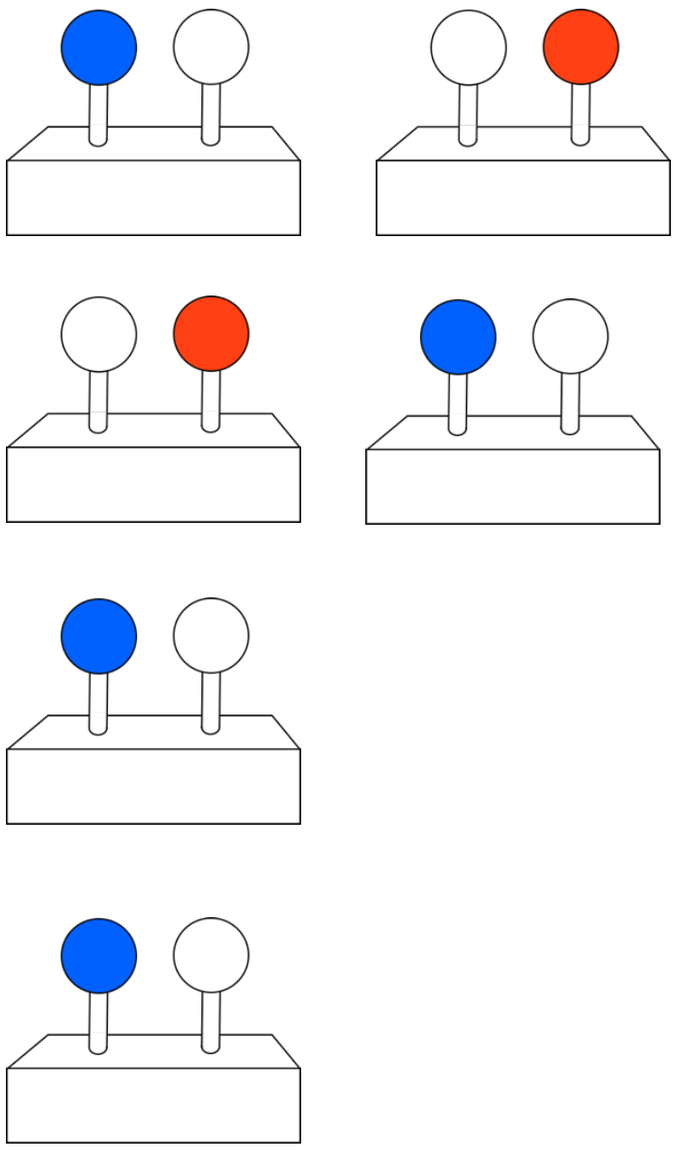


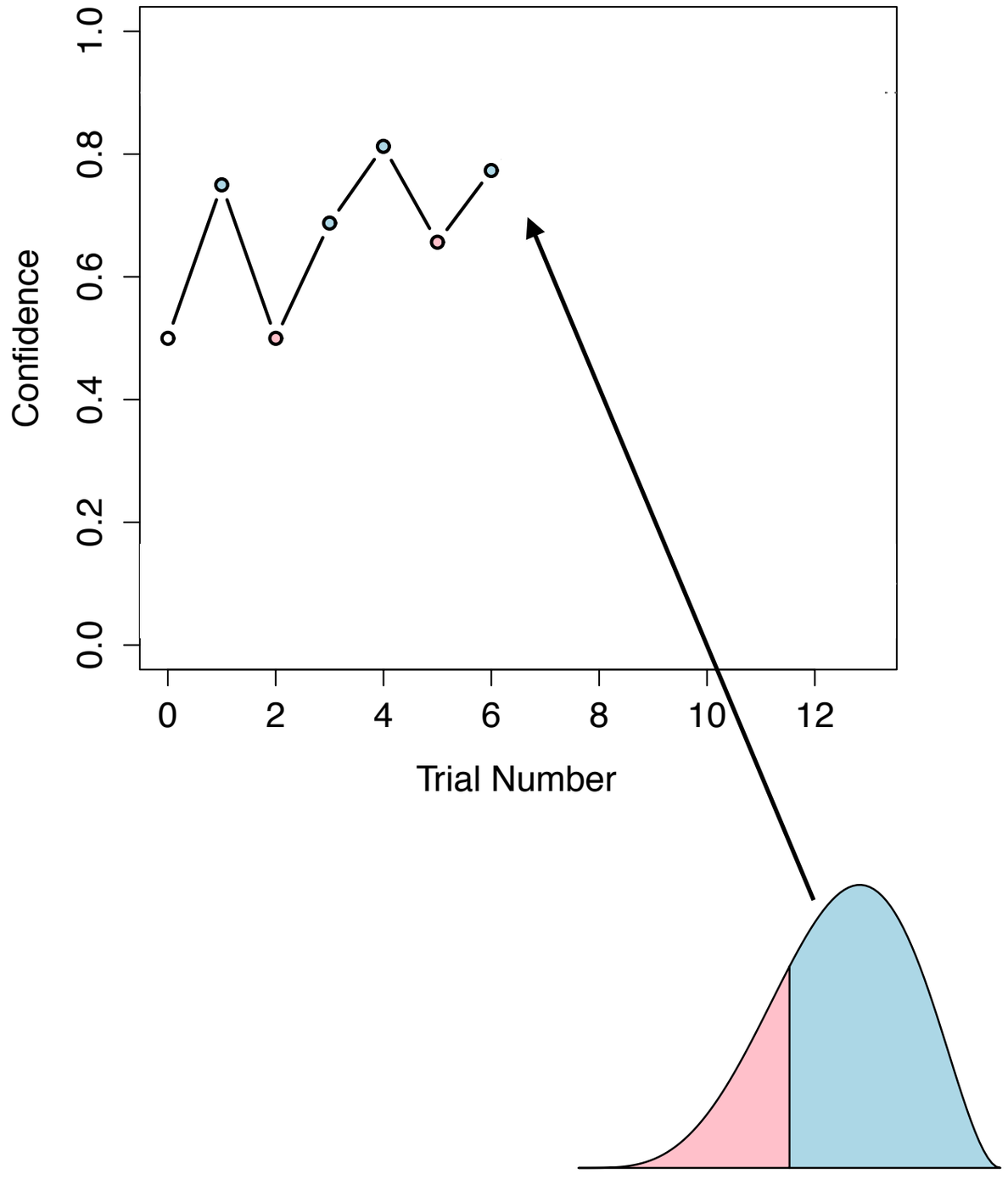

Confidence
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
12
Trial Number

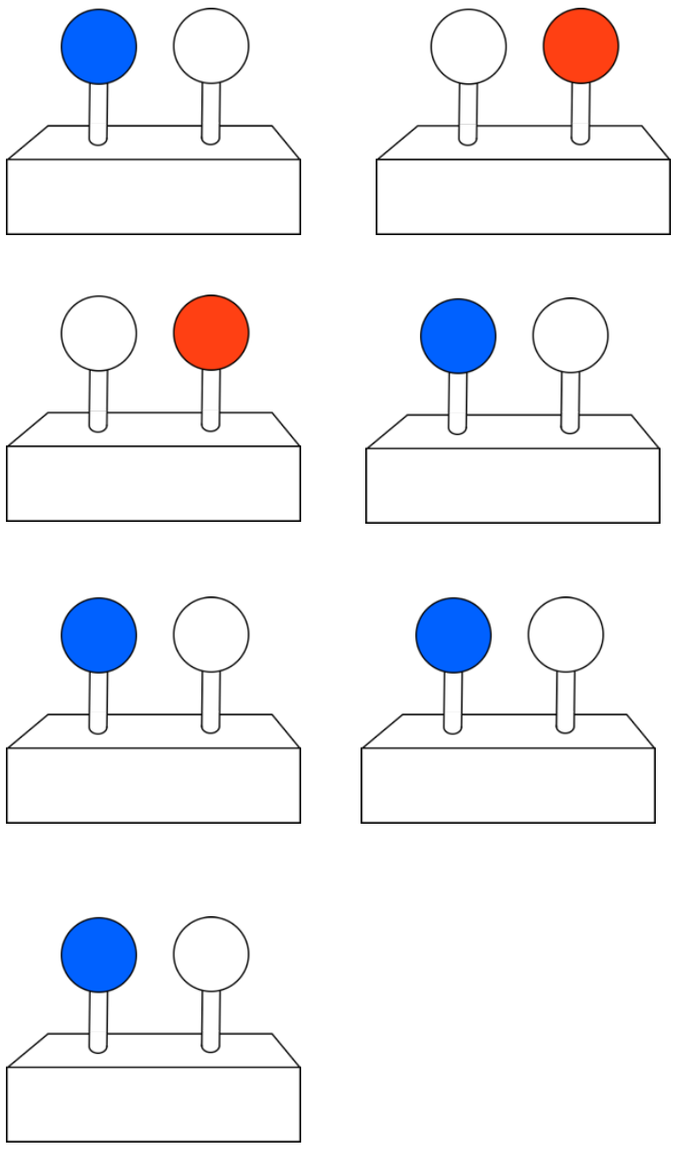


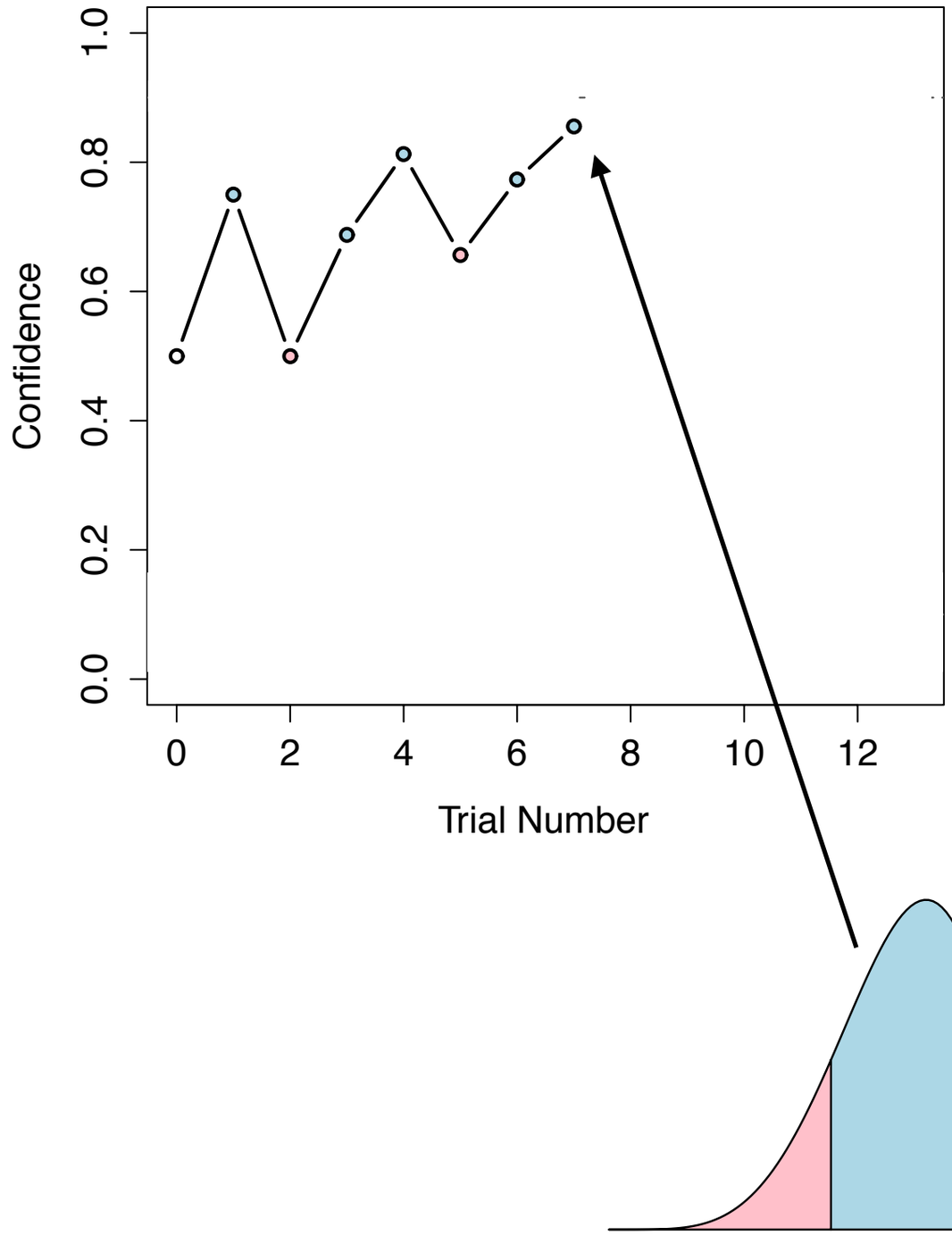



Confidence
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
12
Trial Number

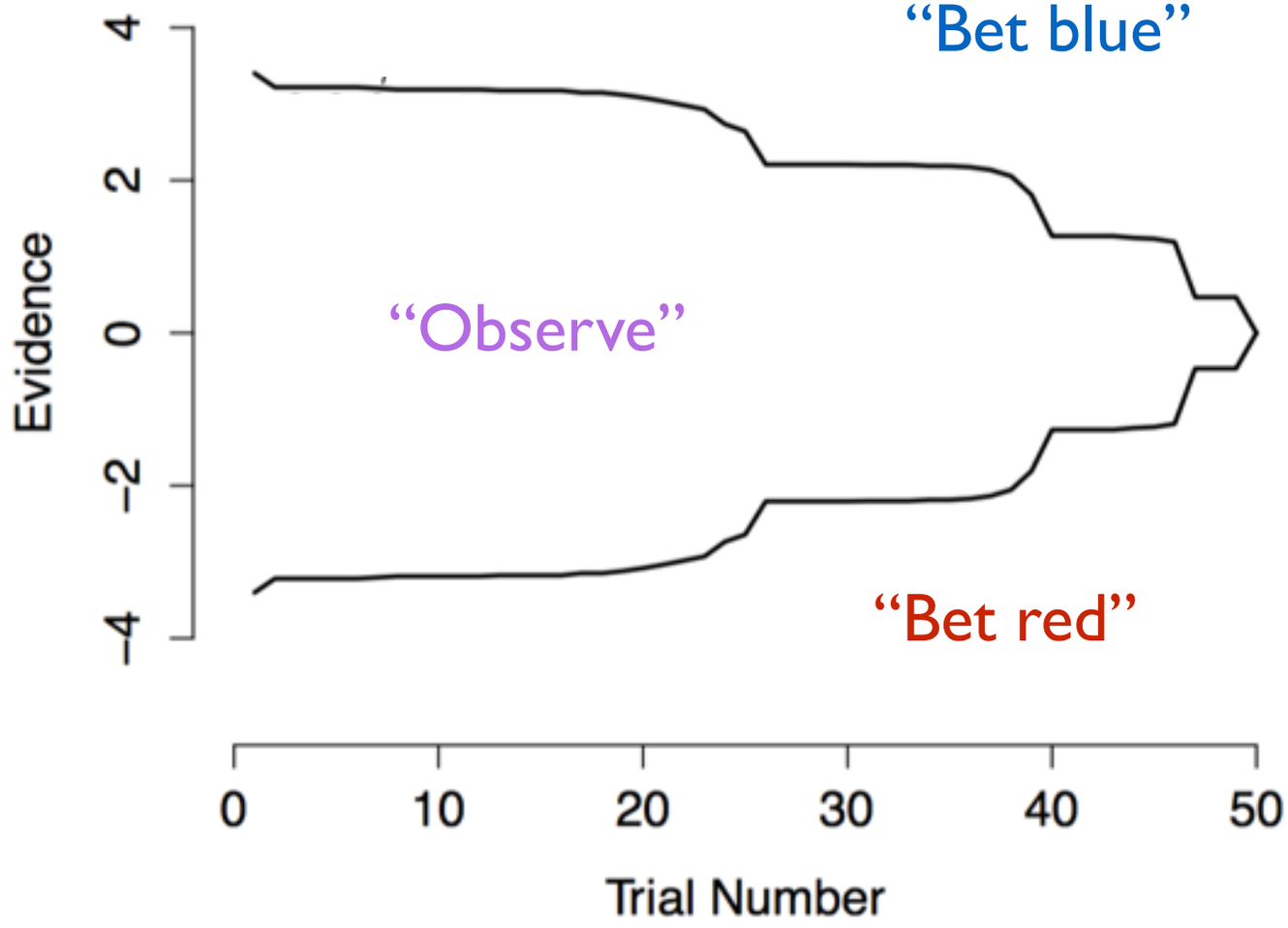



Confidence
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
12
Trial Number



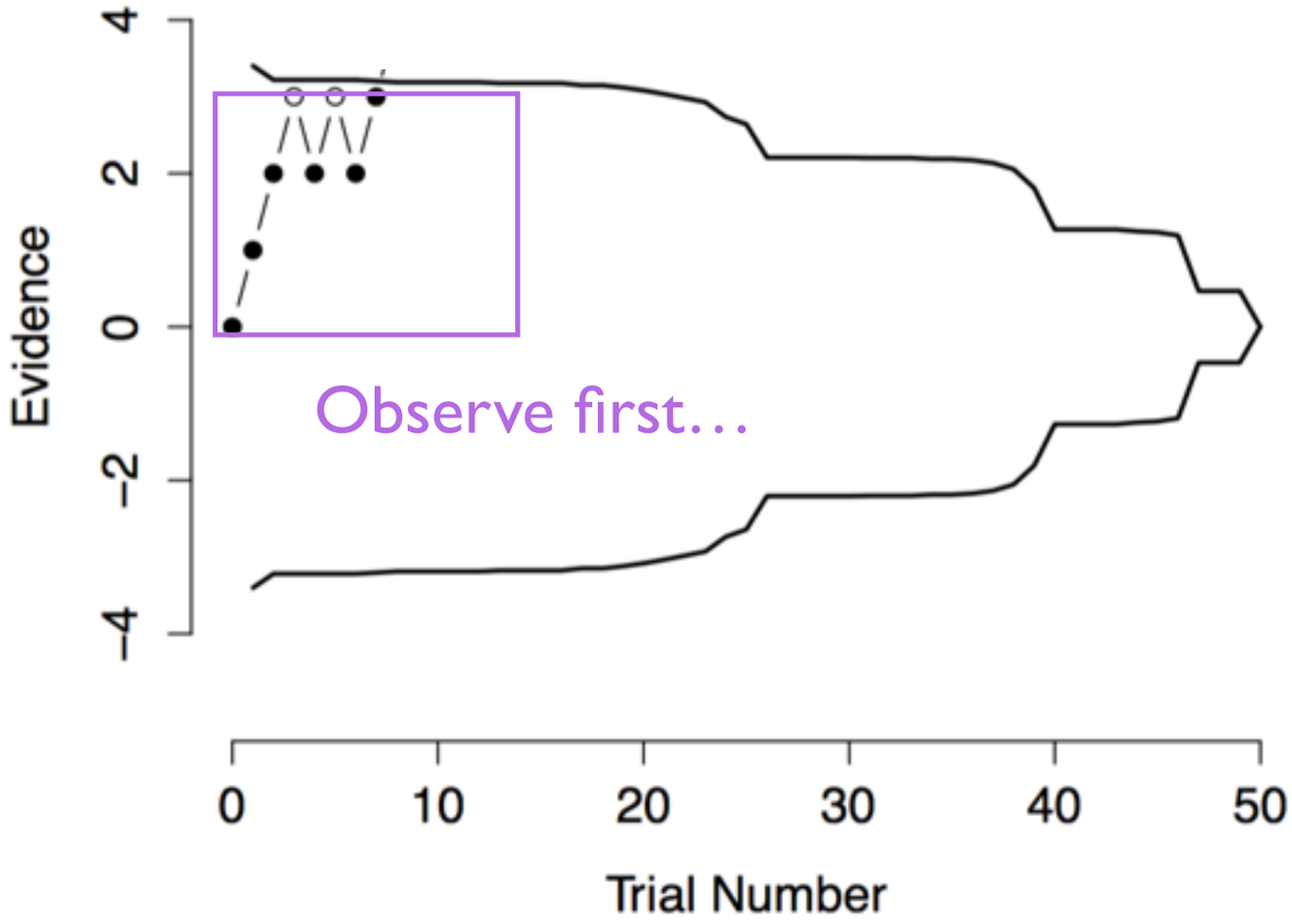

Confidence
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
12
Trial Number

Optimal decision policy for time-homogeneous problems

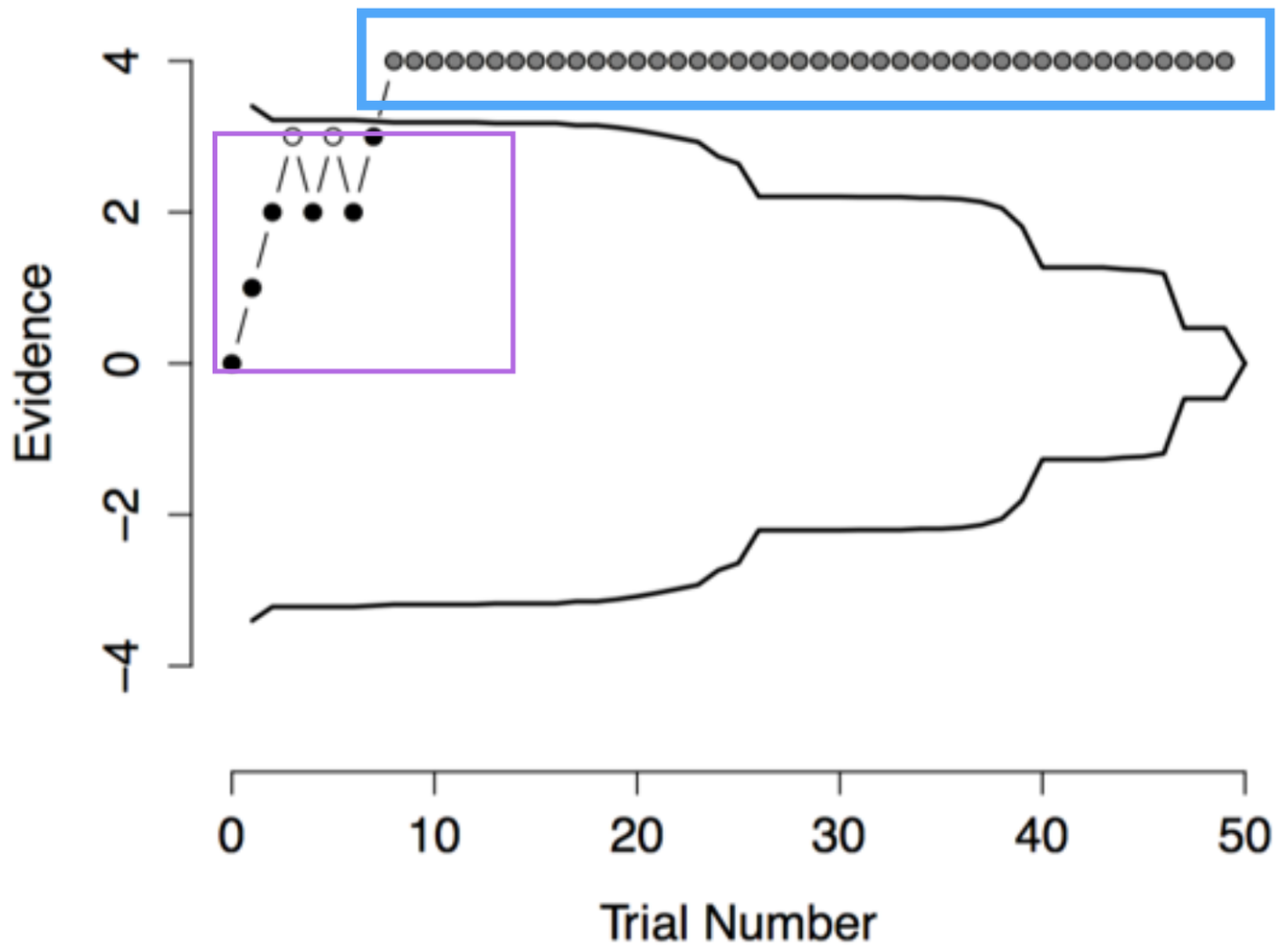


Bellman equation over belief states:

$$u(\boldsymbol{b}_t) = r(\boldsymbol{b}_t) + \max_{a_t} \sum_{\boldsymbol{b}_{t+1}} u(\boldsymbol{b}_{t+1}) P(\boldsymbol{b}_{t+1} | a_t, \boldsymbol{b}_t)$$


Evidence
4
2
0
−2
−4
Observe first…
0
10
20
30
40
50
Trial Number

… then bet (blindly) for the rest



“Optimal” policy: all the observations are front loaded…

Humans don't do this…

(Tversky & Edwards 1966)

Optimal policy: all observations are "front loaded"…

O O O O O O O B B B B B B B B B B B B B B B B

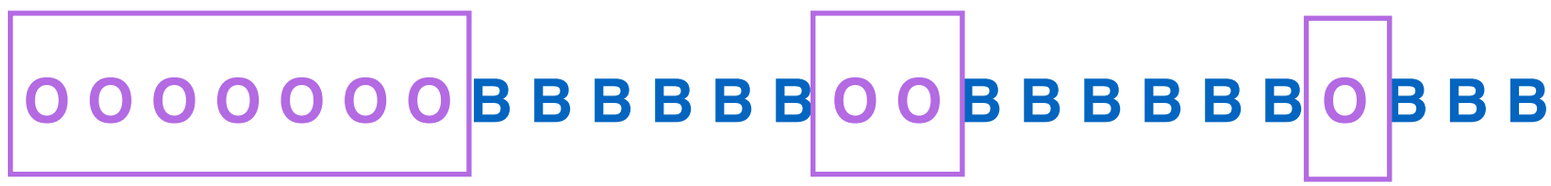
Humans don't do this…

(Tversky & Edwards 1966)

Optimal policy: all observations are "front loaded"…

O O O O O O O B B B B B B B B B B B B B B B B

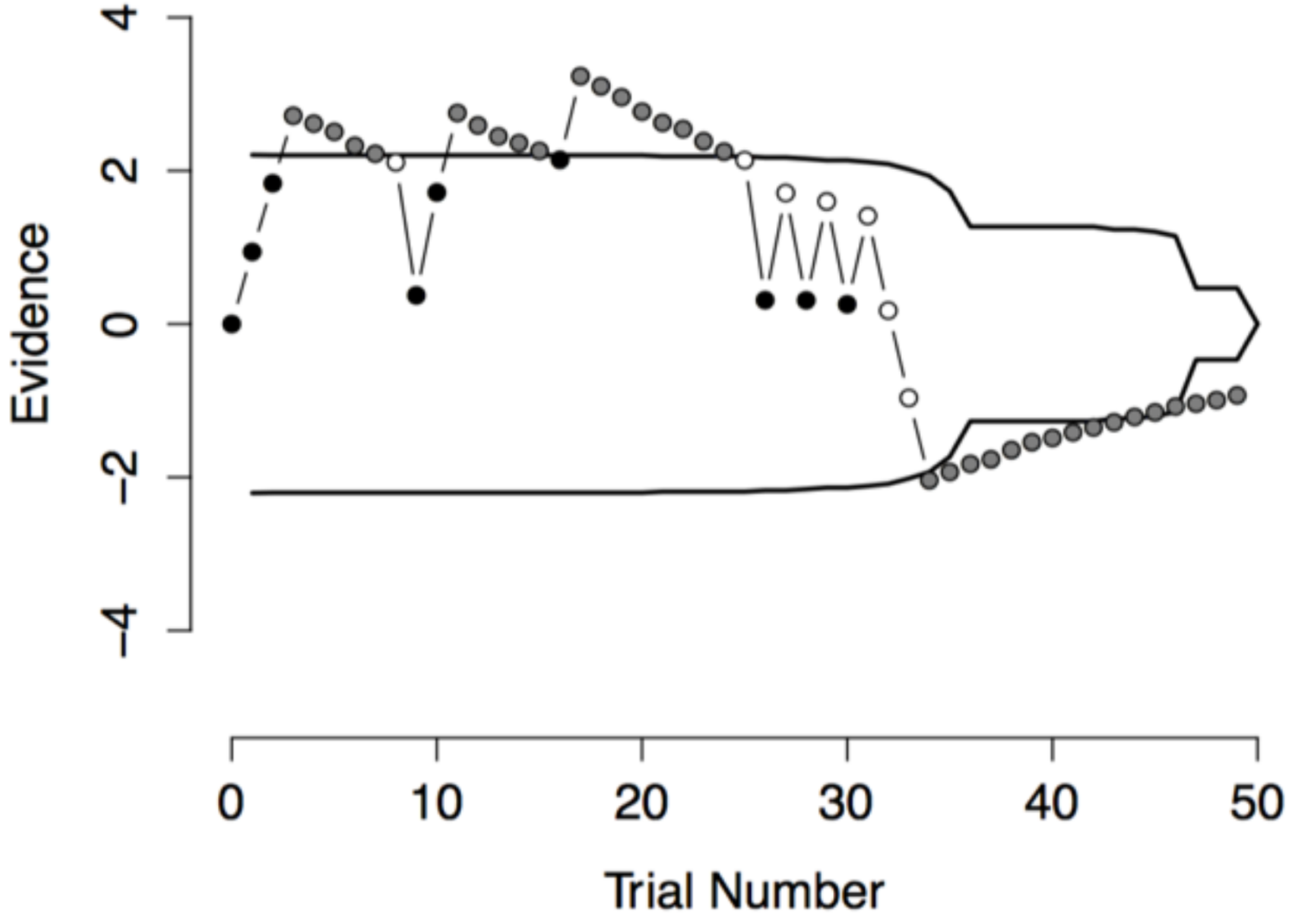
Typical human pattern shows switching:
only some observations get front loaded…



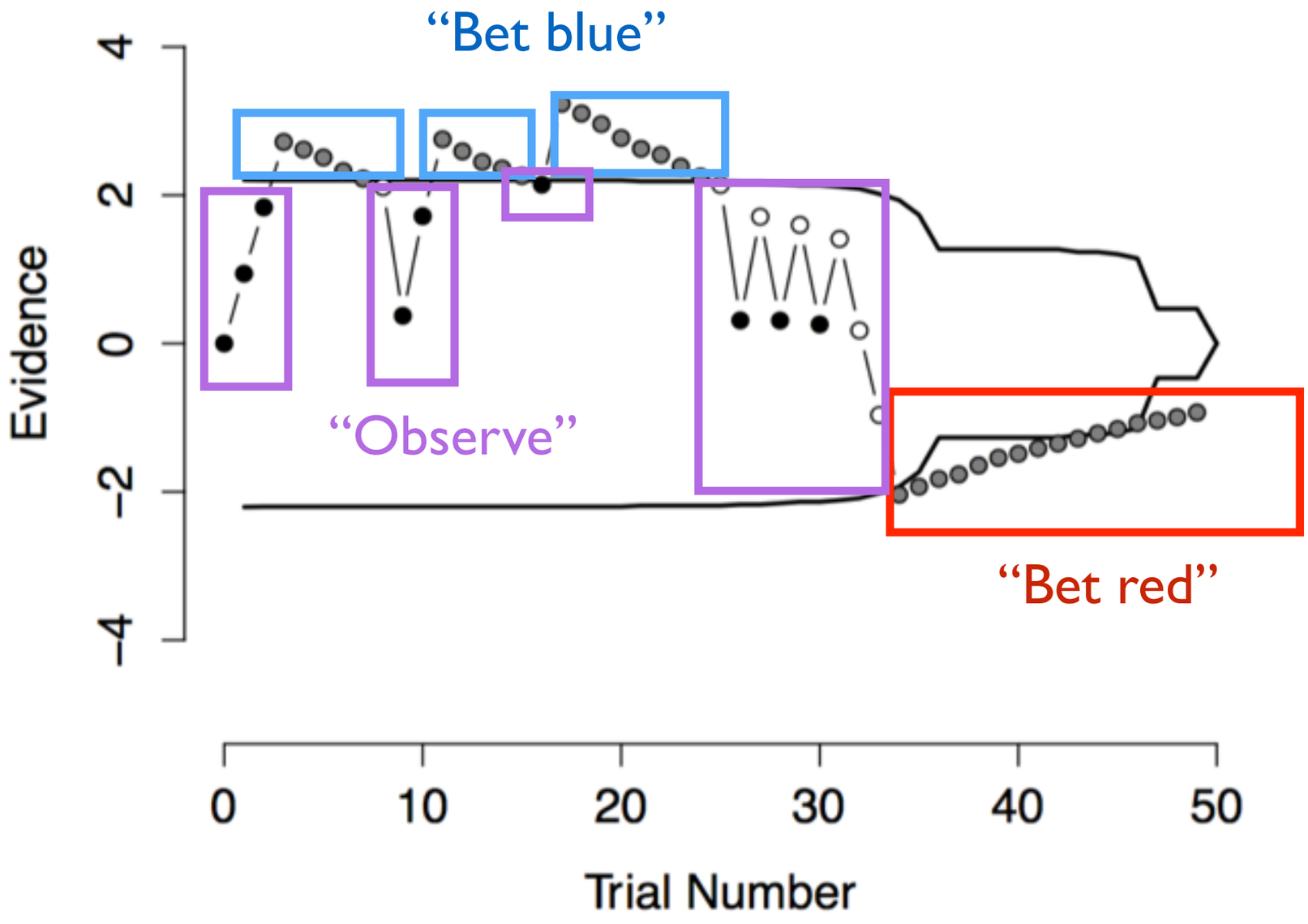
… so either we're stupid or we are solving a different problem

Optimal decision policy for* time-*inhomogeneous* problems

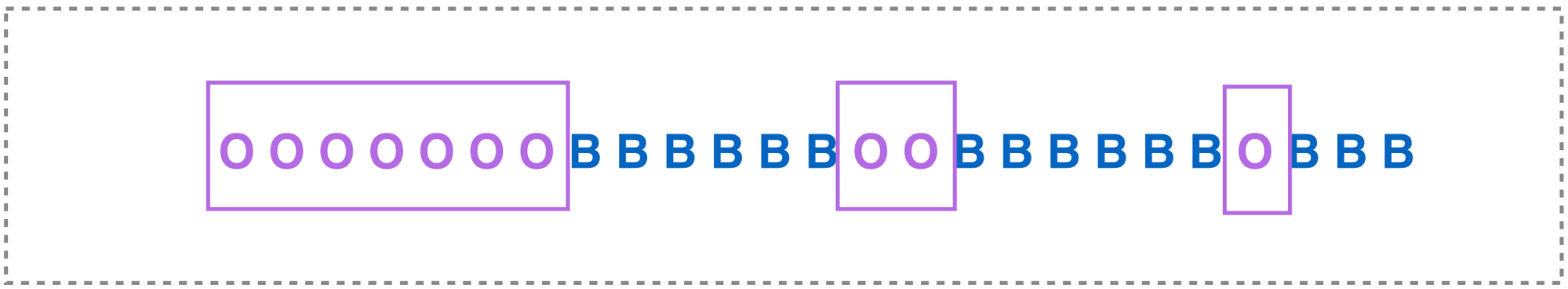
(* a specific class of)


Evidence
4
2
0
−2
−4
0
10
20
30
40
50
Trial Number

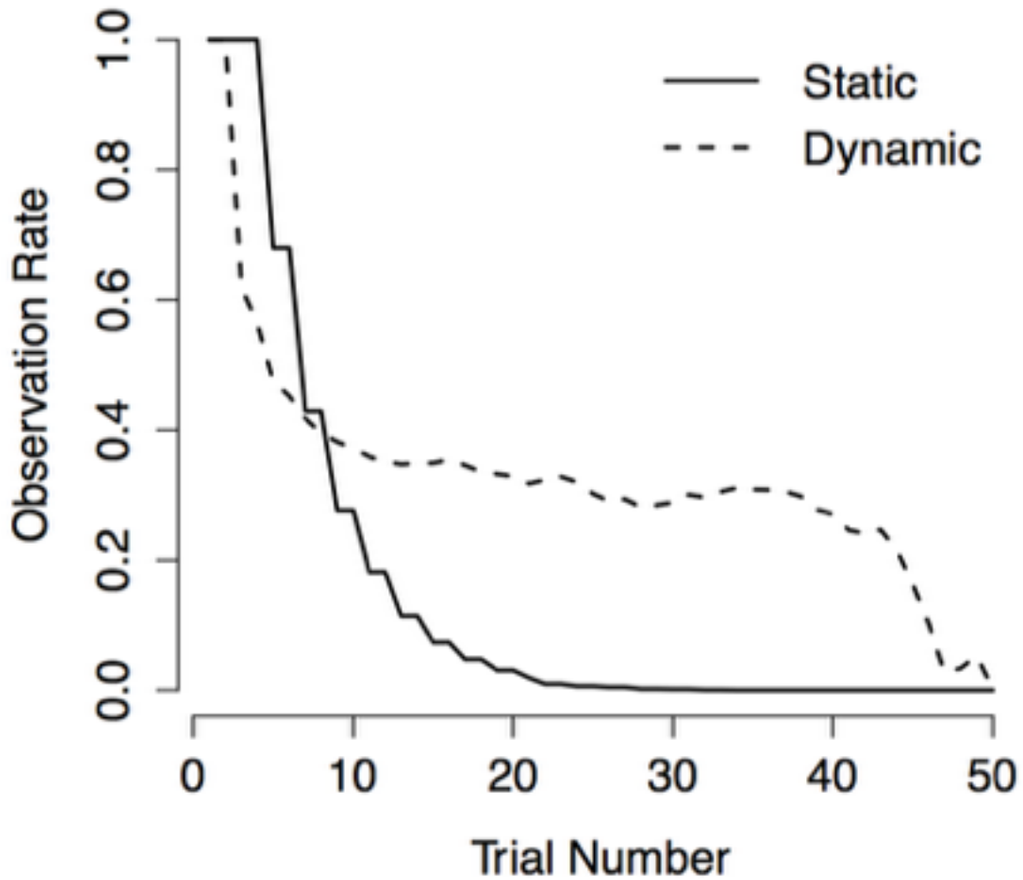
Older observations lose relevance, confidence decays, and the MDP looks more human-like



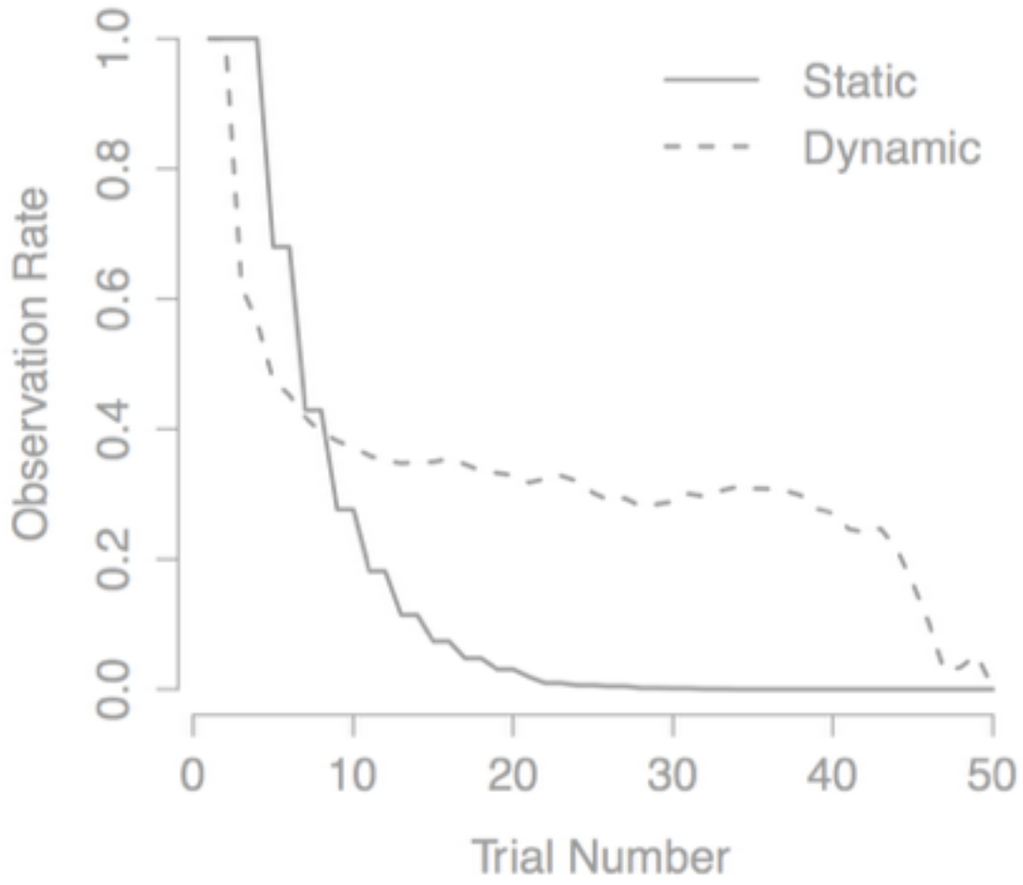
This pattern makes sense if the agent assumes that reward contingencies change over time

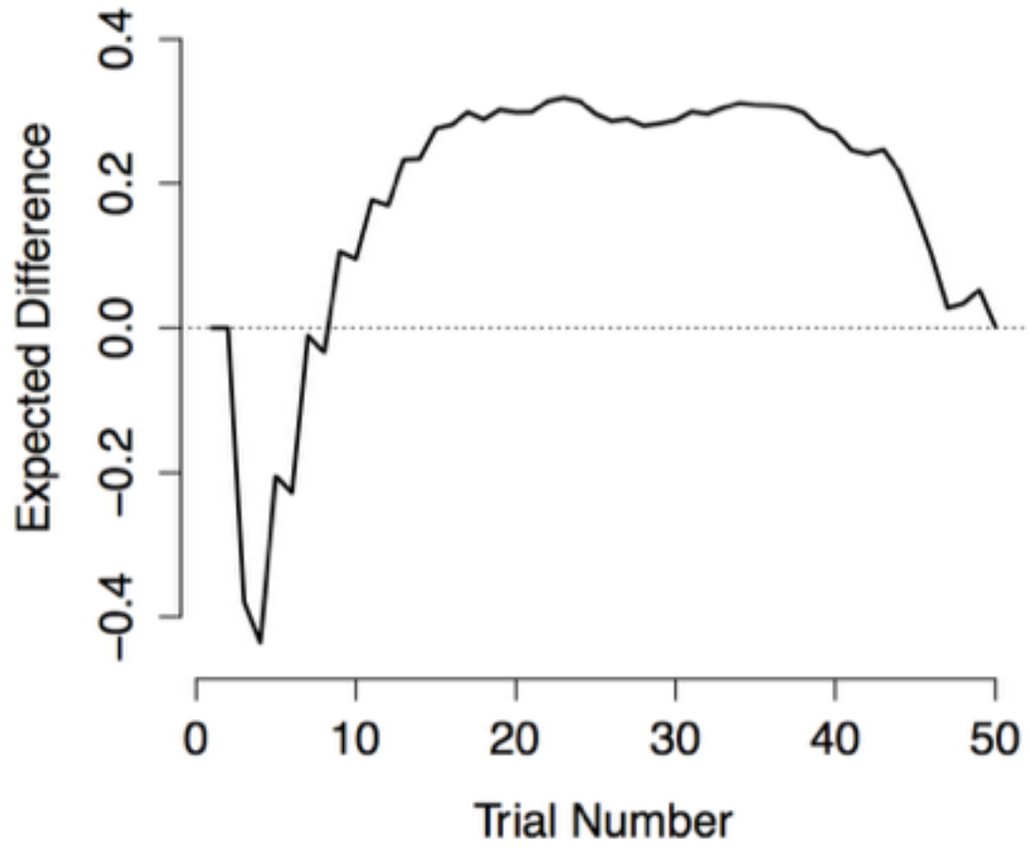


POMDP analysis predicts a qualitative shift in the observation pattern



Dynamic environments force a shift from observe to bet earlier… but switch back often

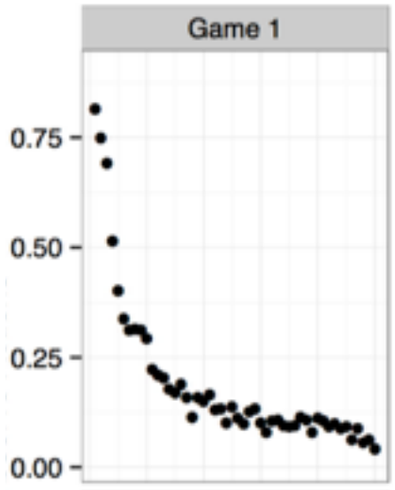




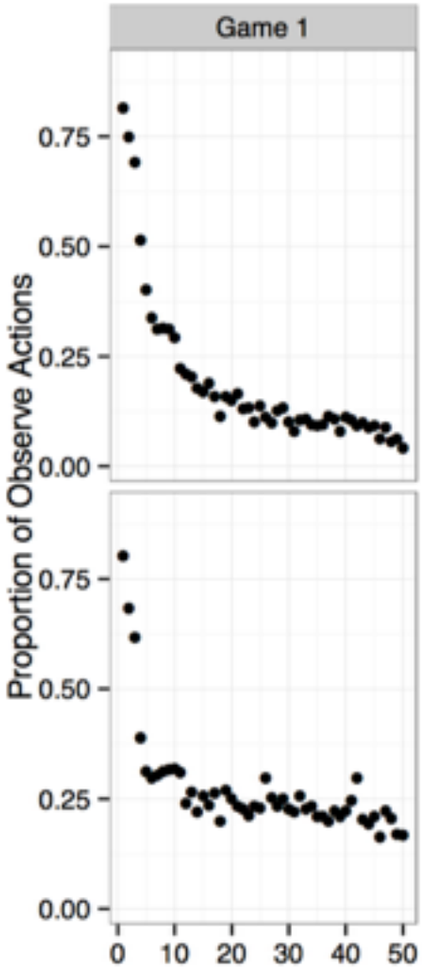
So we ran some experiments…

What do humans do?

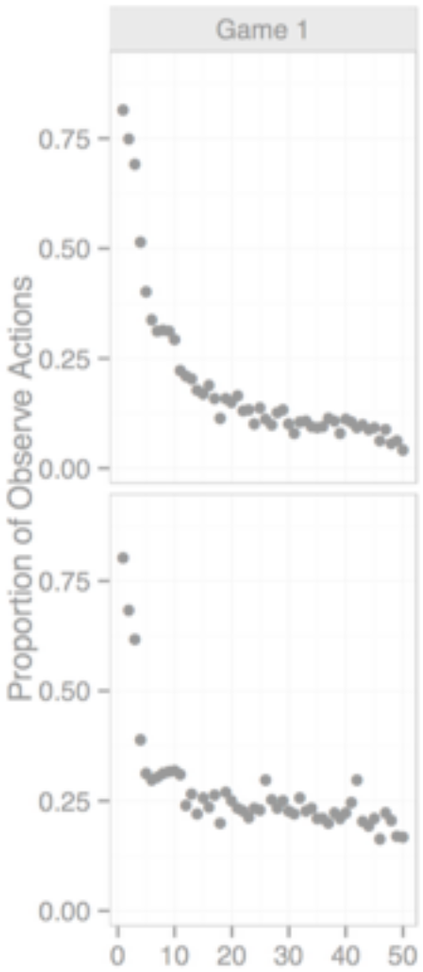
(614 participants on MTurk)



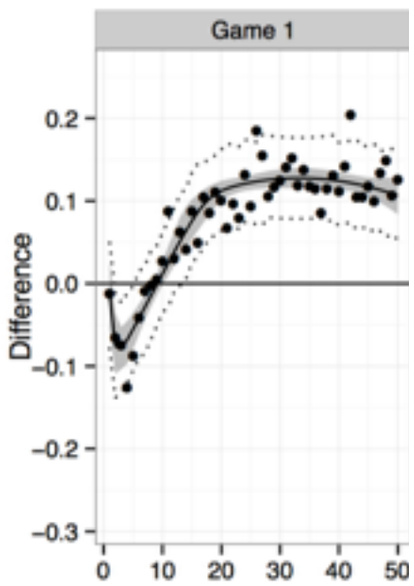
Probability of making an observation as a function of trial number, in a static environment

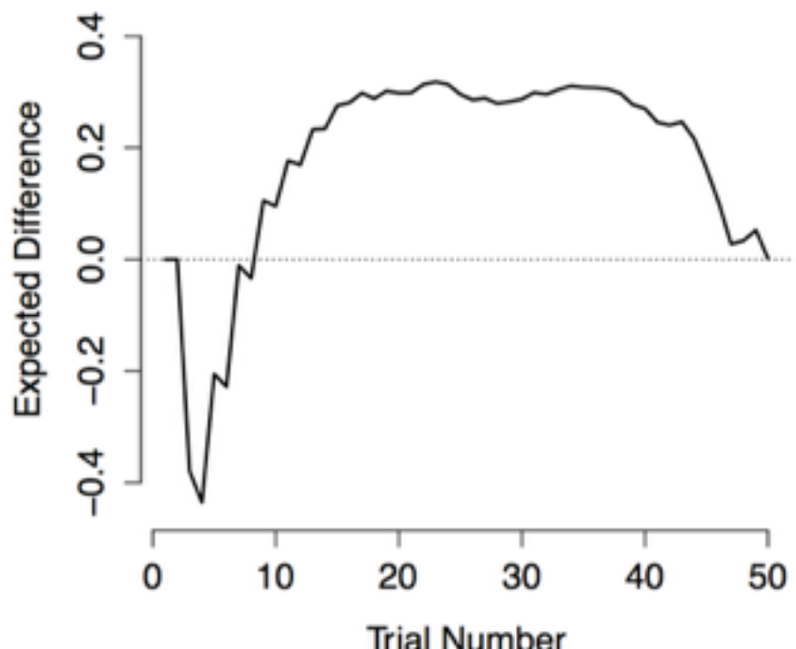


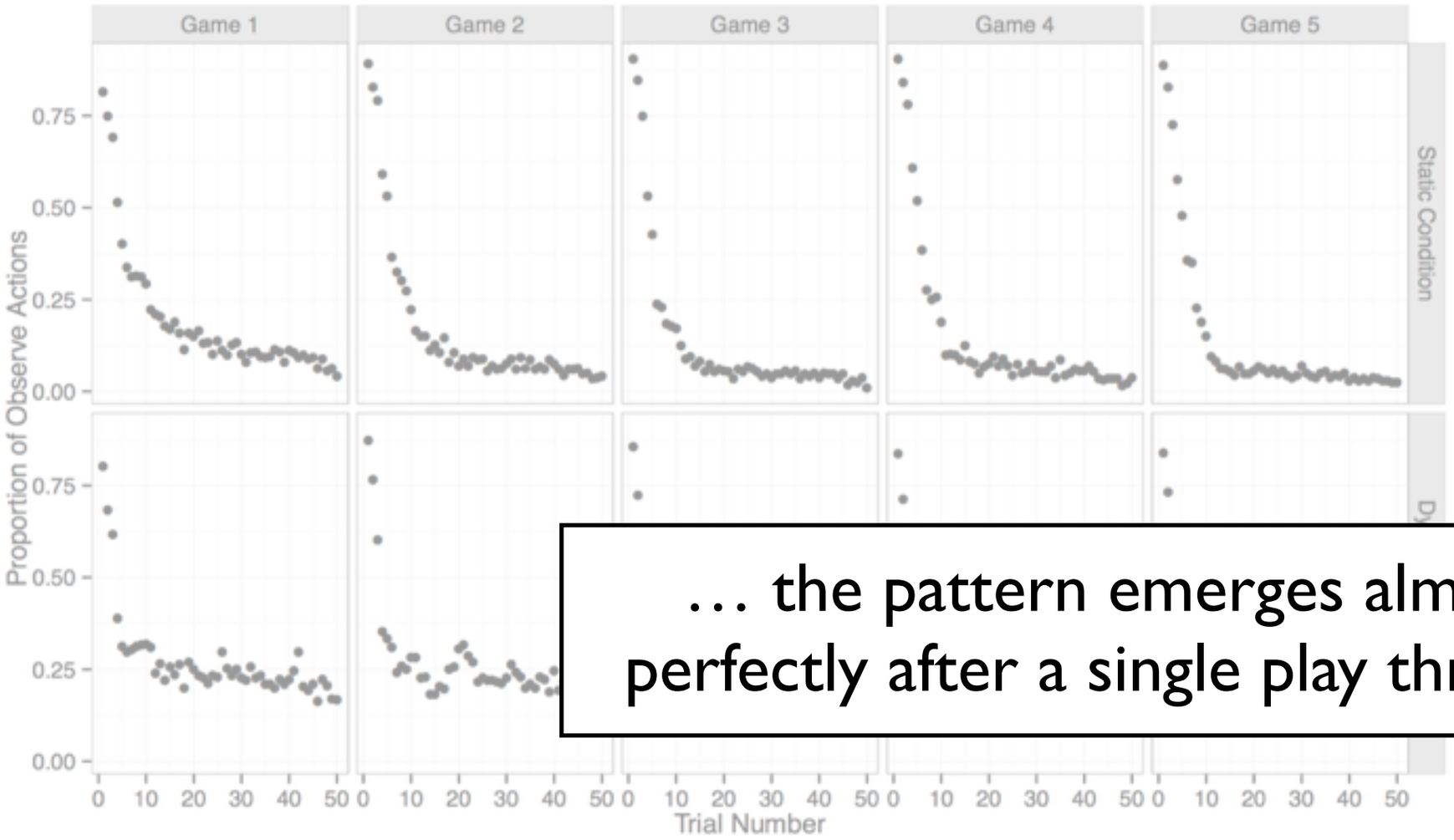
Probability of making an observation as a function of trial number, in a dynamic environment



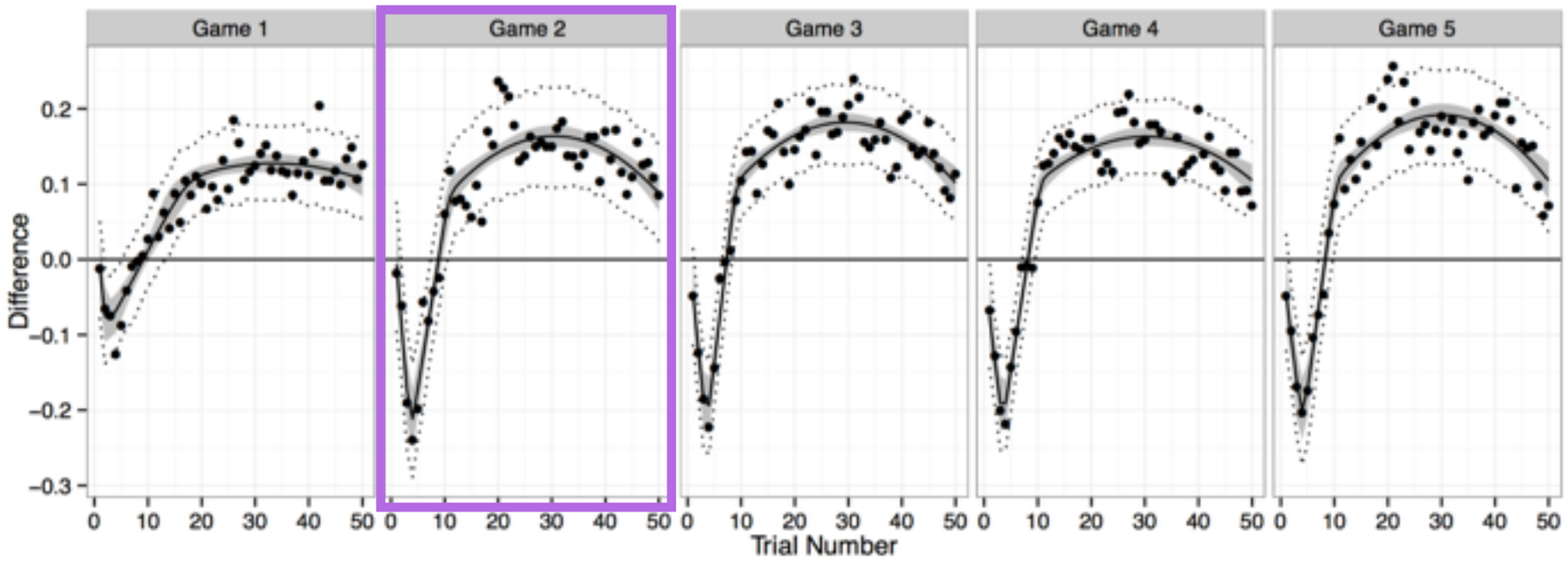
The difference between the two is *kind of* consistent with the POMDP analysis, but at first it's not convincing

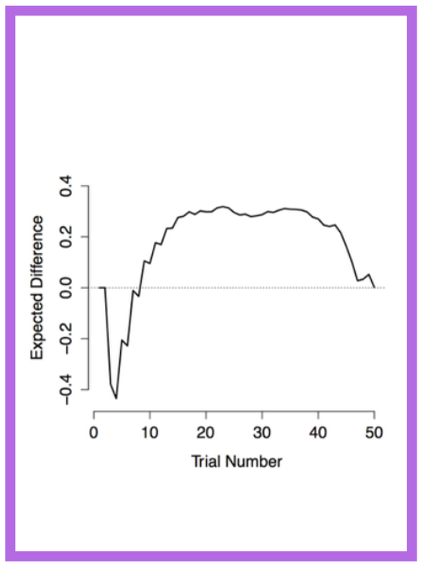




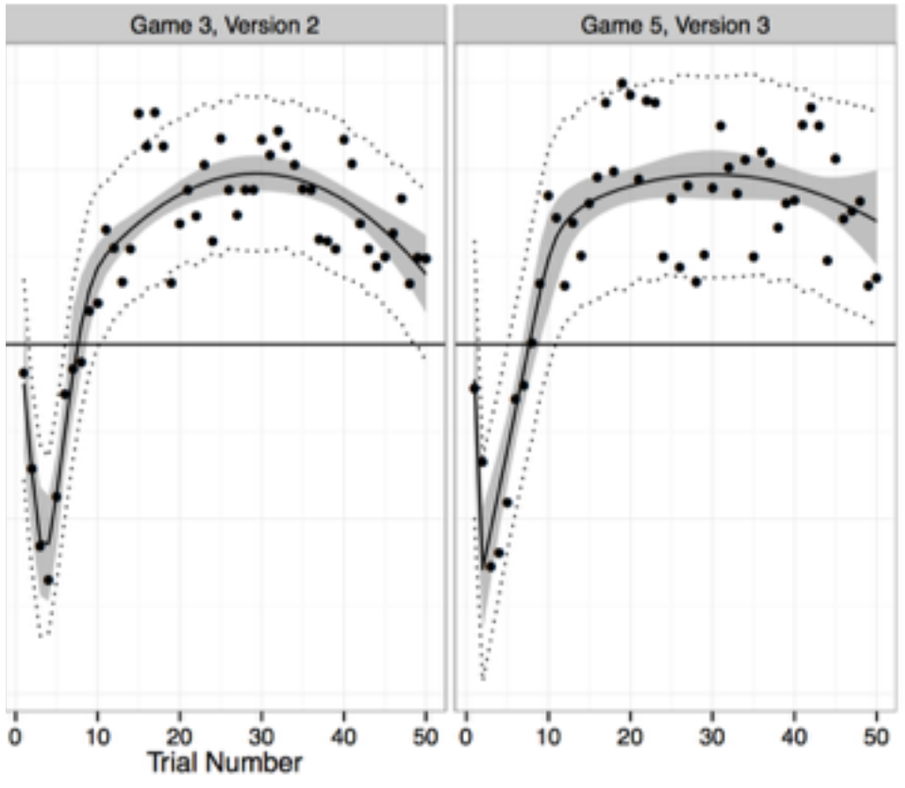


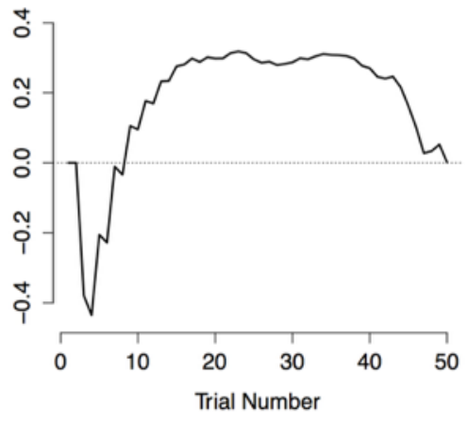
… the pattern emerges almost perfectly after a single play through



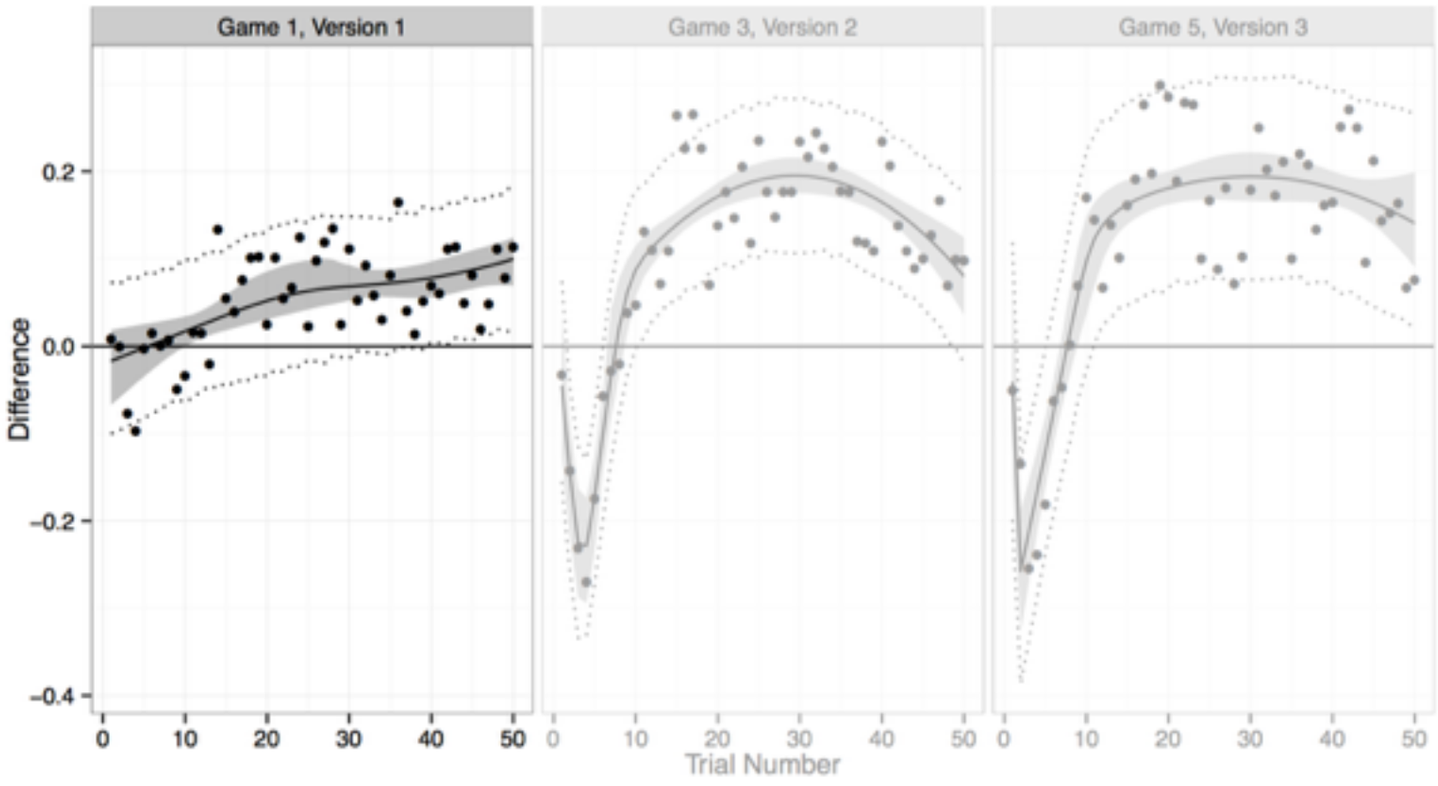


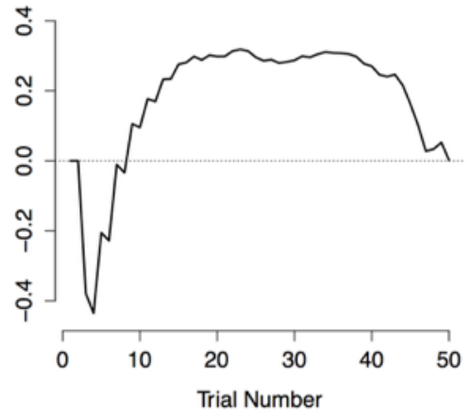
(Methodological control: in some instances the stimulus sequences were identical, and the effect still occurs driven solely by people's expectations…)





But not if the *only* difference is the instruction set



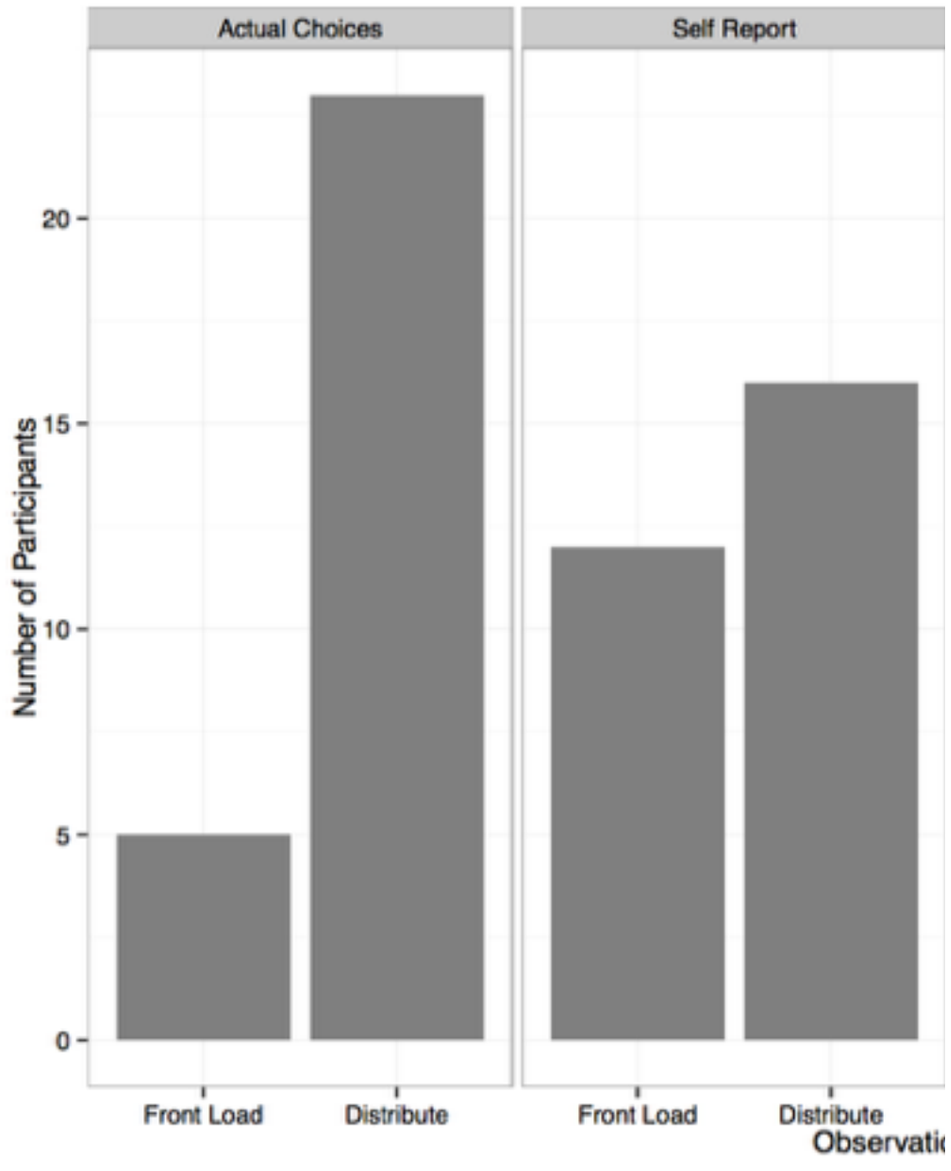


… people need *some* experience to work out what "static" vs "dynamic" really means here, but a *single* game is sufficient

What strategies do people follow and how do they adjust them?

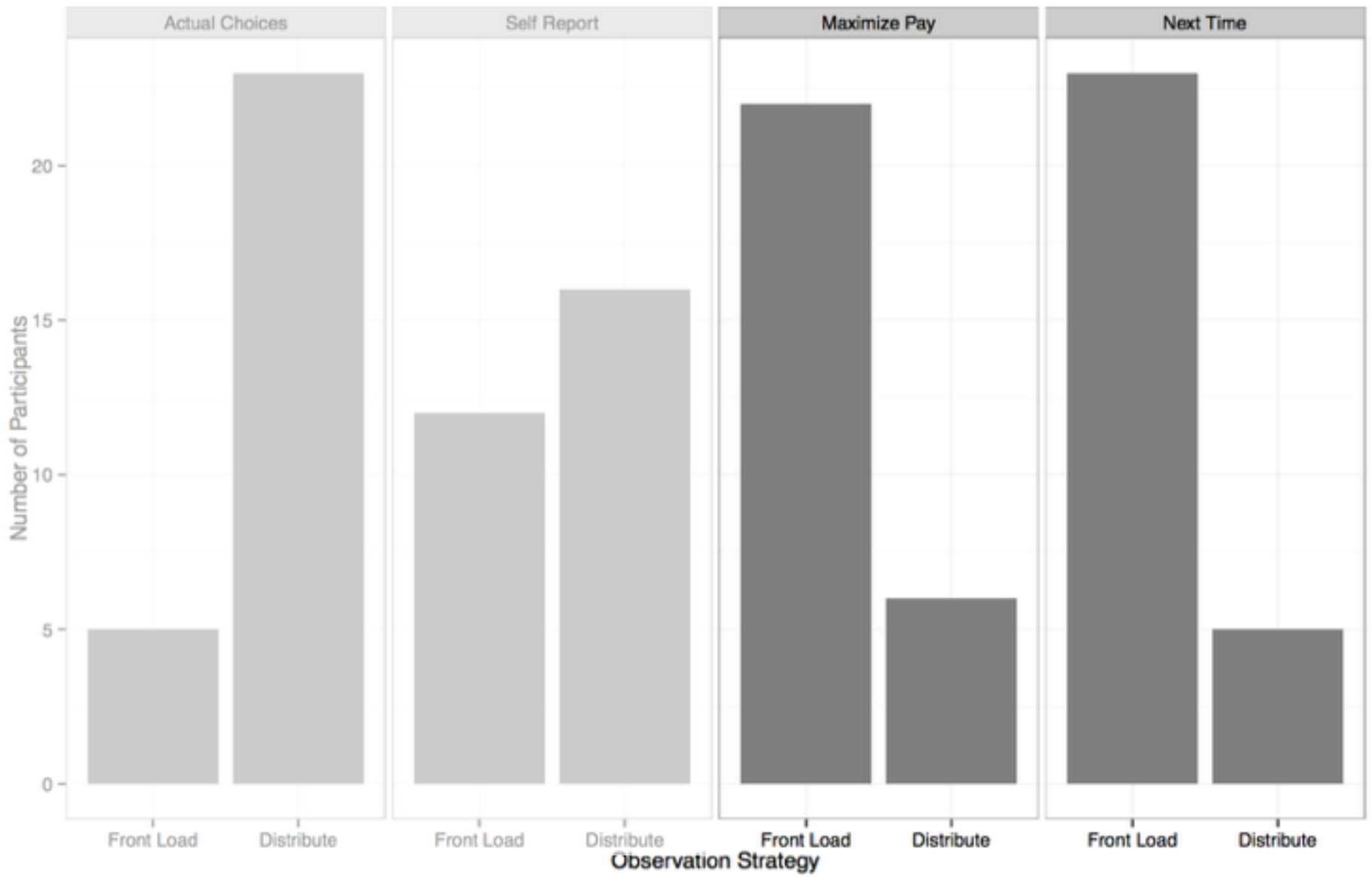
You could ask?

(<u>different</u> experiment, after game 1, static only)



People don't front load their observations, and they (mostly) know that

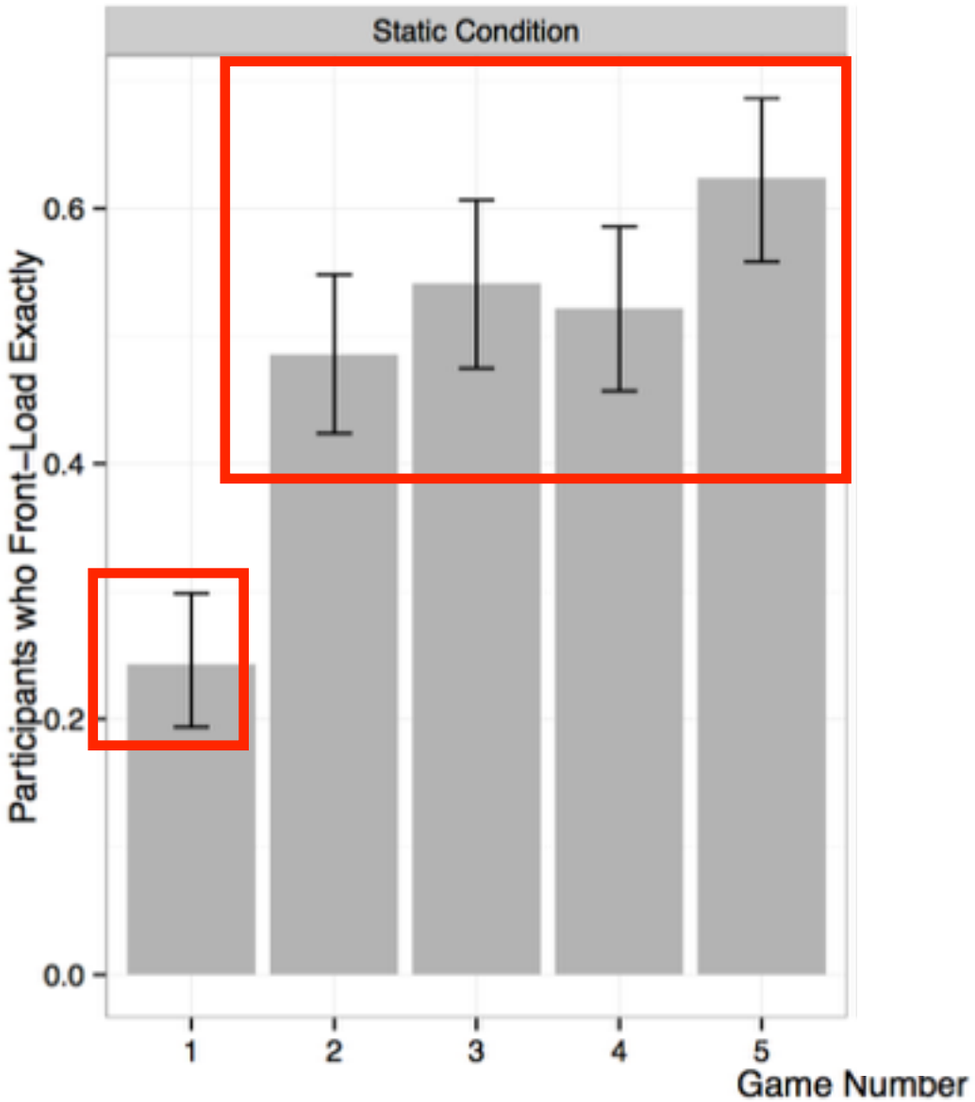
They recognise that front loading is optimal for the task and *claim* that's what they'll do next time…



And they do!

(back to the original expt)

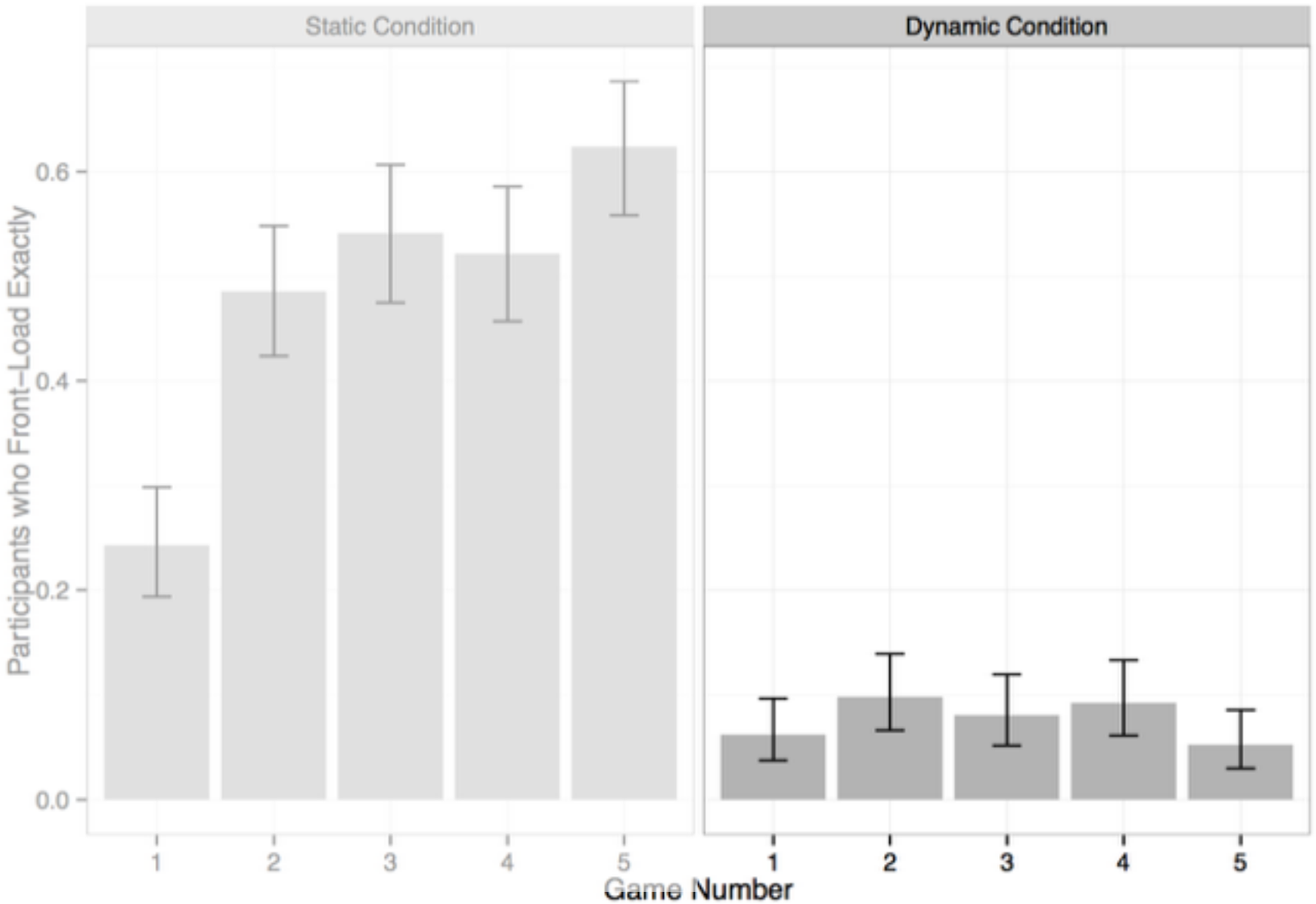
Static



… but only when relevant

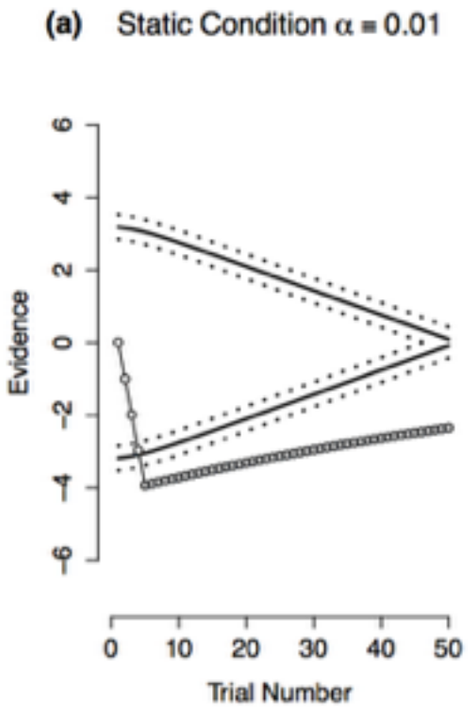
Static

Dynamic

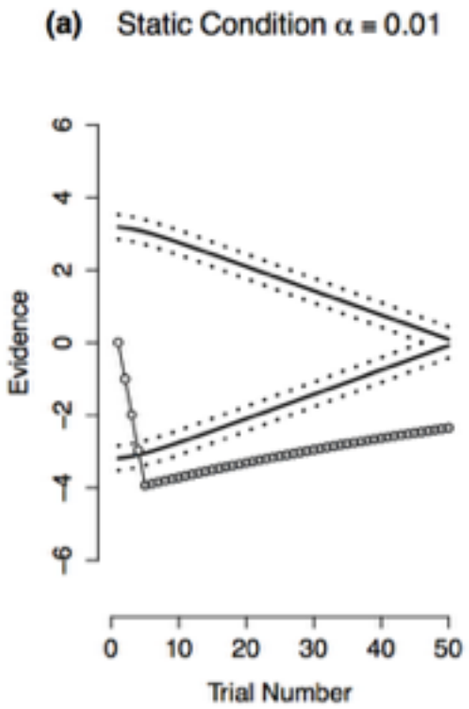


Estimating individual subject decision policies

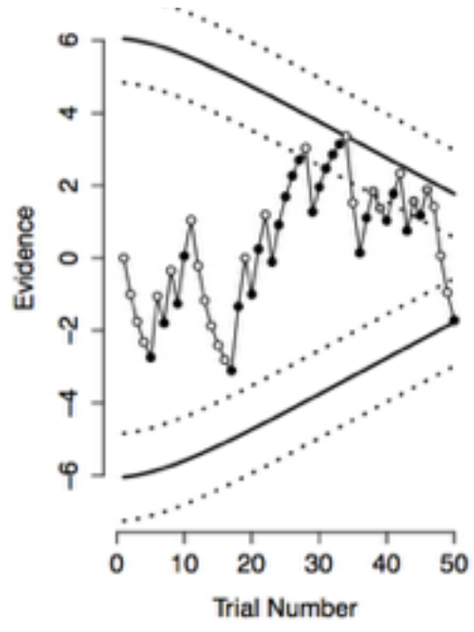
(using simpler evidence accumulation models
based loosely on drift diffusion models)



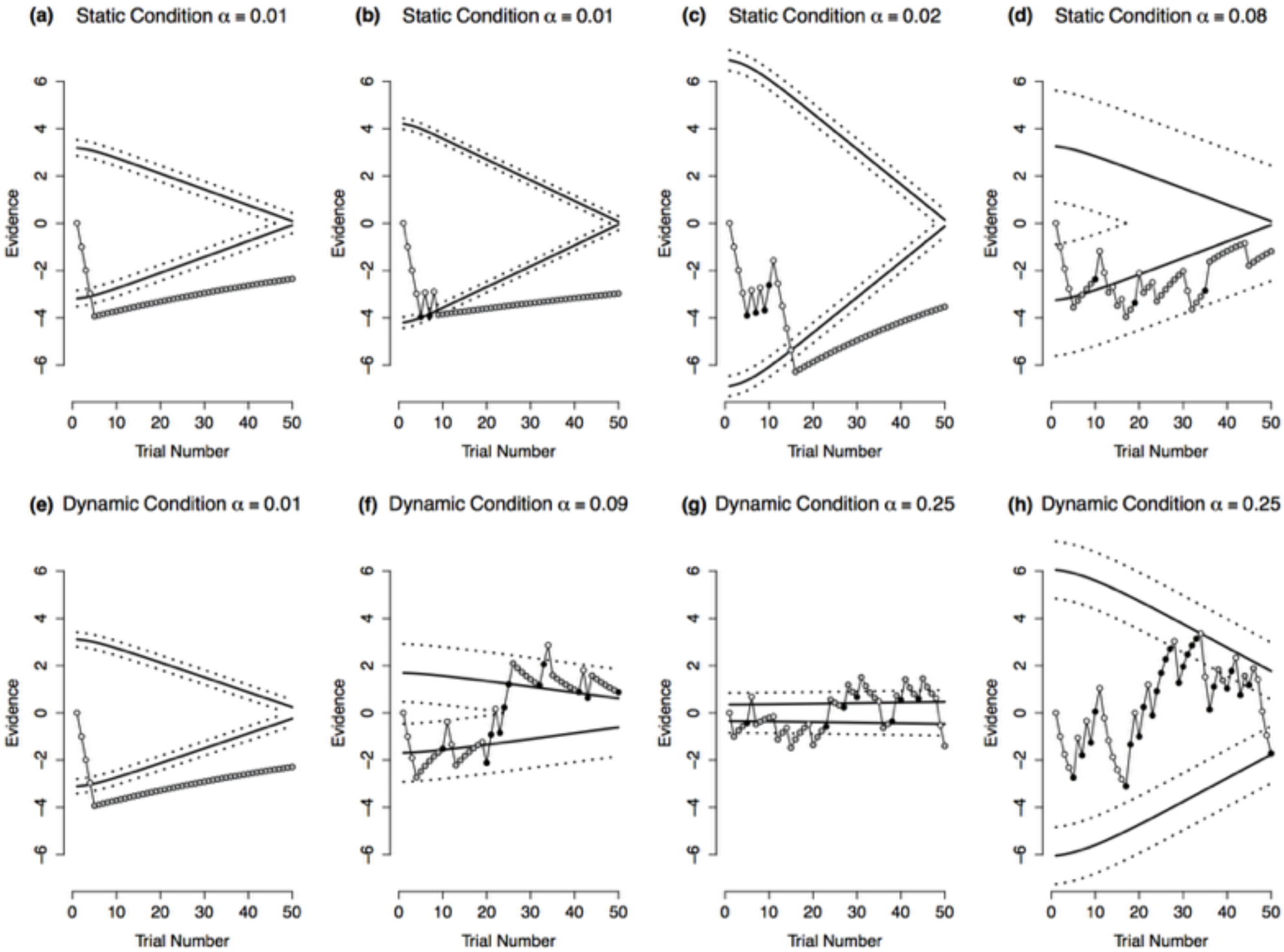
One subject doing a static task



Someone solving a dynamic problem

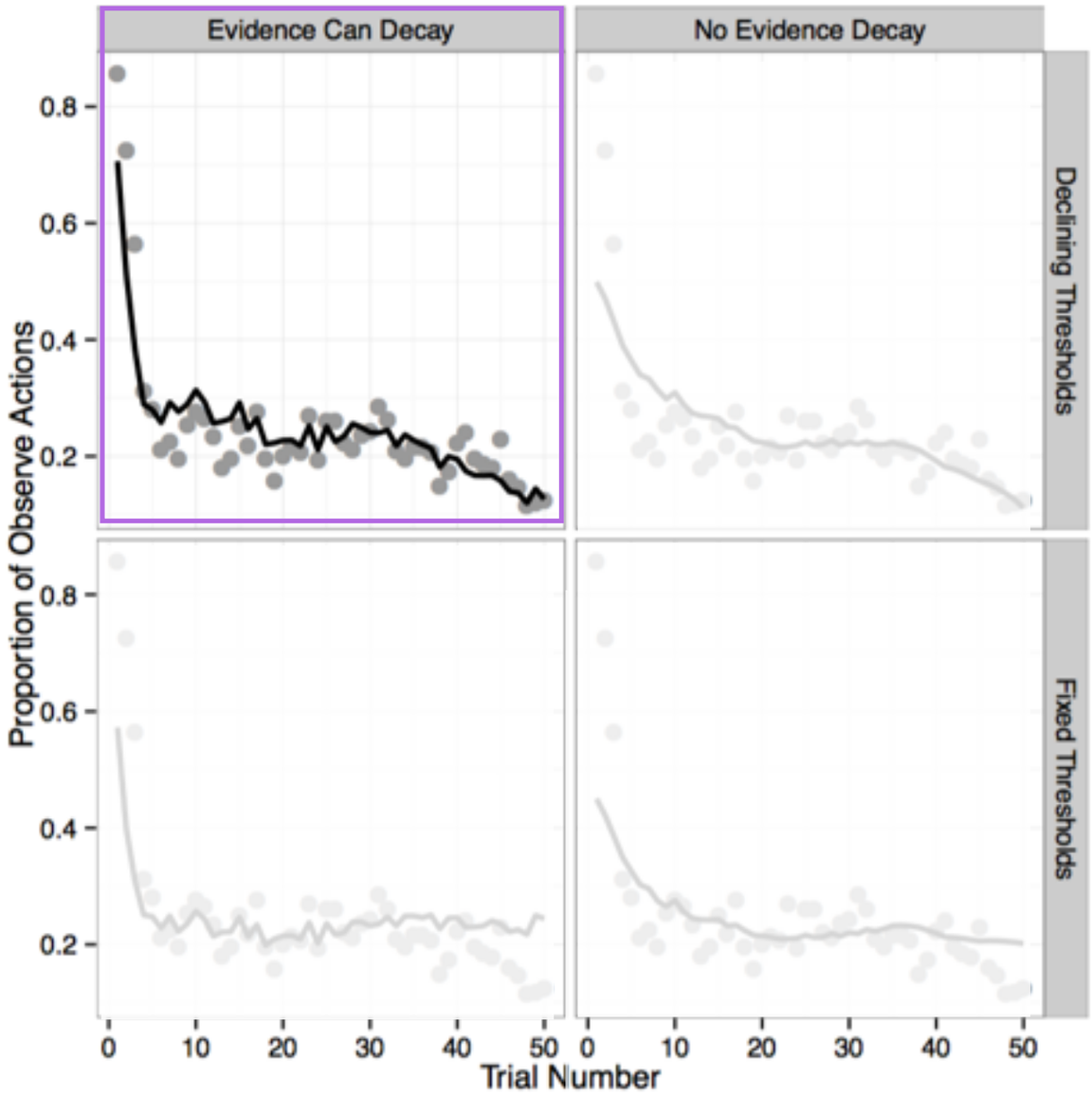


There's considerable variability…



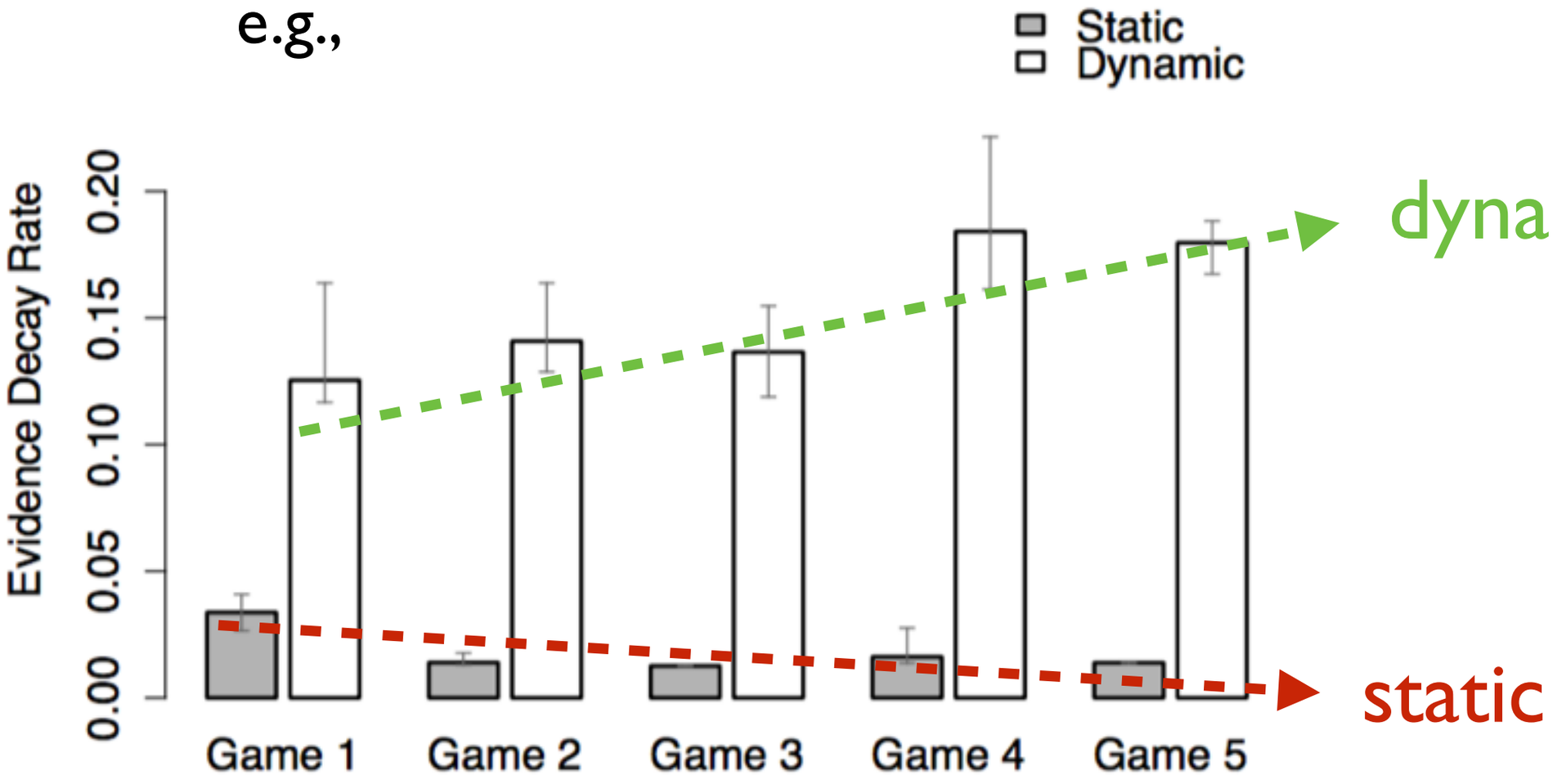
… but …

There are systematic patterns:
the policies have collapsing bounds (*finite horizon*) and evidence decay (*dynamic world*)



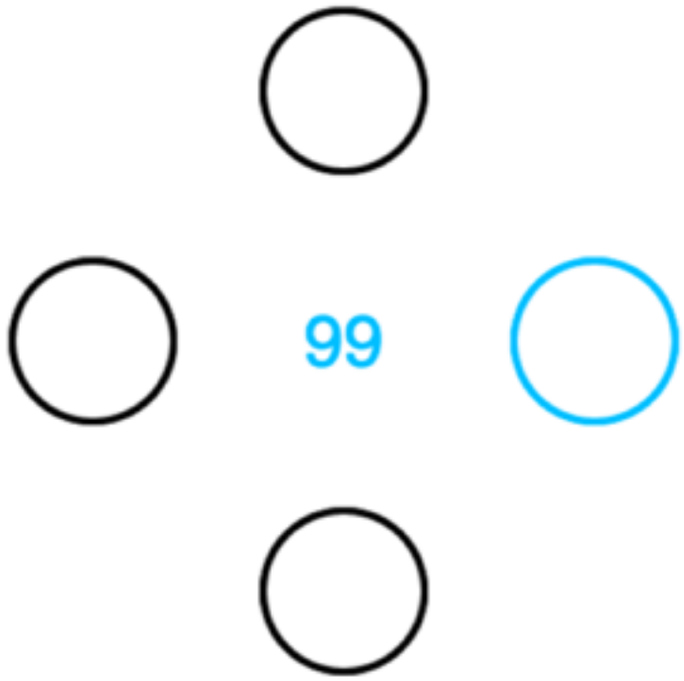
People *learn* the parameters of the task environment?

e.g.,



What to make of this?

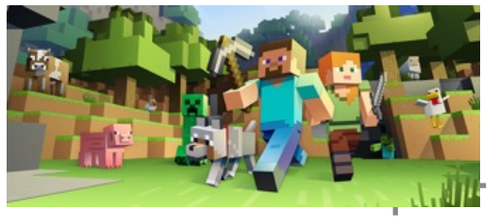
One shot structure learning?

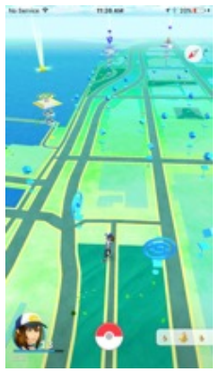


No idea what to do…
so use default strategy

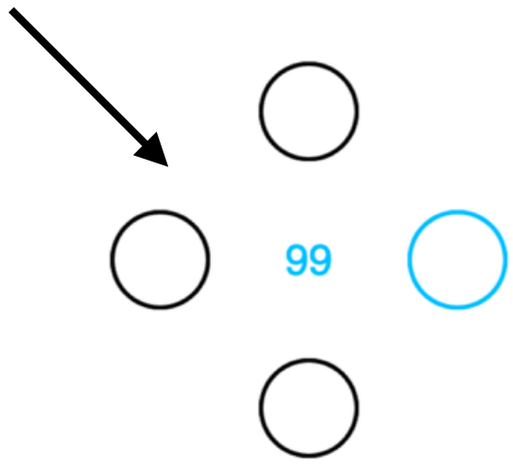
One shot structure learning?

irrelevant

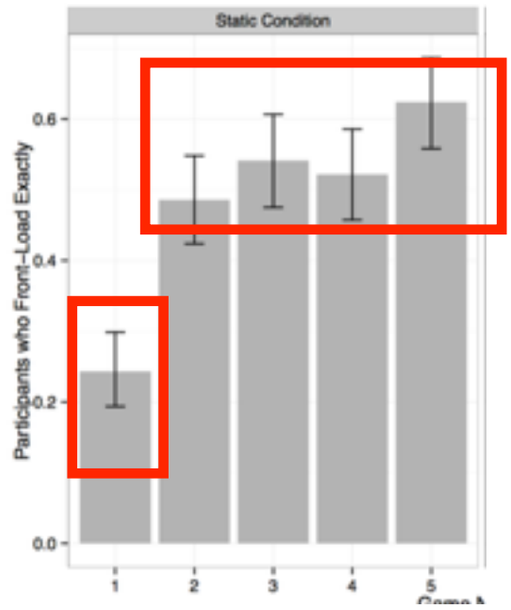




relevant



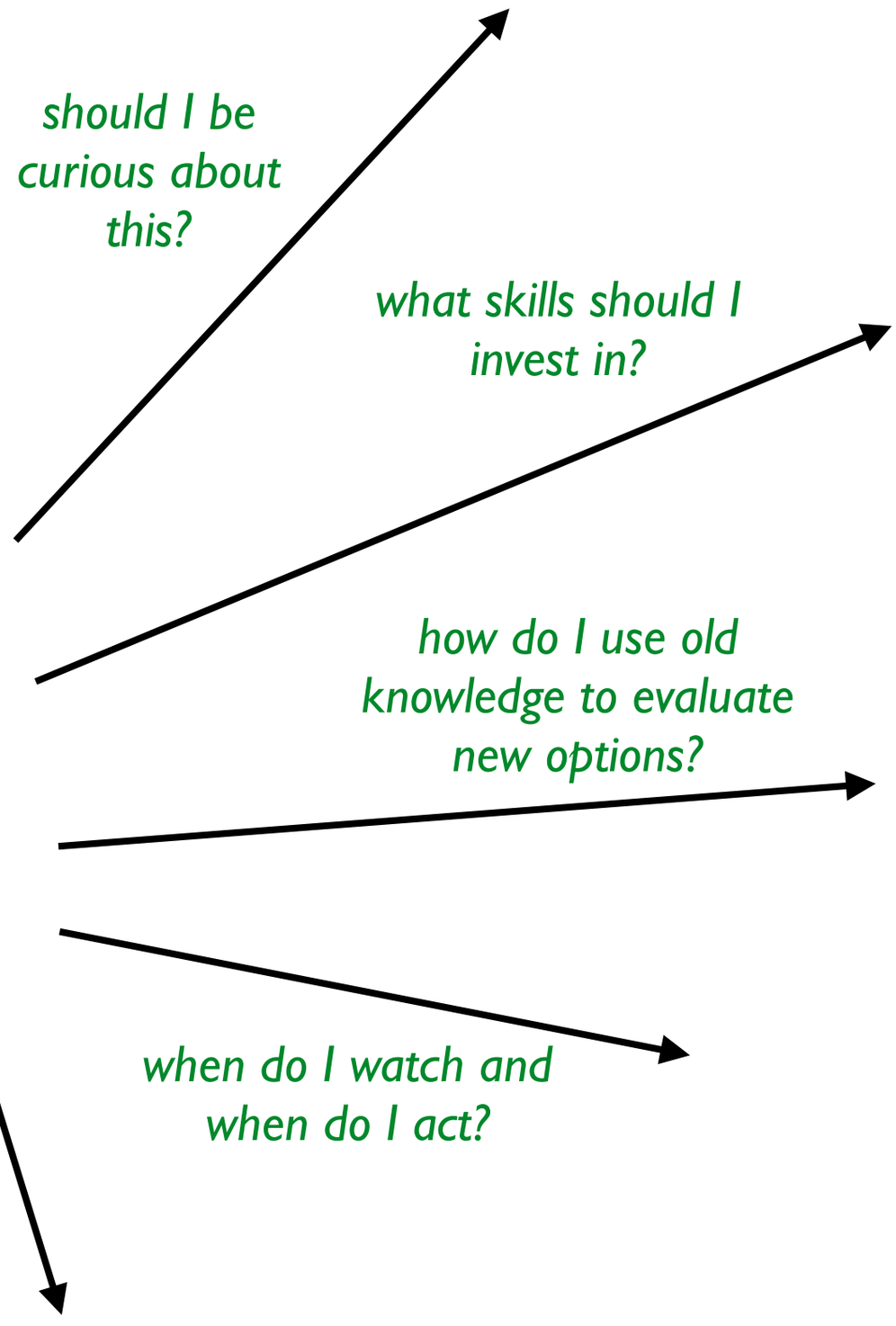
We need one play through to work out which “kind” of rule set is required?



What kinds of "task models"
do people use?

(Towards a richer class of
explore exploit dilemmas)

Everyday life motivates many different variants on sequential decision problems



Everyday life motivates many different variants on sequential decision problems

should I be curious about this?

Learning the value (or irrelevance) of novelty

what skills should I invest in?

Sequential decision making in a reactive environment

how do I use old knowledge to evaluate new options?

Contextual bandits - learning to use stimulus features to guide choices

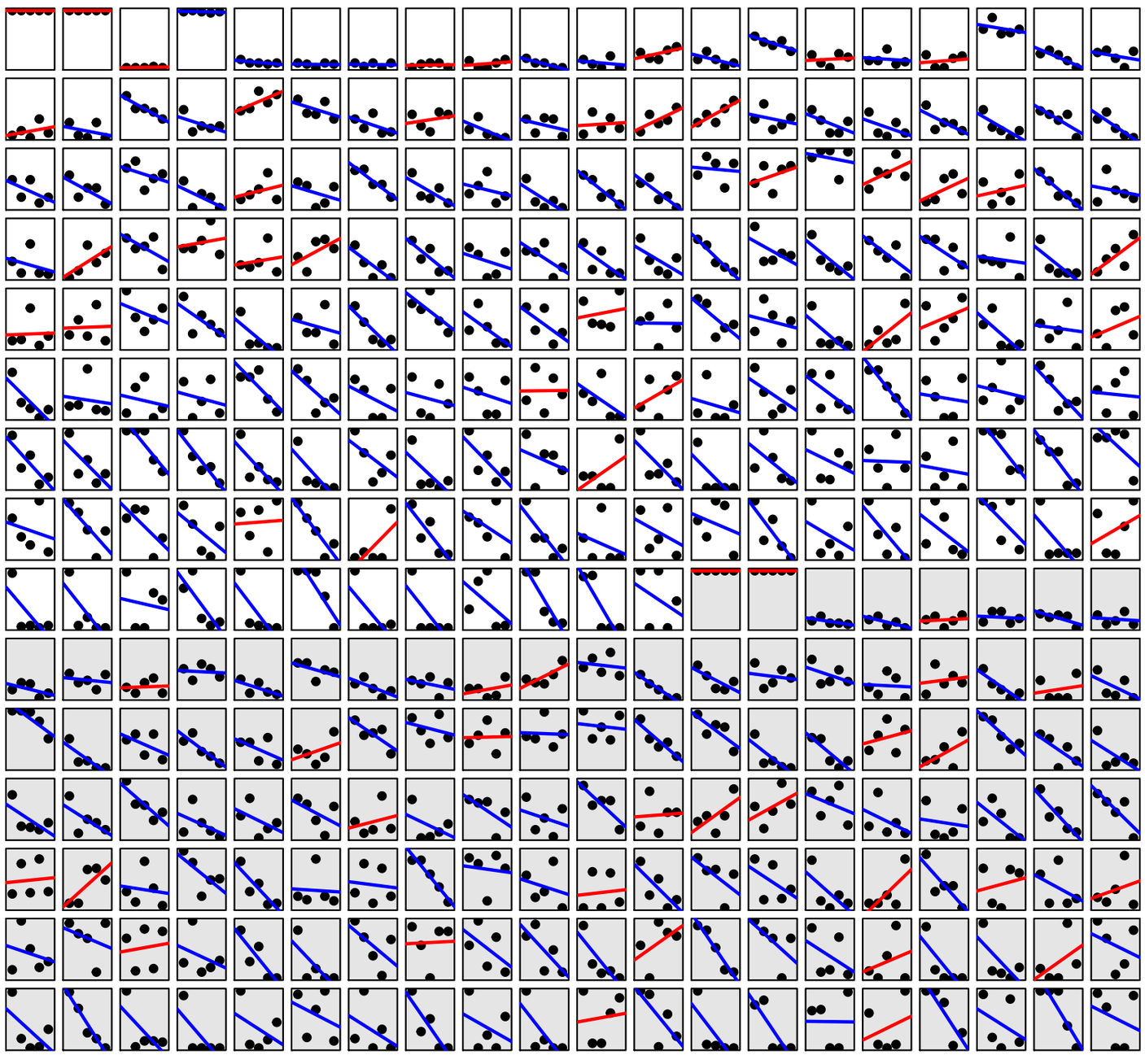
when should I hedge my bets?

Vanishing bandits: Keeping doors open in uncertain worlds

when do I watch and when do I act?

Peeks and keeps: Blending observe-or-bet problems with bandit problems

Each task seems* to show rapid strategy adaptation after a single short game



(* preliminary)

Which problem am I solving? Rule re-use across tasks supports rich transfer? Priors over environments?

	OB1	OB2	SB	VB1	VB2	PK	CoB	CuB
Changing rewards		■	■		■	■		■
Reactive environment			■	■	■			
Allows I/R separation	■	■				■		
Option turnover				■	■		■	
Predictive features							■	

. . .

⋮

Thanks!

Ben Newell

Christin Schulze

Sean Tauber

Dan Bennett

Nathaniel Phillips

Michael Lee

Amy Perfors

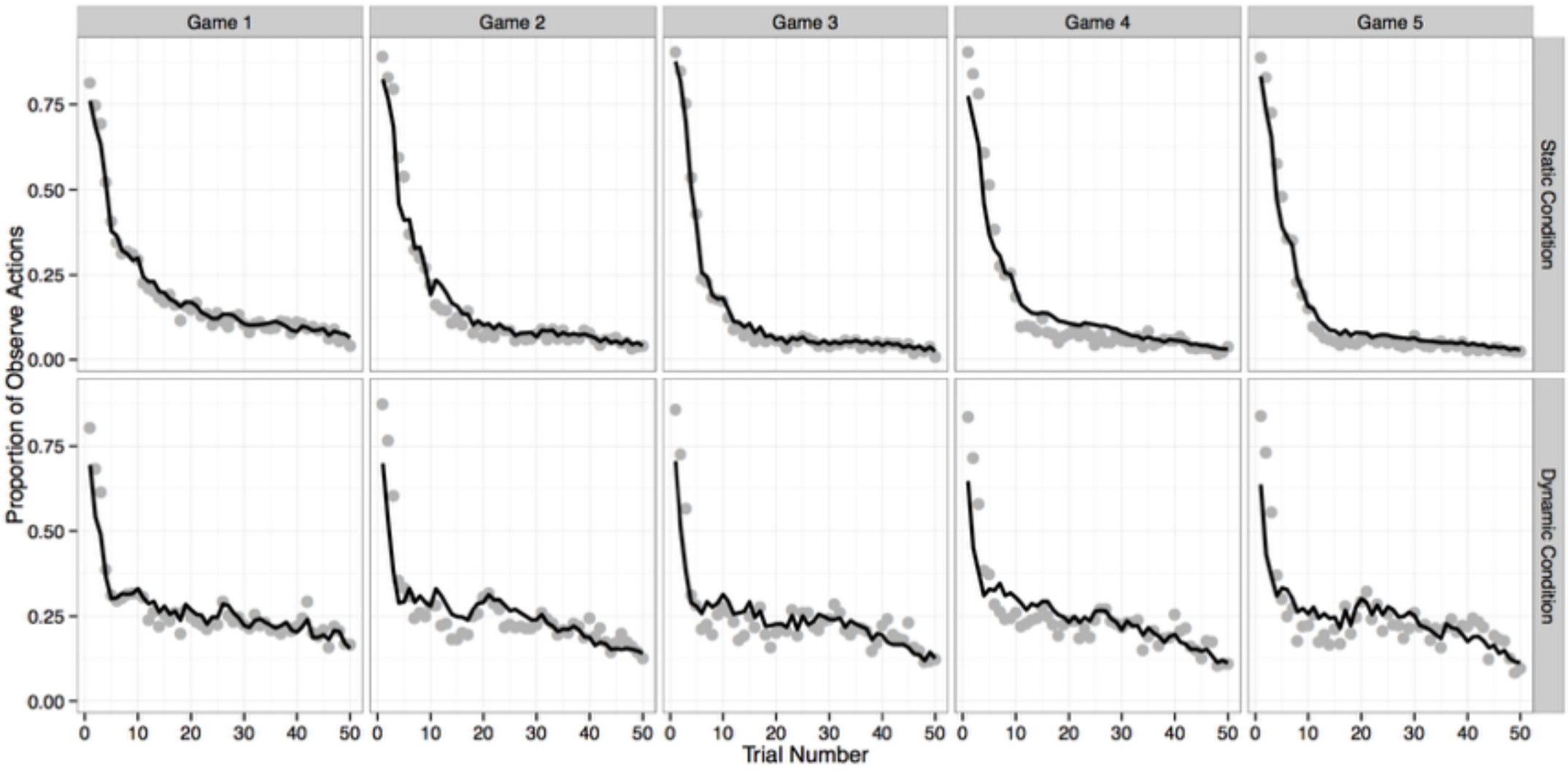
Keith Ransom

Wouter Voorspoels

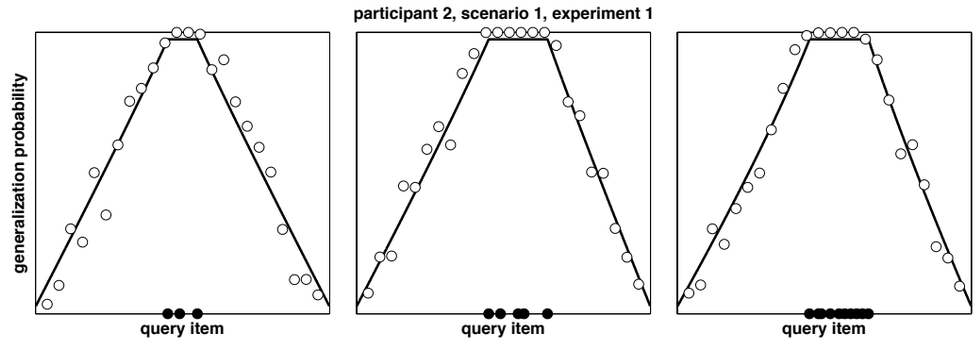
Drew Hendrickson

Australian Government
Australian Research Council

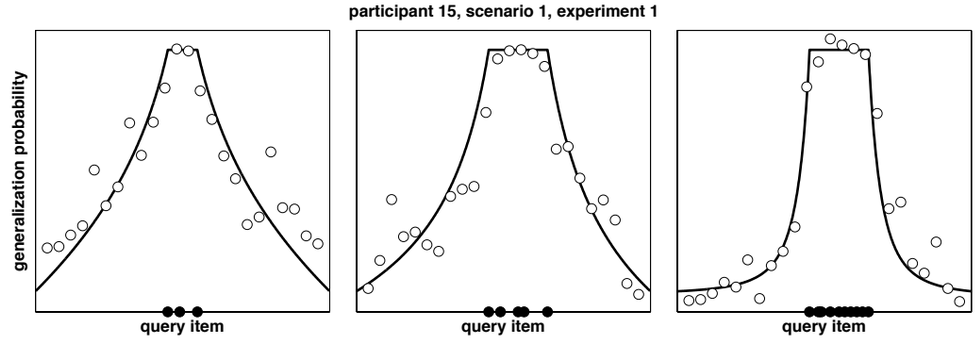
(Quick sanity check - model fits)

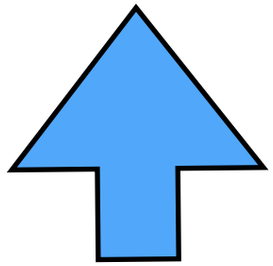


Randomly generated data licence weak inferences →



Helpfully generated data licence strong inferences →





Effect of sample size in simple generalisation depends on sampling assumption

Learner's theory of the data generating mechanism induces qualitative shifts in reasoning

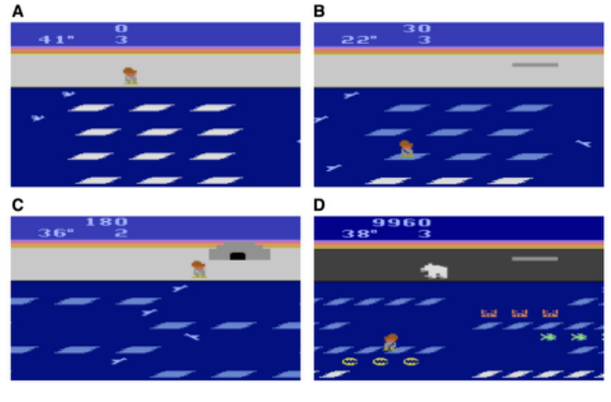
Navarro, Dry & Lee (2012). *Cognitive Science*

Back to the puzzle…

Human RL needs to infer *which model* to apply in which context, solve problems with *large state spaces*, using *limited computational resources* and with *minimal training* data. How is this done?

Answer? Flexible *re-use* of old knowledge?

- Get closest to 100, or 300, or 1000, or 3000, or any level, without going over.
- Beat your friend, who's playing next to you, but just barely, not by too much, so as not to embarrass them.
- Go as long as you can without dying.
- Die as quickly as you can.
- Pass each level at the last possible minute, right before the temperature timer hits zero and you die (i.e., come as close as you can to dying from frostbite without actually dying).
- Get to the furthest unexplored level without regard for your score.
- See if you can discover secret Easter eggs.
- Get as many fish as you can.
- Touch all the individual ice floes on screen once and only once.
- Teach your friend how to play as efficiently as possible.



Lake, Ullman, Tenenbaum & Gershman (in press). *BBS*

How do people acquire new knowledge?

(categorisation & reasoning)

How do we make choices in an uncertain world?

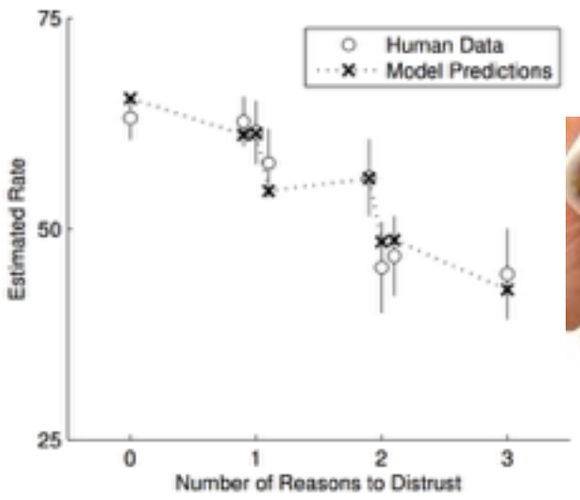
(judgment & decision making)

Sequential decision problems in an *uncertain* environment: people need to *learn* a model of the world and then work out how best to make *use* of it!

How do we make choices in an uncertain world?

(judgment & decision making)

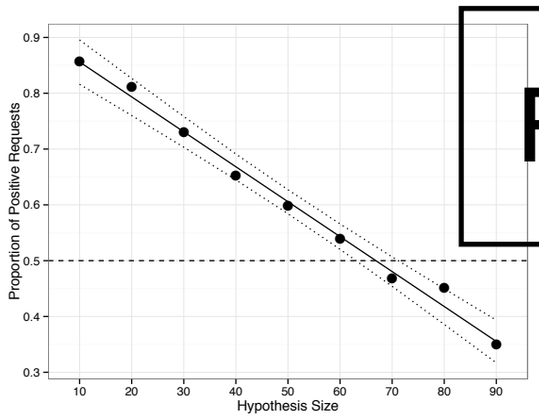
How do people evaluate the quality of evidence?



Welsh & Navarro (2012). *Org. Behavior & Human Dec. Making*

How do we make choices in an uncertain world?

(judgment & decision making)



P D 2 4

In what way do the statistics of the decision problem matter?

Navarro & Perfors (2011). *Psych Review*
Hendrickson, Perfors & Navarro (2016) *Decision*