

Reverse Engineering Mind and Society

Xuechunzi Bai

Assistant Professor of Psychology

University of Chicago

Reverse Engineering **Mind and Society**



tufas



tufas





Squeaky



Squeaky



Squeaky



Squeaky



Squeaky



Squeaky



Squeaky



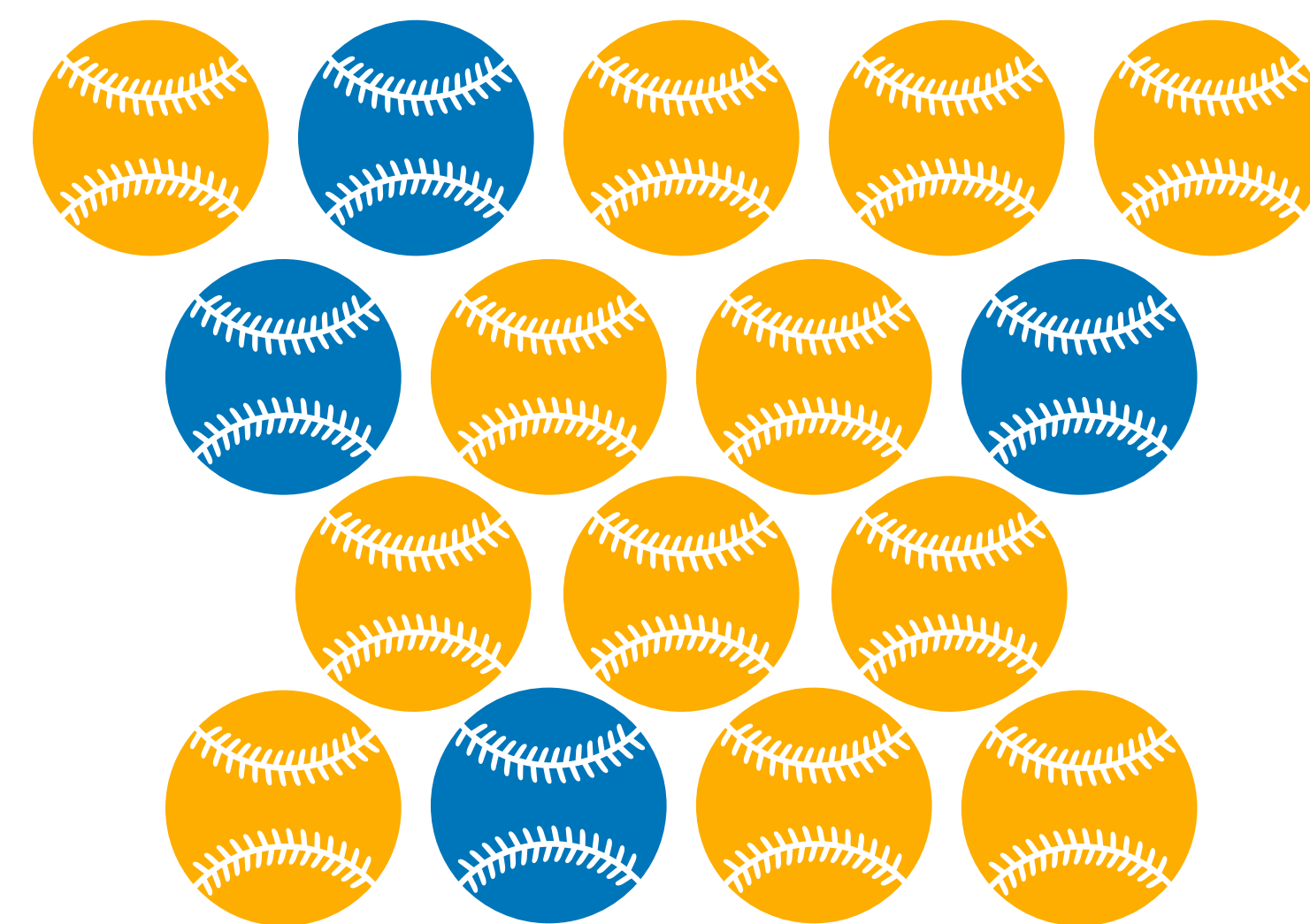
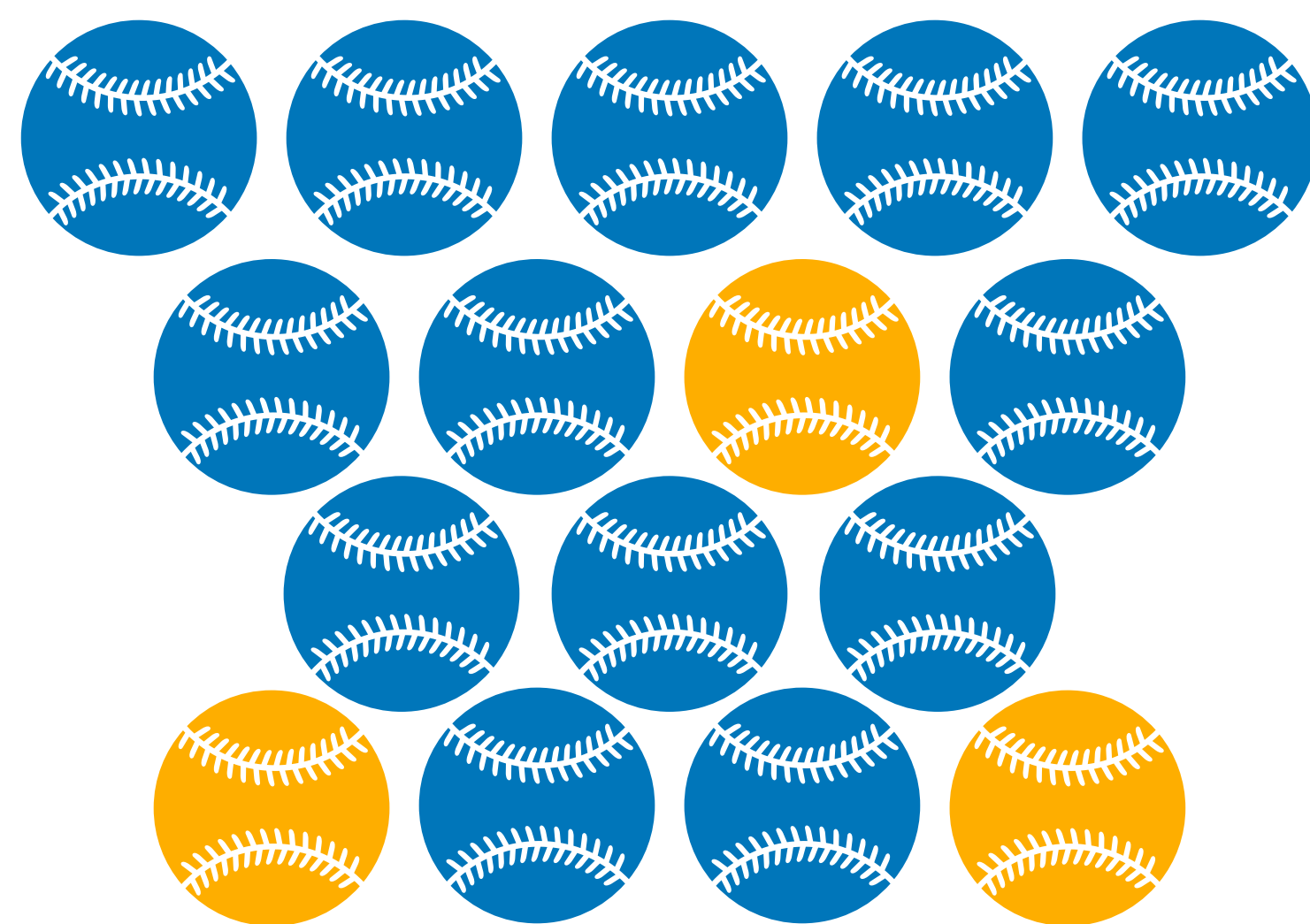
Squeaky

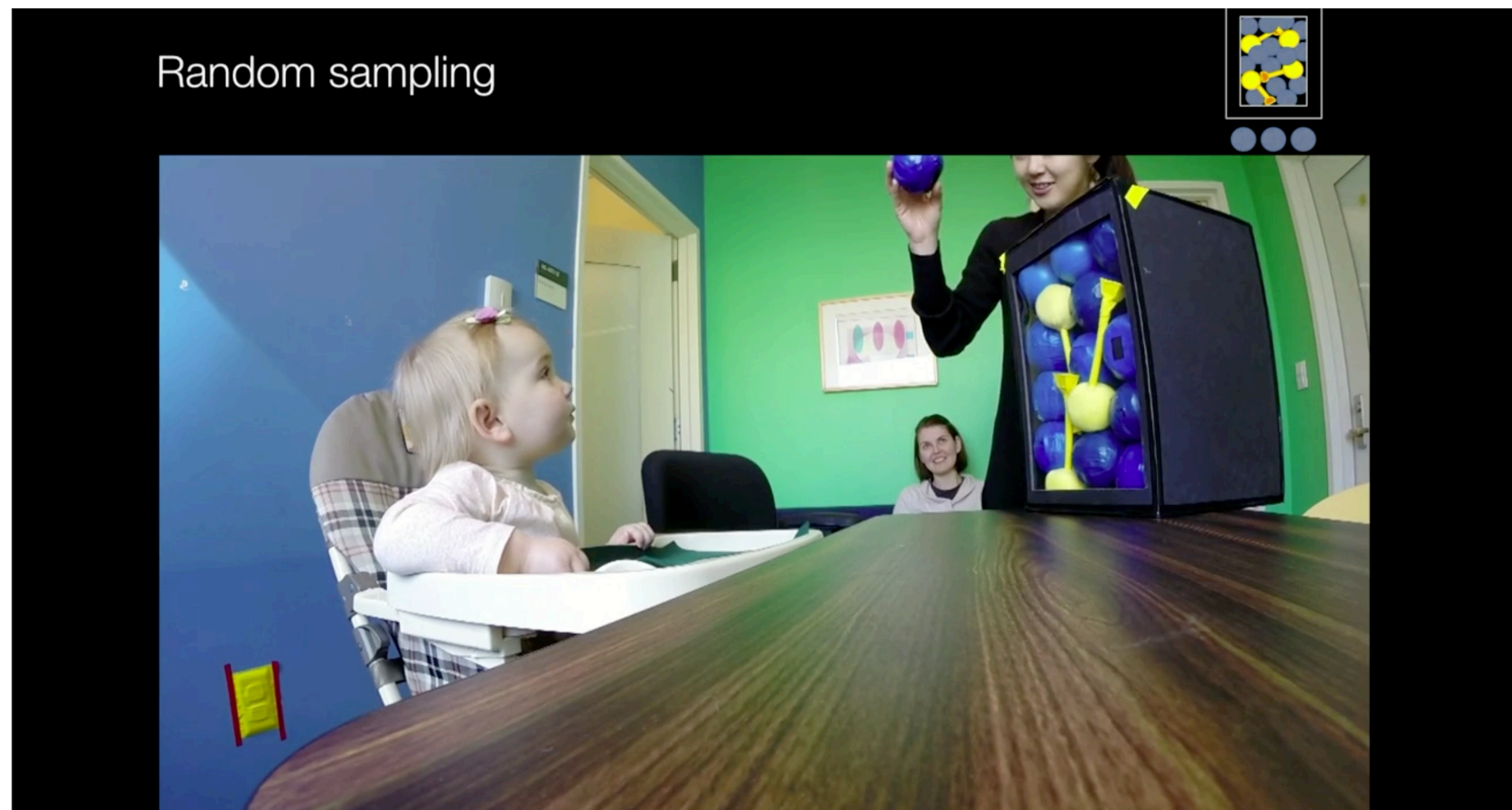
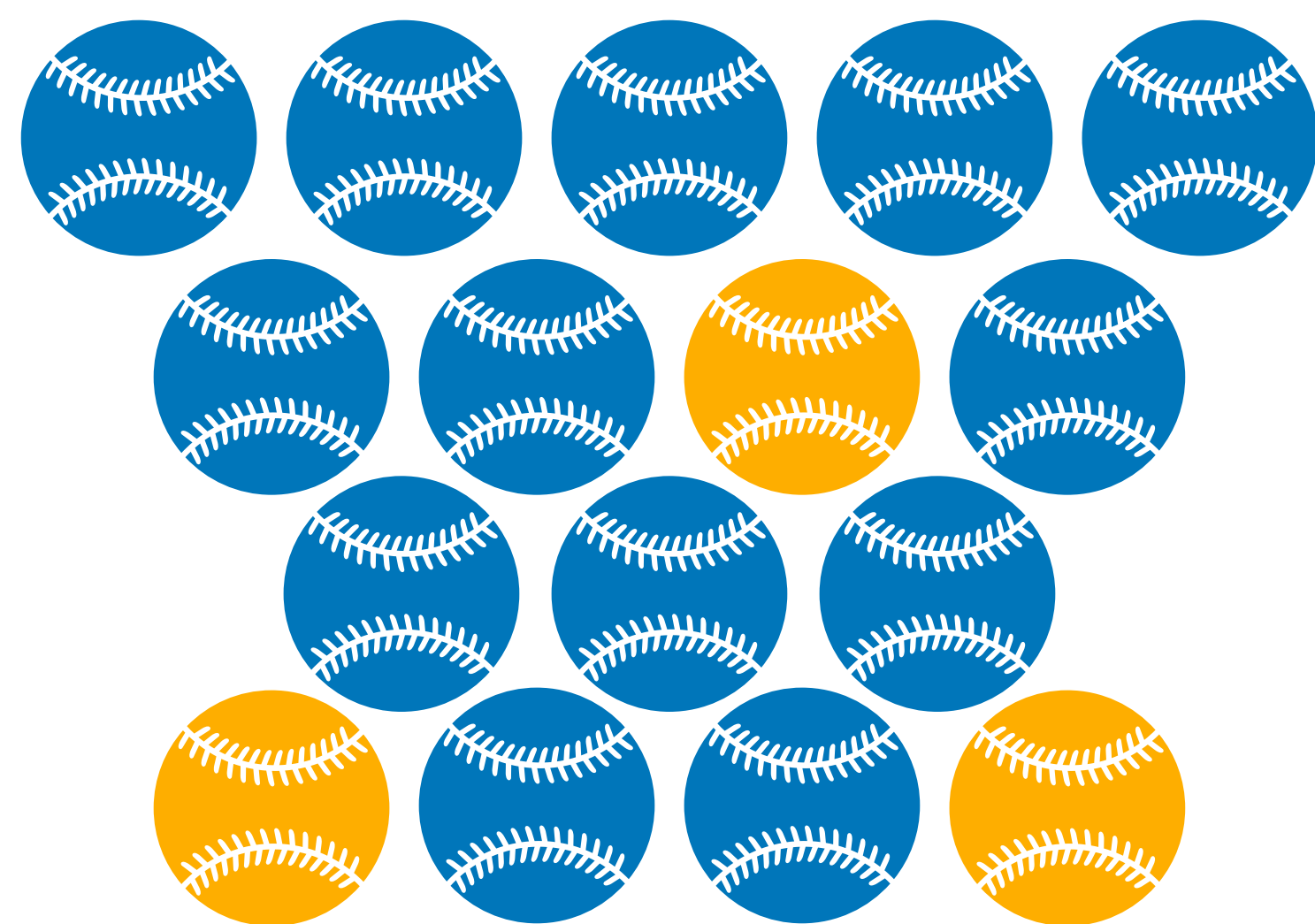


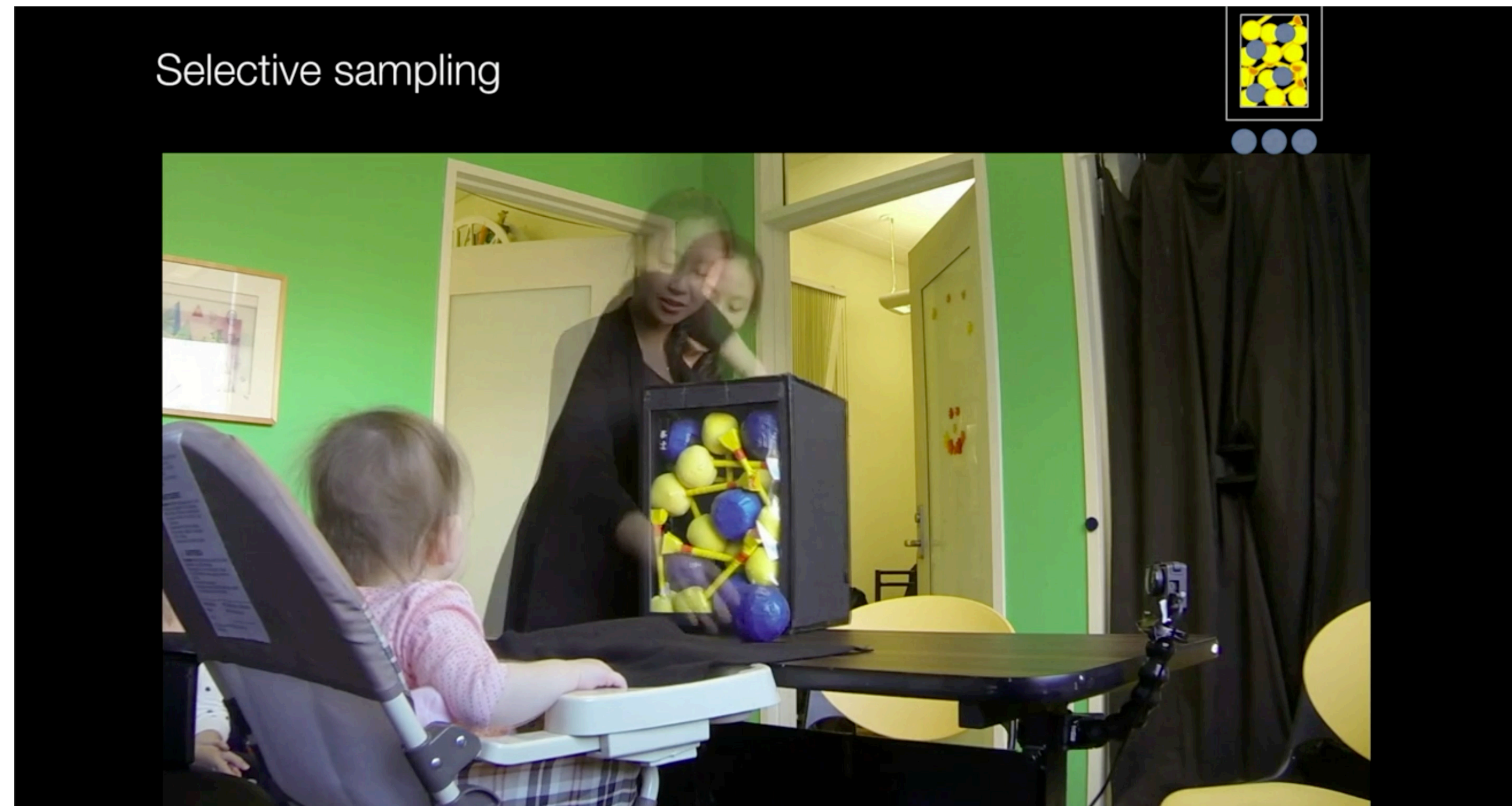
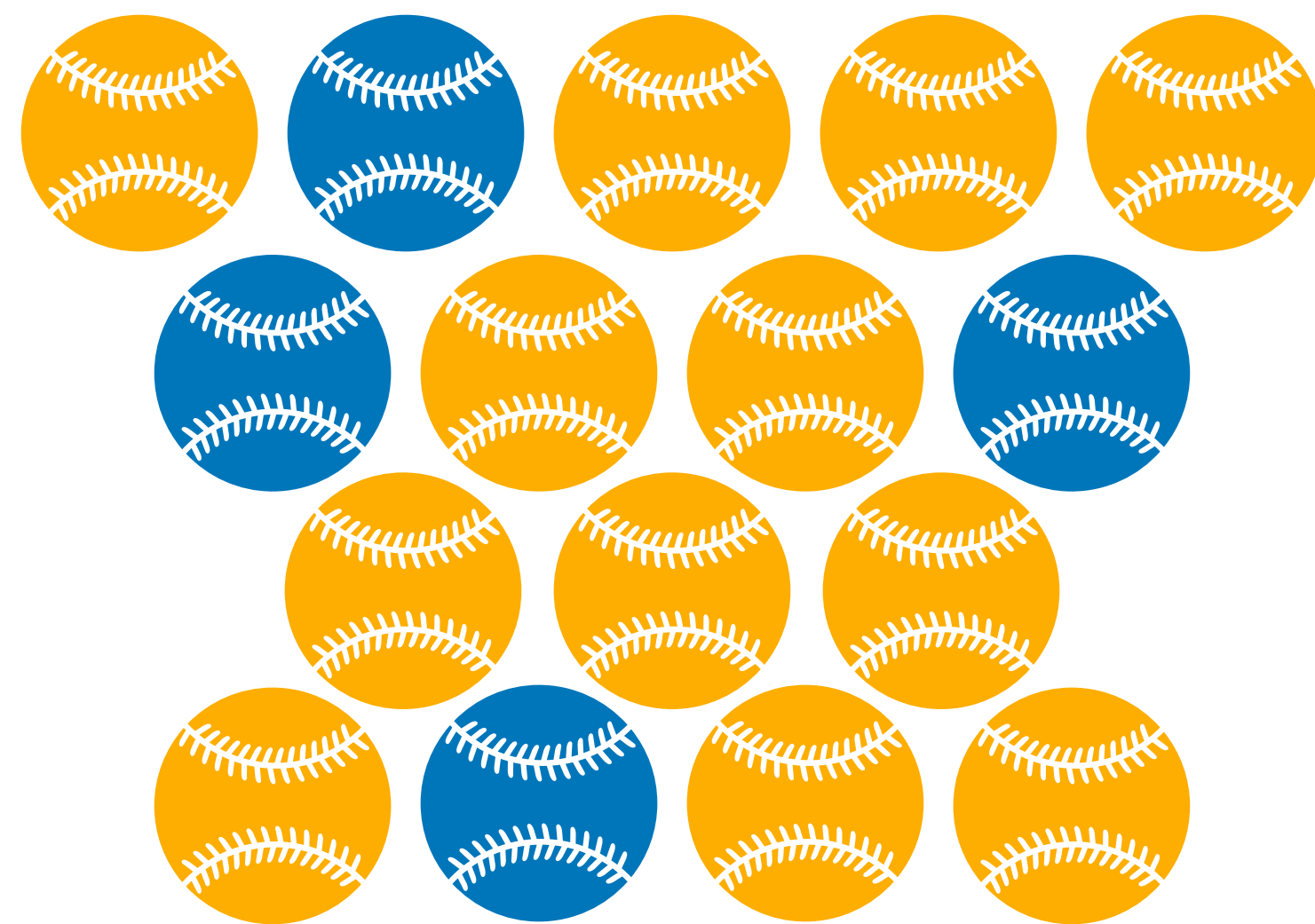
Squeaky



Does this ball squeak?



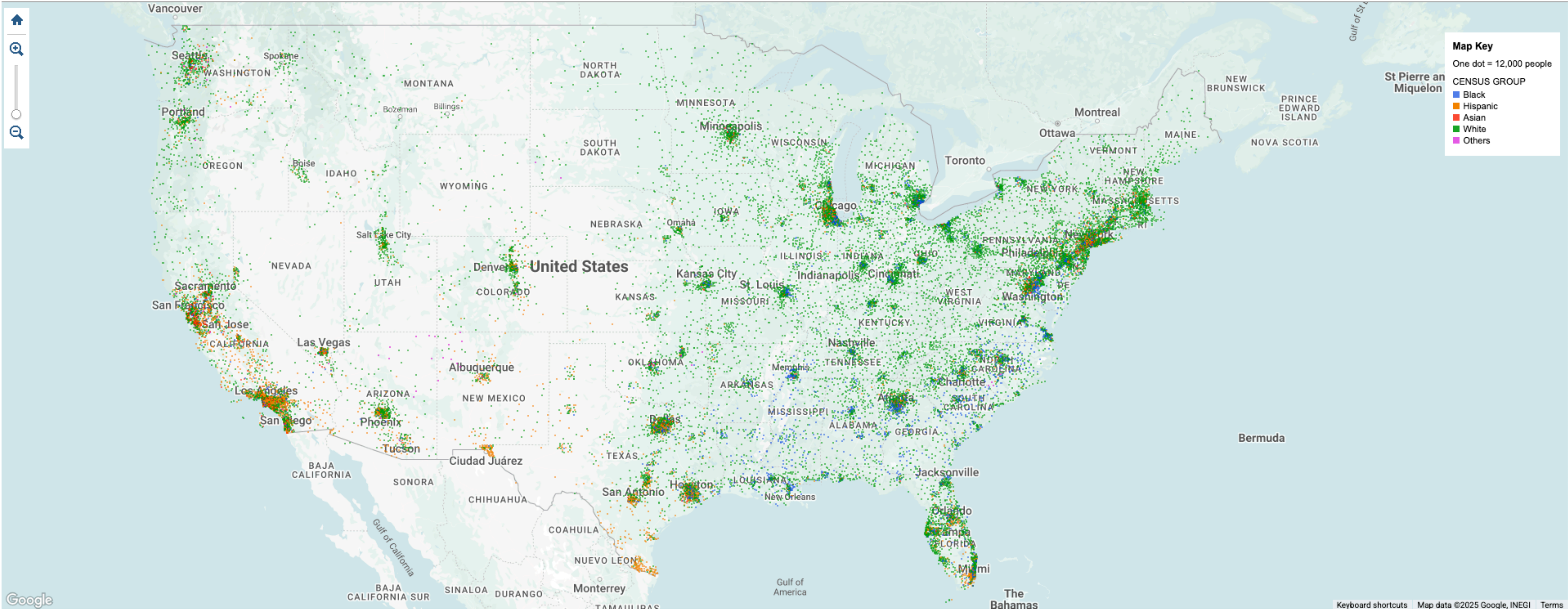




Reverse Engineering Mind and Society

Mapping Segregation

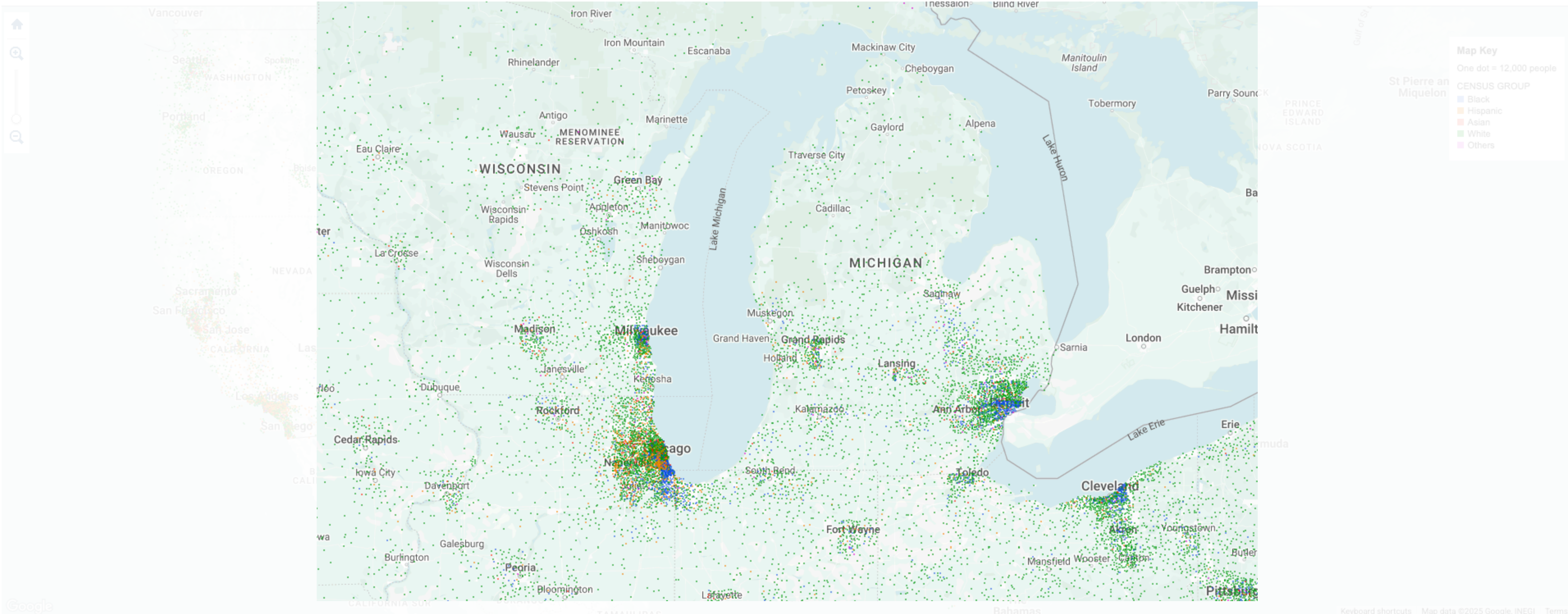
New government rules will require all cities and towns receiving federal housing funds to assess patterns of segregation.



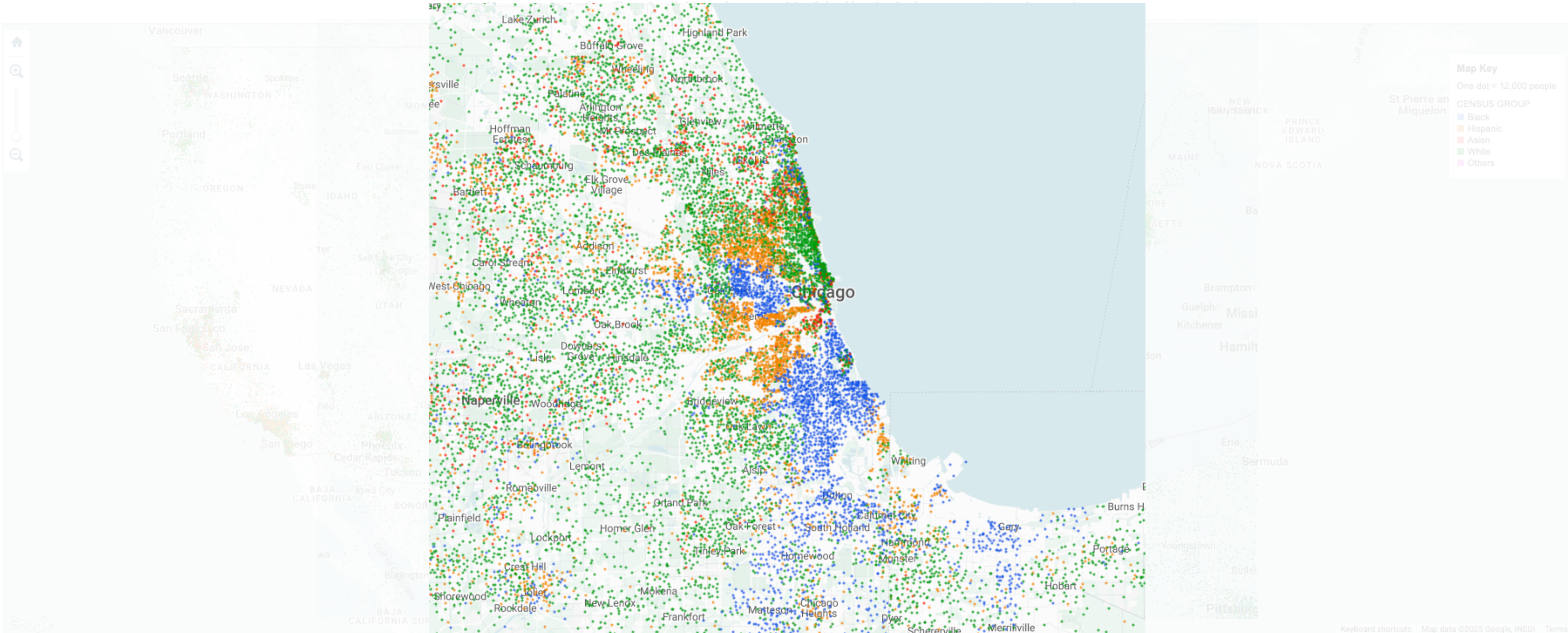
Mapping Segregation, 2015, *The New York Times*

Mapping Segregation

New government rules will require all cities and towns receiving federal housing funds to assess patterns of segregation.



Mapping Segregation



Reverse Engineering Mind and Society



Richard Feynman, 1918 - 1988

“What I cannot create, I do not understand.”

A Brief History

1913



John Watson
1878 - 1958



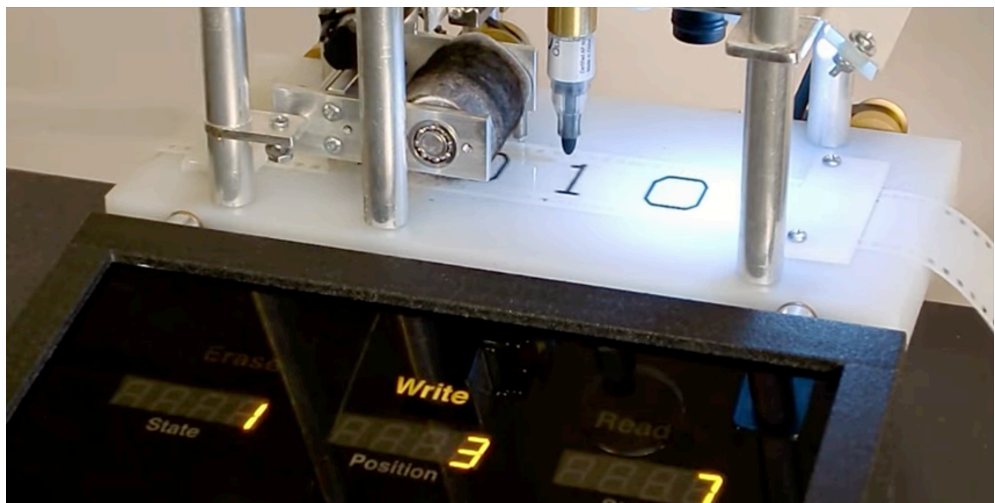
B.F. Skinner
1904 - 1990

Psychology as the **behaviorist** views it is a purely **objective** natural science. Its theoretical goal is the prediction and control of behavior. **Introspection forms no essential part of its method nor is the scientific value** of its data dependent upon...

A Brief History



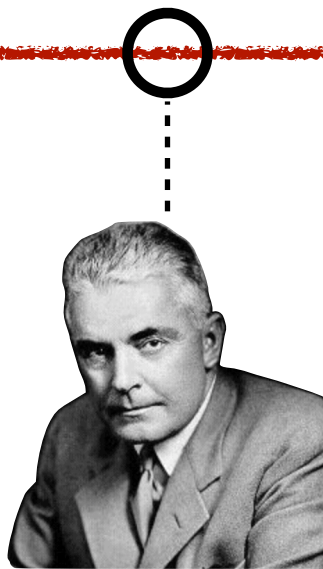
Alan Turing
1912 - 1954



Turing Machine

1936

1913



John Watson
1878 - 1958

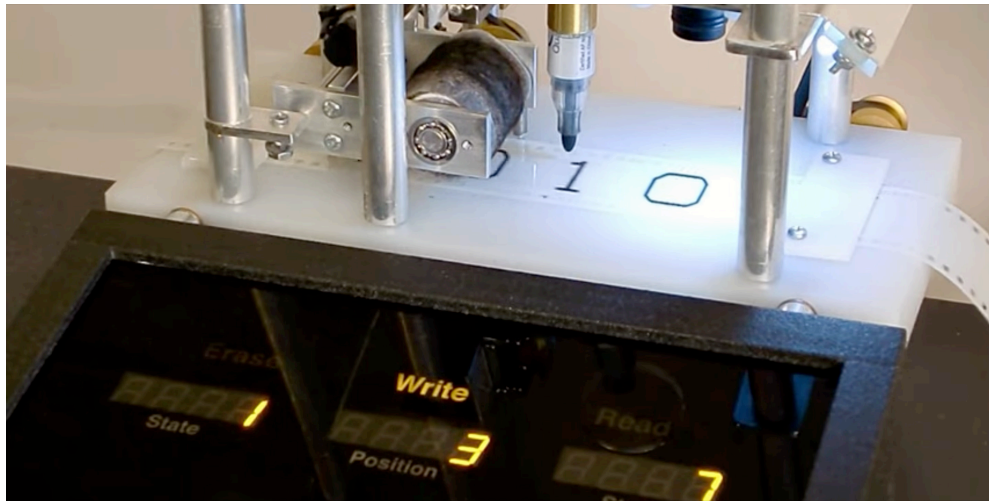


B.F. Skinner
1904 - 1990

A Brief History



Alan Turing
1912 - 1954



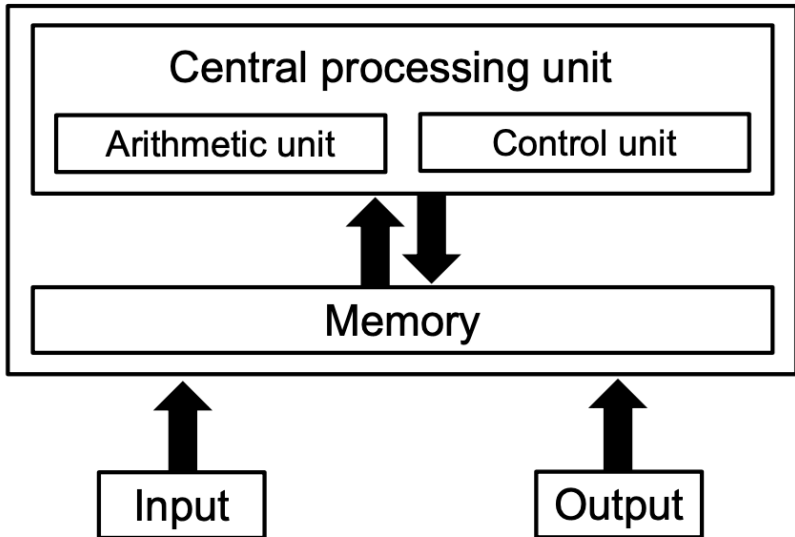
Turing Machine

1936

1945



John von Neumann
1903 - 1957



The von Neumann Architecture

1913



John Watson
1878 - 1958

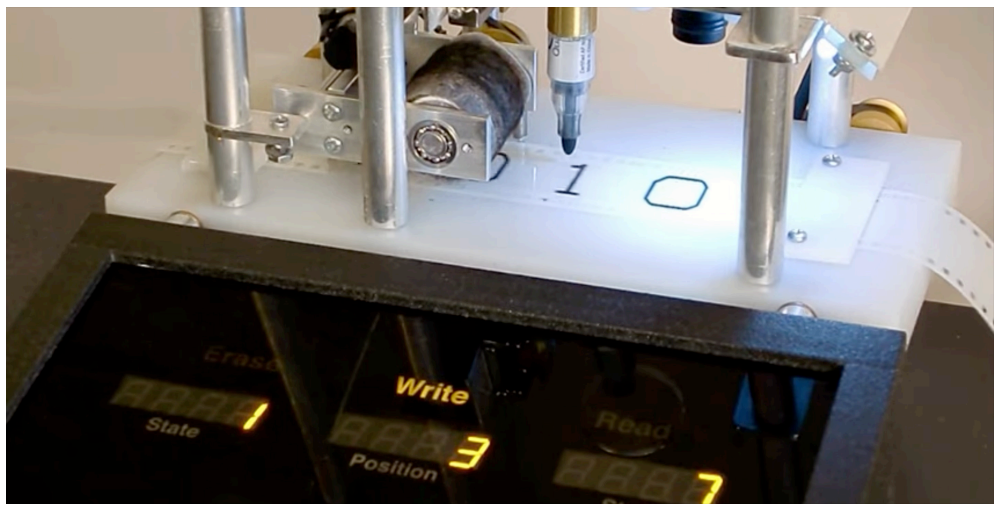


B.F. Skinner
1904 - 1990

A Brief History



Alan Turing
1912 - 1954



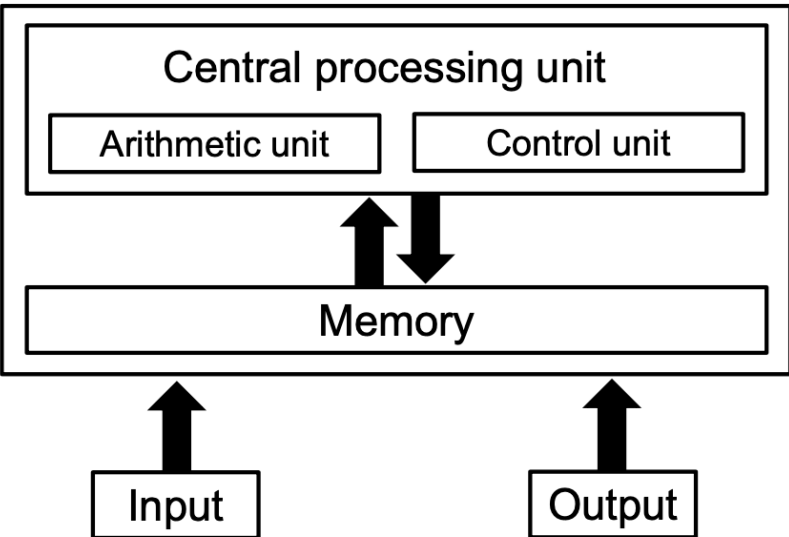
Turing Machine

1936

1945



John von Neumann
1903 - 1957



The von Neumann Architecture

1913



John Watson
1878 - 1958



B.F. Skinner
1904 - 1990

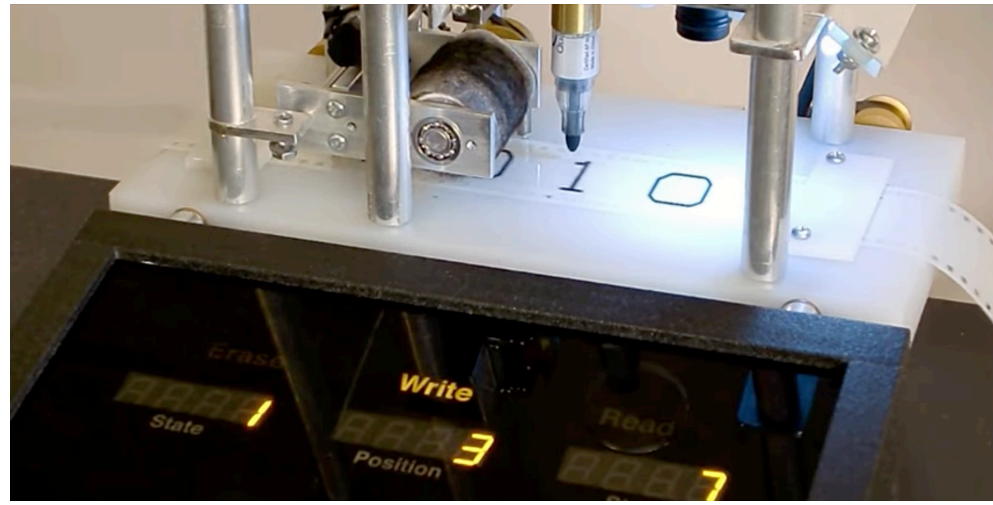
1956

Symposium on Information Theory @ MIT
(Birth of Cognitive Science)

A Brief History



Alan Turing
1912 - 1954



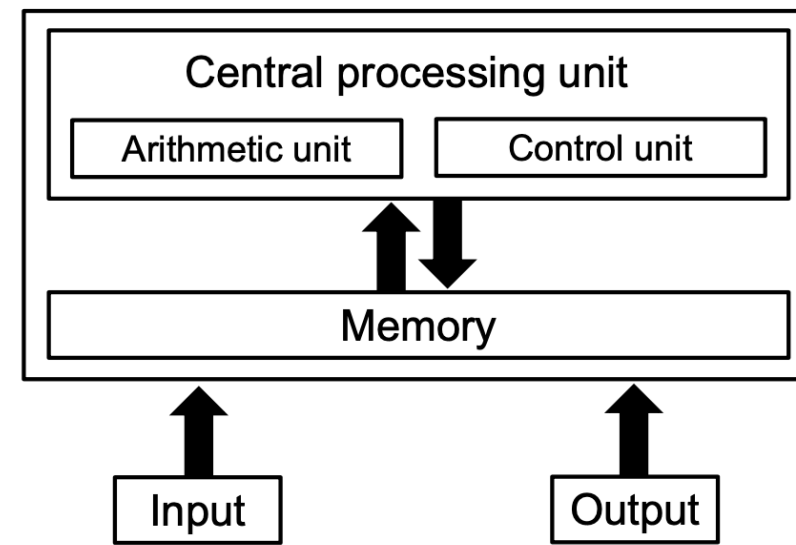
Turing Machine

1936

1945

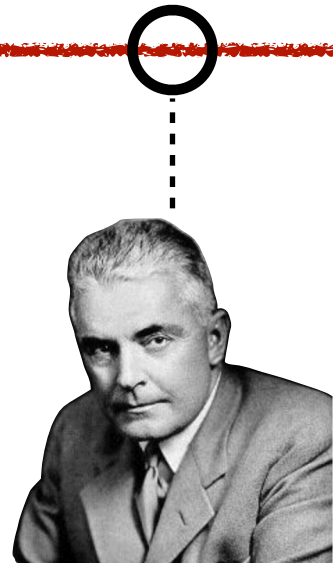


John von Neumann
1903 - 1957

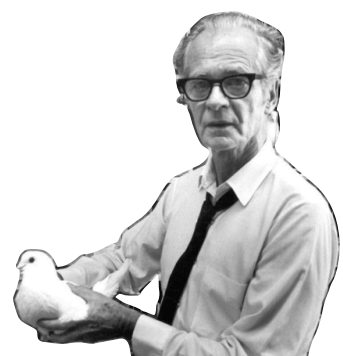


The von Neumann Architecture

1913



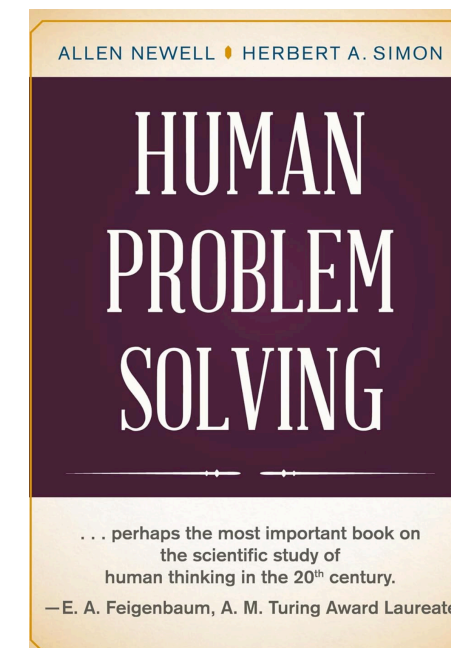
John Watson
1878 - 1958



B.F. Skinner
1904 - 1990

1956

Symposium on Information Theory @ MIT
(Birth of Cognitive Science)



1972



Herb Simon (1916 - 2001)
Allen Newell (1927 - 1992)

A Brief History

1956



Symposium on Information Theory @ MIT
(Birth of Cognitive Science)

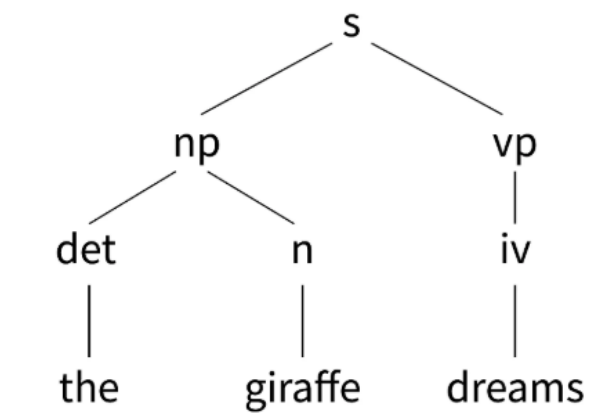
1972



Herb Simon (1916 - 2001)
Allen Newell (1927 - 1992)

A Brief History

Symbolic



Universal Grammar



Noam Chomsky

1956

1957

1972

Symposium on Information Theory @ MIT
(Birth of Cognitive Science)

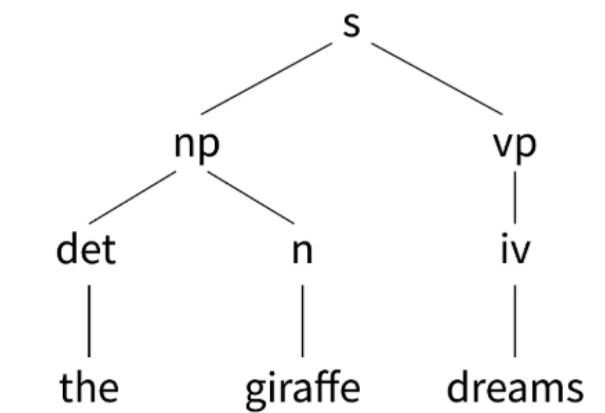


Herb Simon (1916 - 2001)
Allen Newell (1927 - 1992)

A Brief History

Symbolic

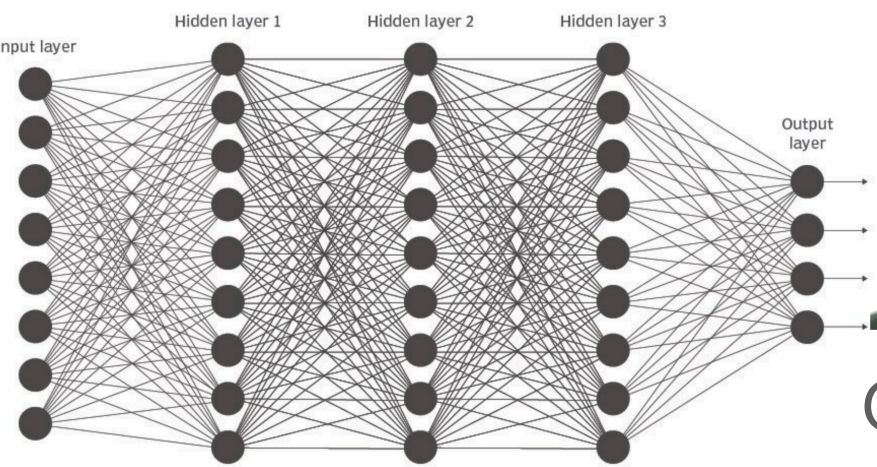
Connectionist



Universal Grammar



Noam Chomsky



Neural Network



Geoffrey Hinton

1956

1957

1972

1986

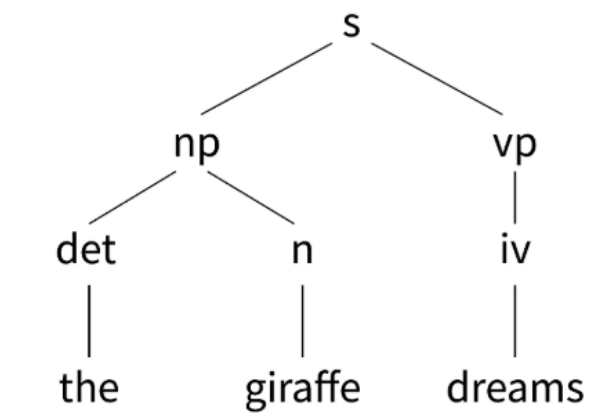
Symposium on Information Theory @ MIT
(Birth of Cognitive Science)



Herb Simon (1916 - 2001)
Allen Newell (1927 - 1992)

A Brief History

Symbolic

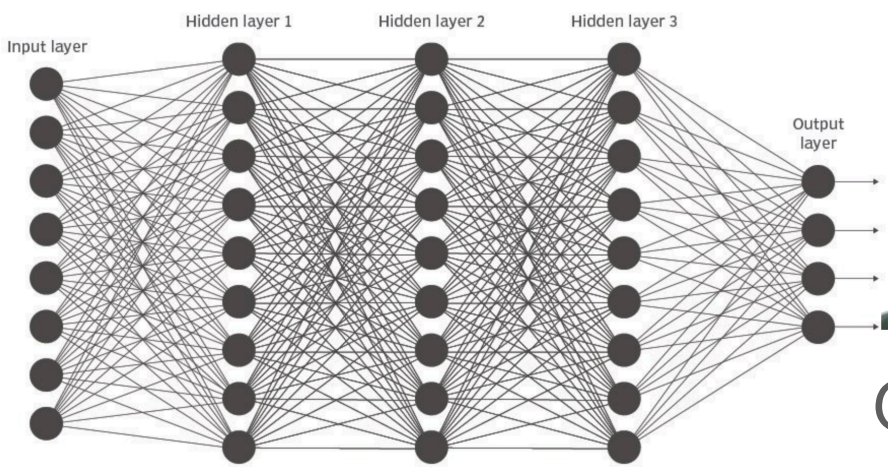


Universal Grammar



Noam Chomsky

Connectionist



Neural Network



Geoffrey Hinton

Bayesian

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Josh Tenenbaum et al.

1956

1957

1972

1986

2001

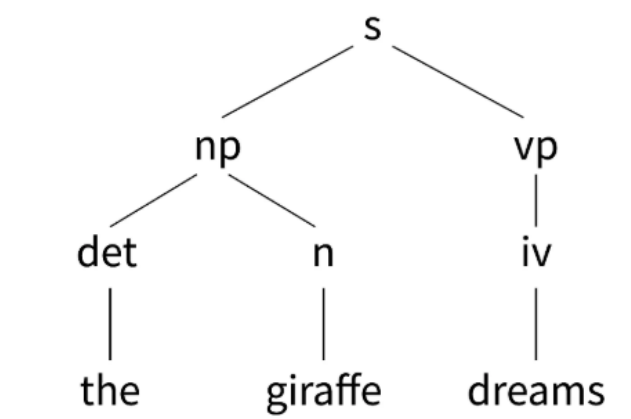
Symposium on Information Theory @ MIT
(Birth of Cognitive Science)



Herb Simon (1916 - 2001)
Allen Newell (1927 - 1992)

A Brief History

Symbolic



Universal Grammar



Noam Chomsky

1956

1957

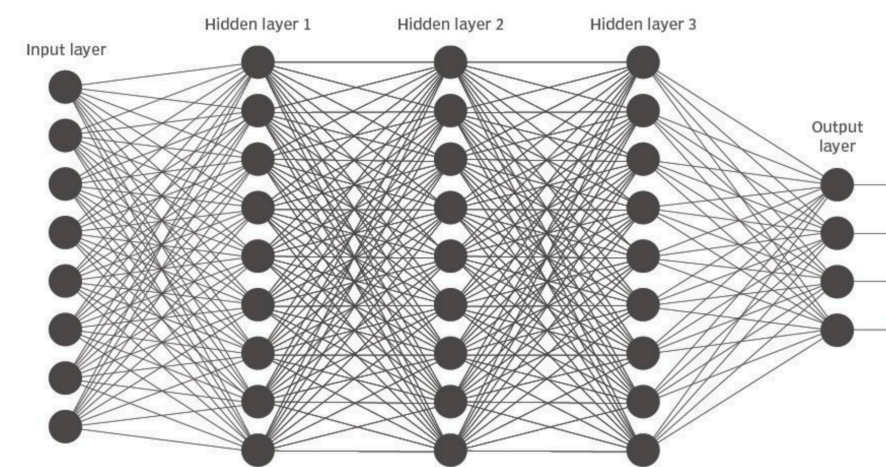
1972

1986

2001

2025

Connectionist



Neural Network



Geoffrey Hinton

Bayesian

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Josh Tenenbaum et al.

LLM...?



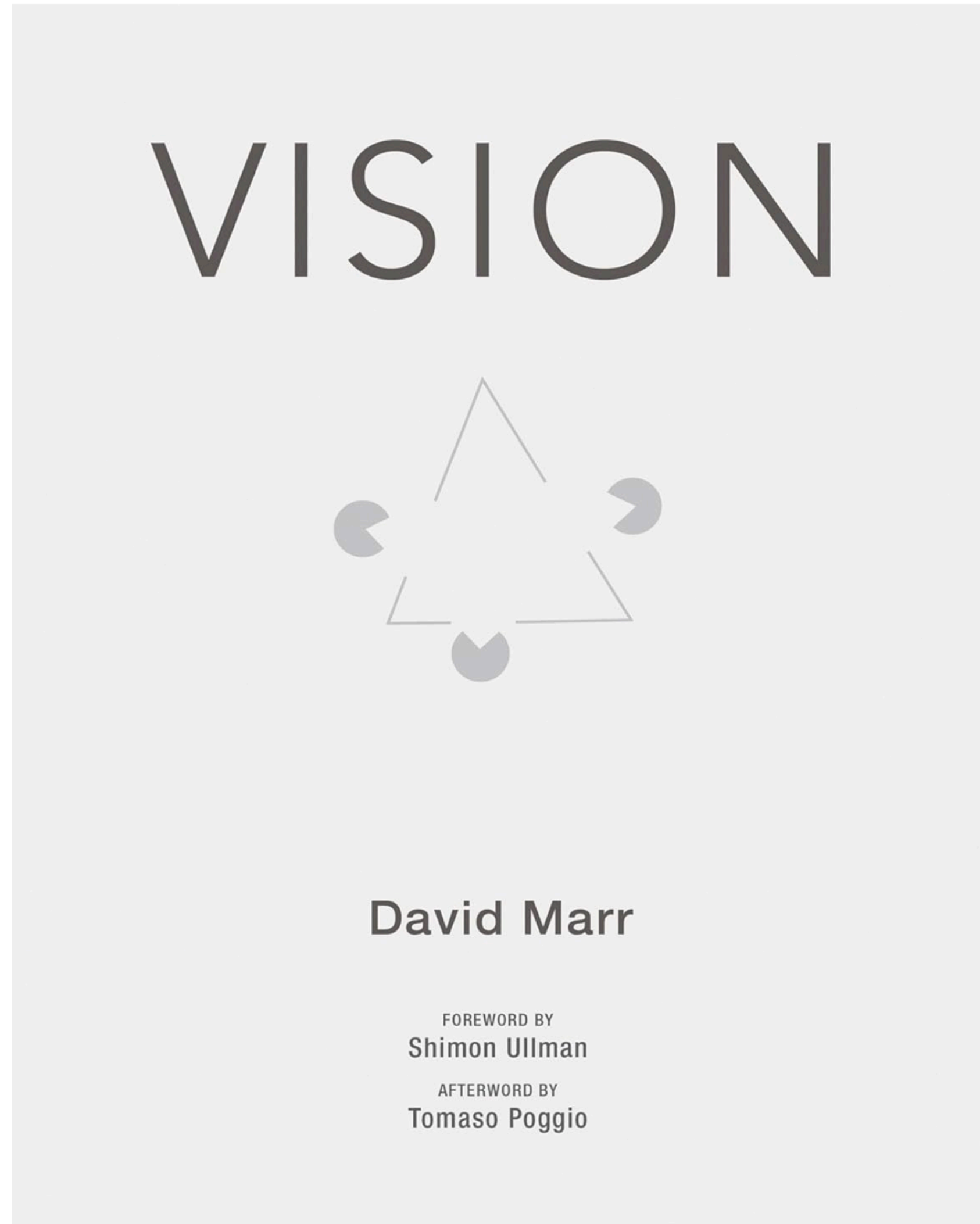
Symposium on Information Theory @ MIT
(Birth of Cognitive Science)



Herb Simon (1916 - 2001)
Allen Newell (1927 - 1992)

A How-to Guide

Marr's **three** levels

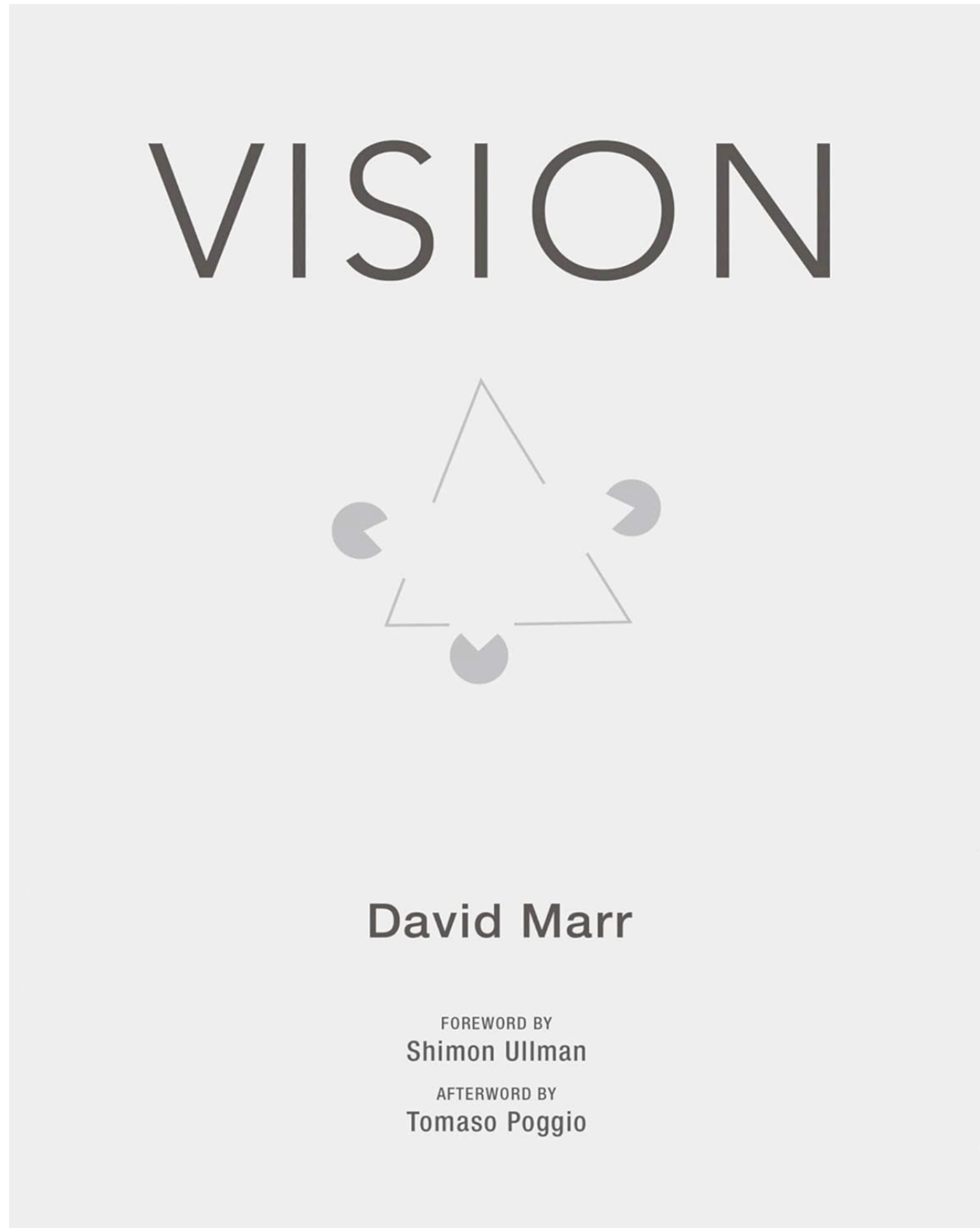


Marr's **three** levels

Computation

Representation and algorithm

Implementation



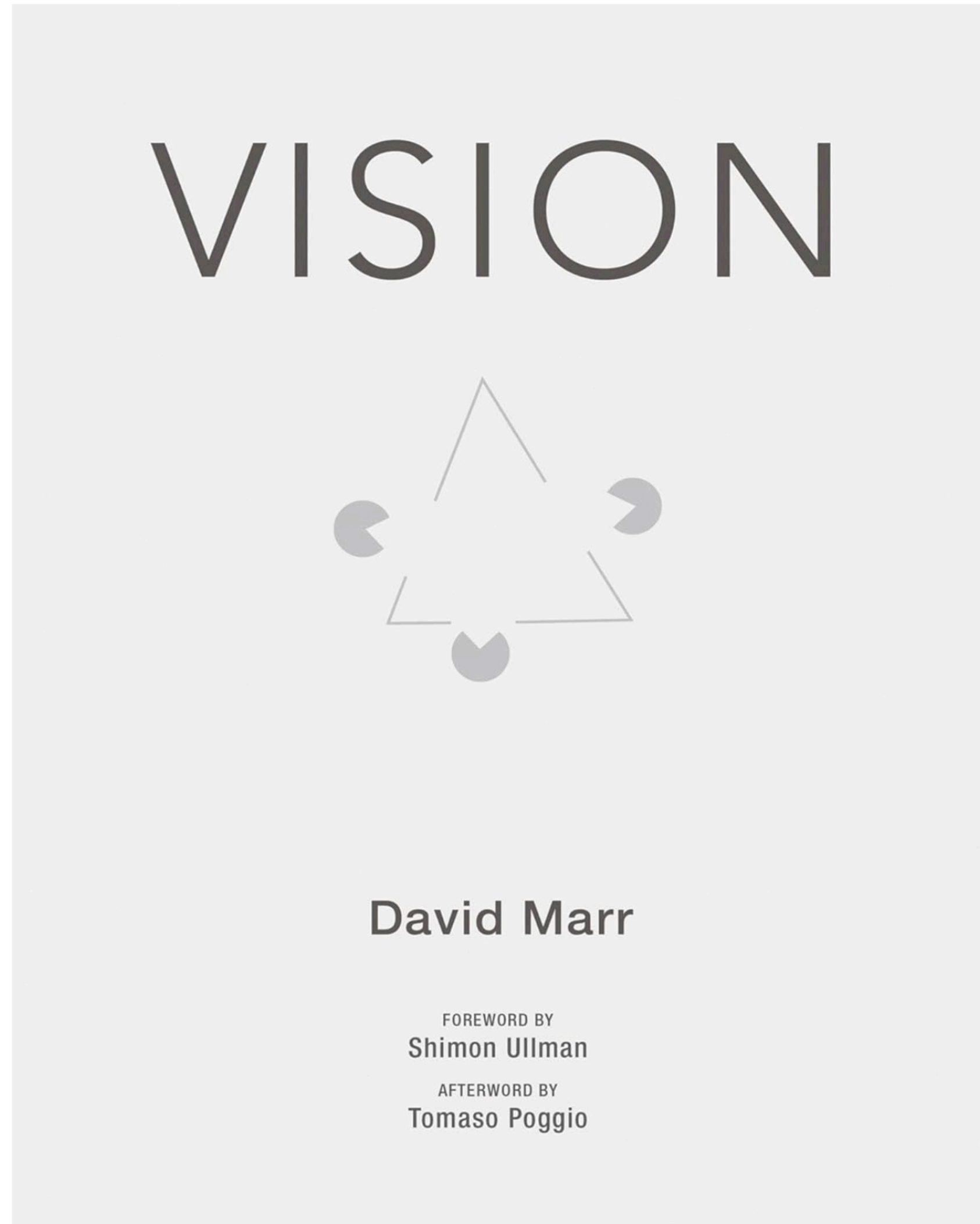
Marr's **three** levels

Computation

What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?

Representation and algorithm

Implementation



Marr's **three** levels

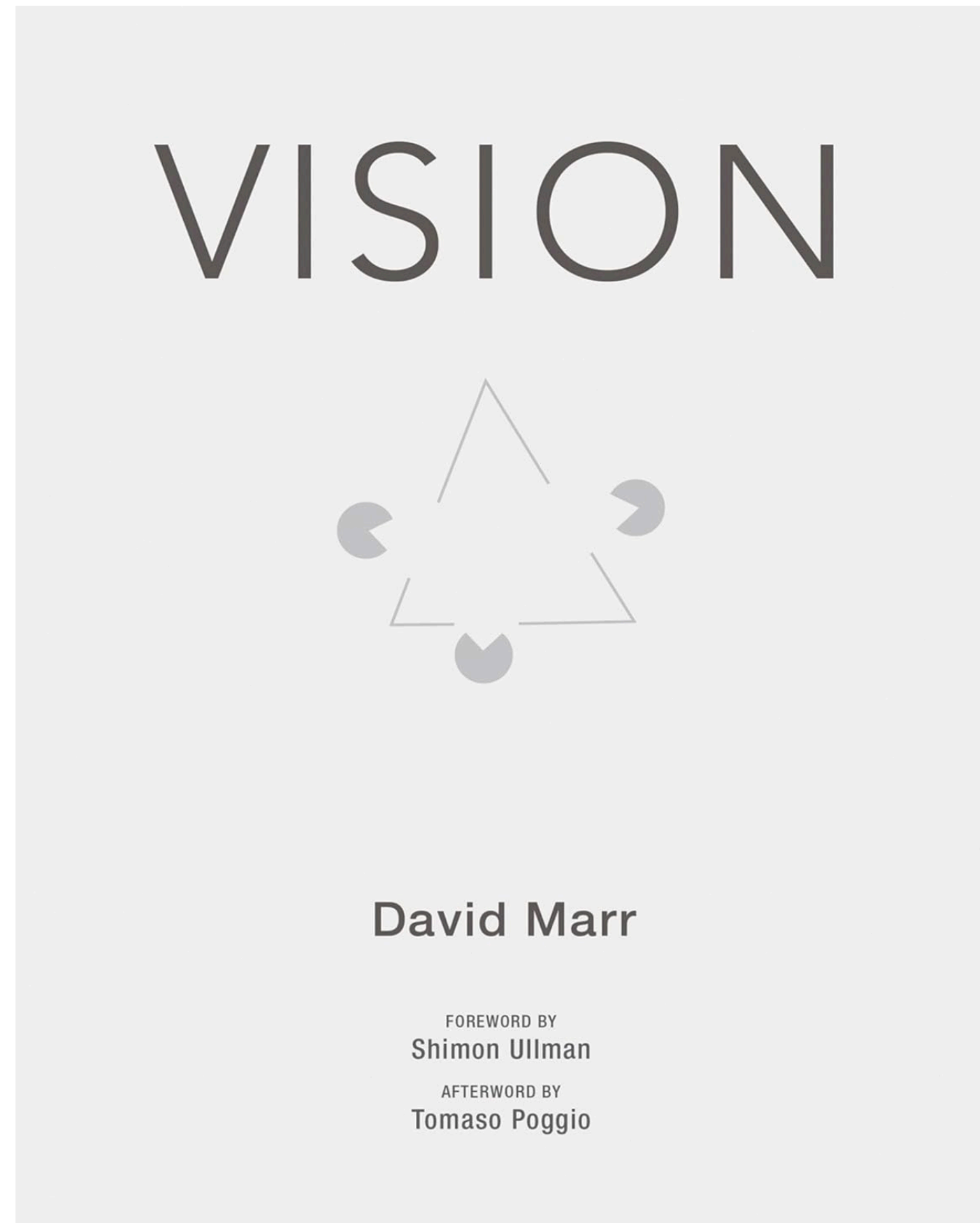
Computation

What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?

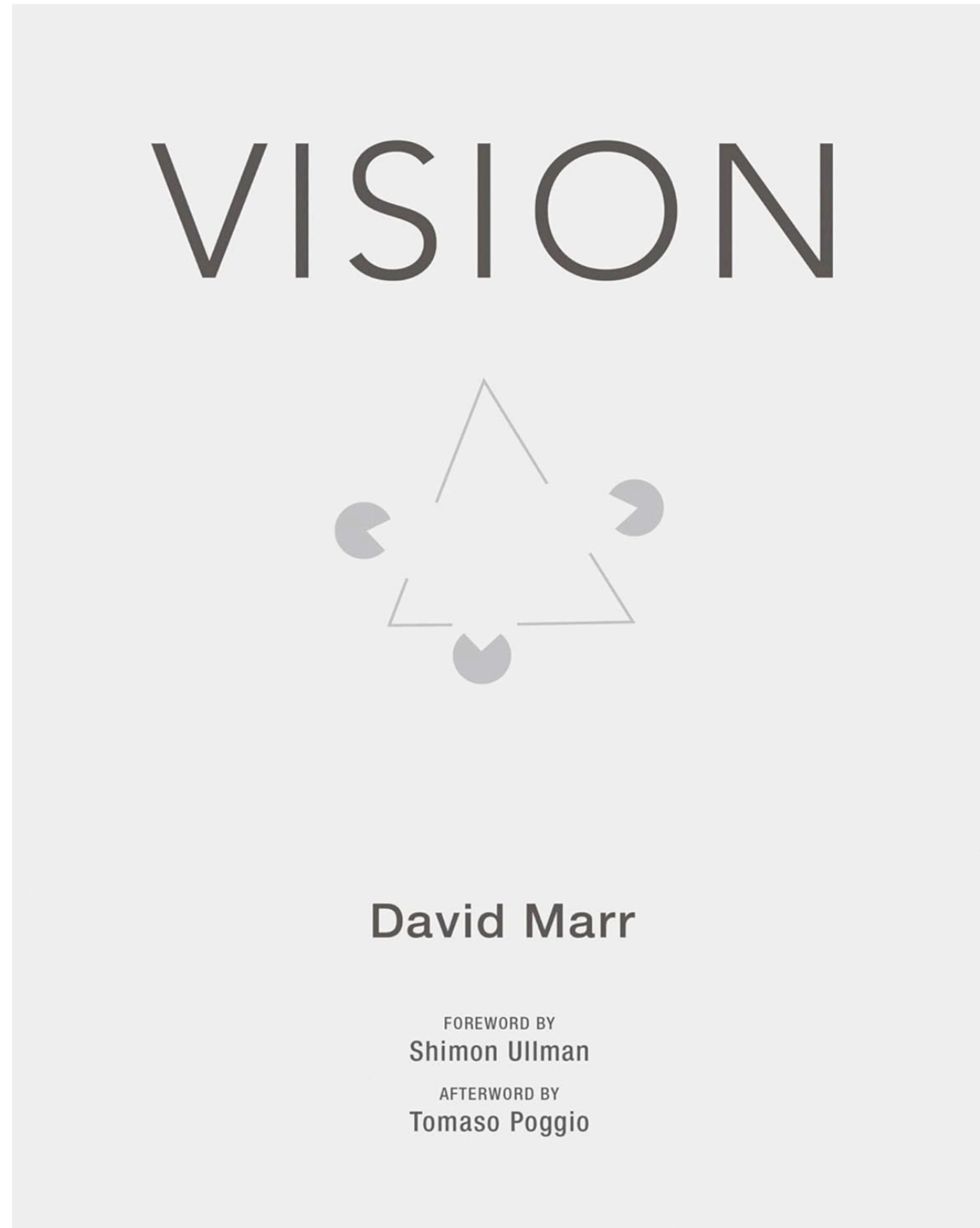
Representation and algorithm

What is the representation for the input and output, and the algorithm for the transformation?

Implementation



Marr's **three** levels



Computation

What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?

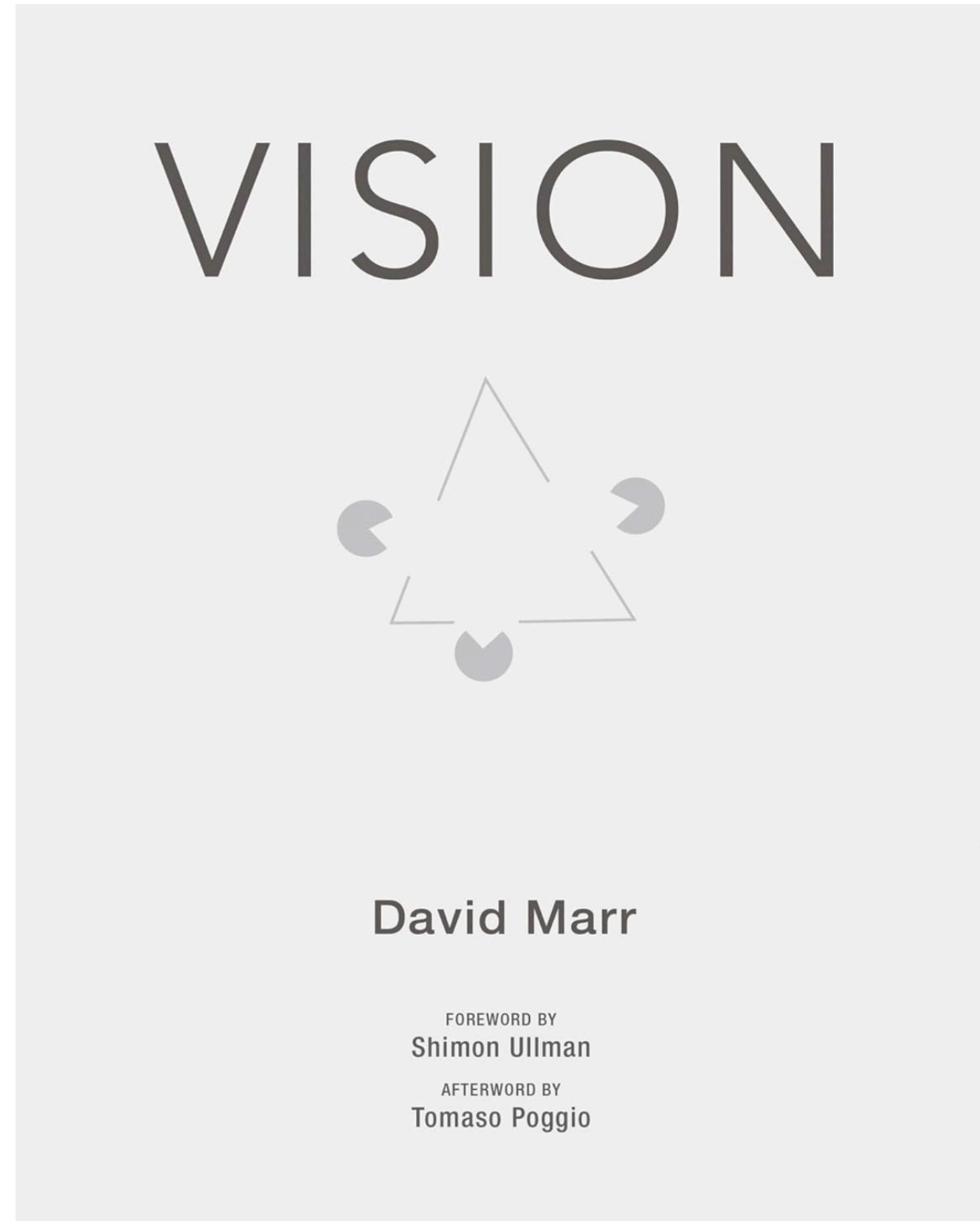
Representation and algorithm

What is the representation for the input and output, and the algorithm for the transformation?

Implementation

How can the representation and algorithm be realized physically in human brain and body?

Marr's **three** levels



Representation and algorithm

What is the representation for the input and output,
and the algorithm for the transformation?

Implementation

How can the representation and algorithm be realized
physically in human brain and body?

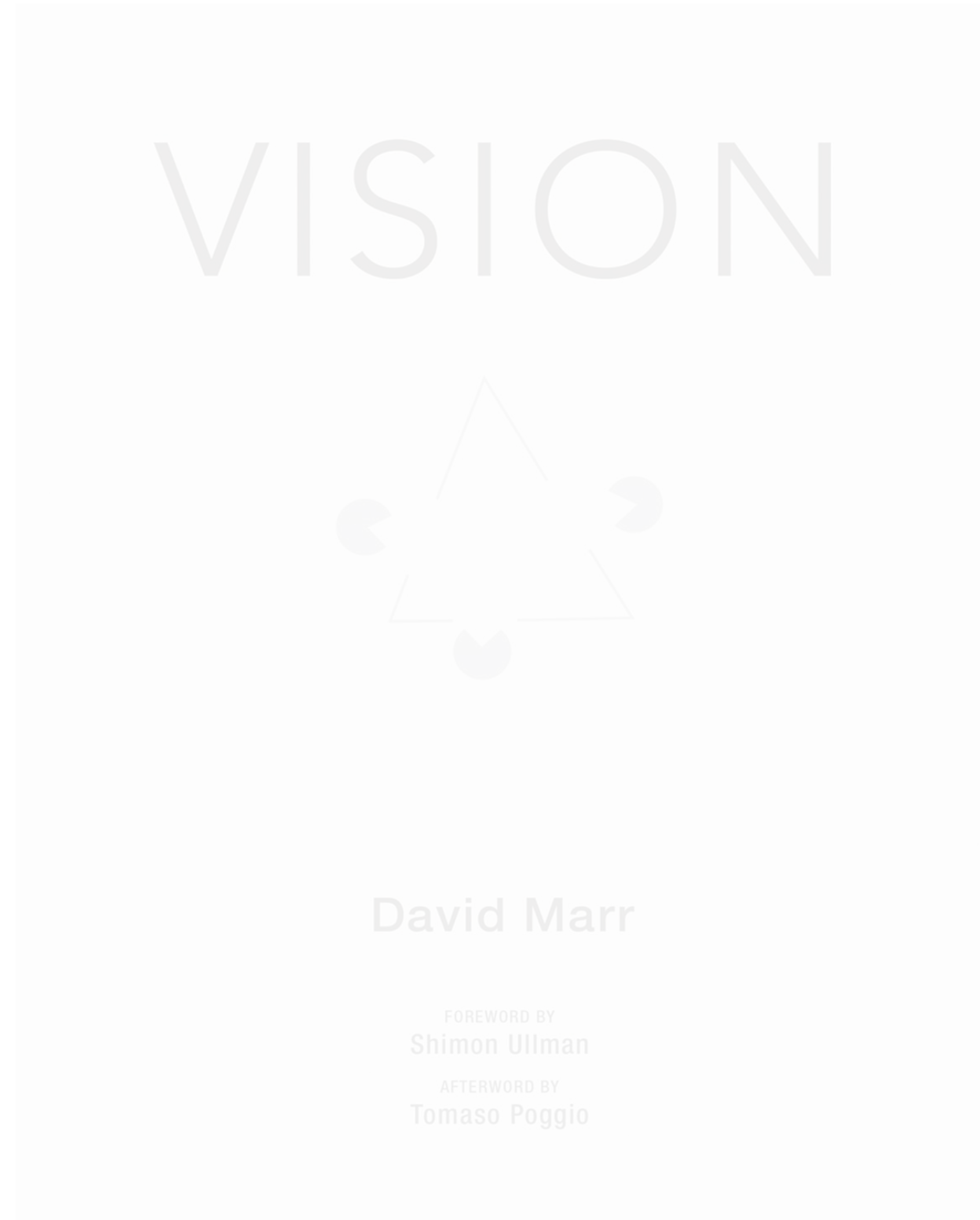
MIND I: Cognitive Psychology

MIND II: Neuroscience

Marr's **three** levels

Computation

What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?



Five Easy Steps

Step 1: Find an interesting aspect of cognition

Step 2: Identify the underlying computational problem, the environment constraints

Step 3: Work out the *optimal* solution to that problem

Step 4: See how well that solution corresponds to human behavior
(read as: do some experiments!)

Step 5: Refine steps 2-4.

Five Easy Steps

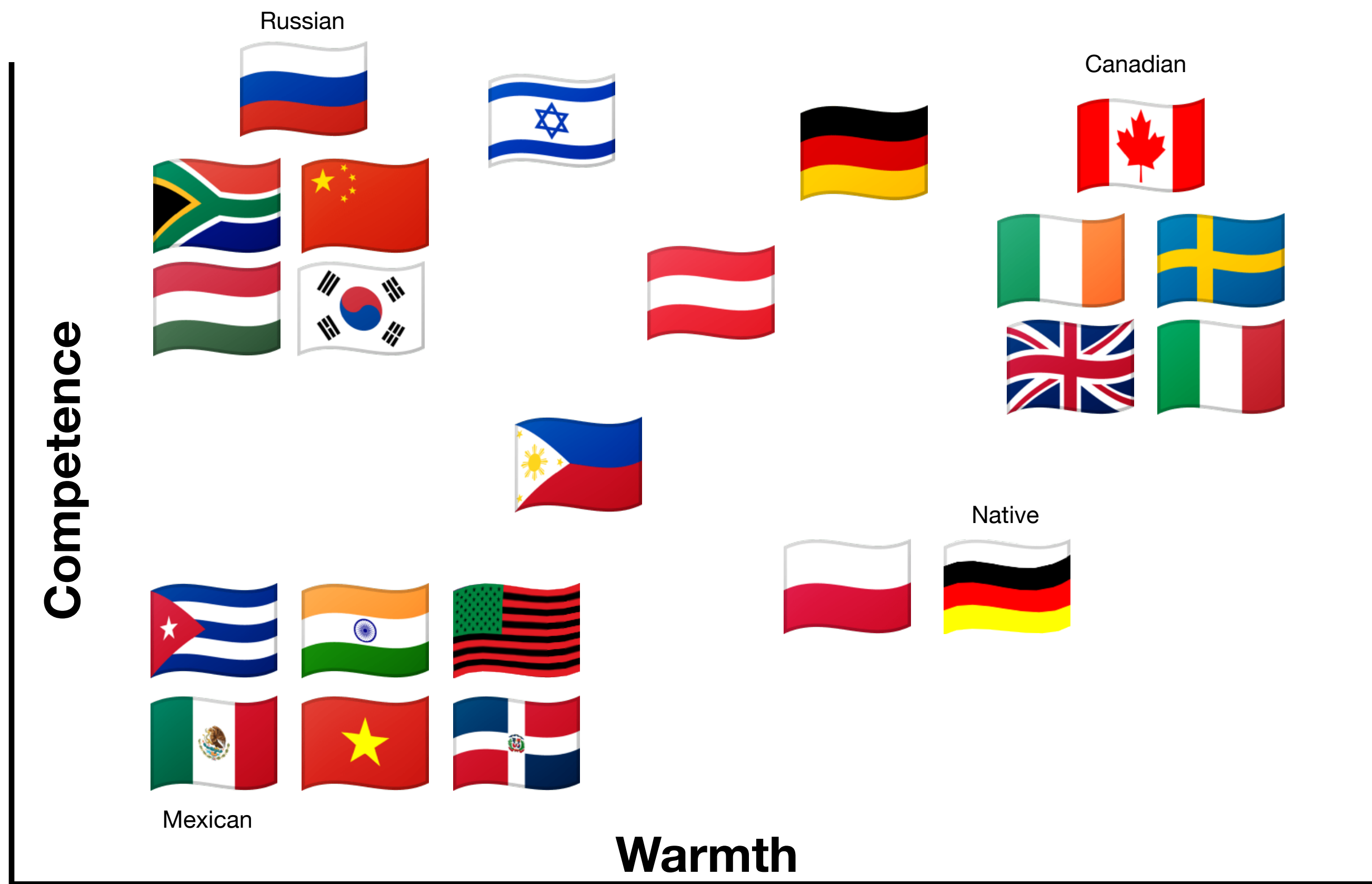
MIND III: Contextual Influence

Step 2: Identify the underlying computational problem, the environment constraints

Step 3: Work out the *optimal* solution to that problem

An Illustration: Stereotype and Exploration

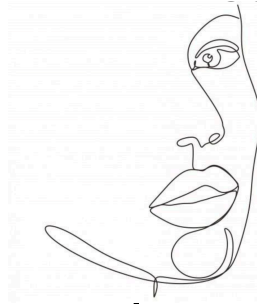






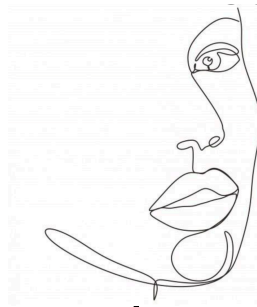
Where do **stereotypes** come from?

Existing explanations:



Where do **stereotypes** come from?

Existing explanations:



Motivational biases

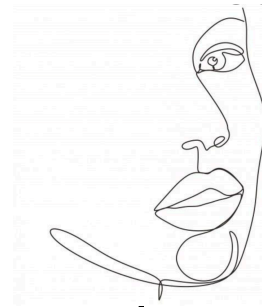
- Identity
- Dominance



Social identity theory (Tajfel & Turner, 1979)
In-group favoritism (Brewer, 1999)
Social dominance theory (Sidanius & Pratto, 1999)
System justification theory (Jost & Banaji, 1994)

Where do **stereotypes** come from?

Existing explanations:



Motivational biases

- Identity
- Dominance

Cognitive biases

- Limited memory
- Selective attention

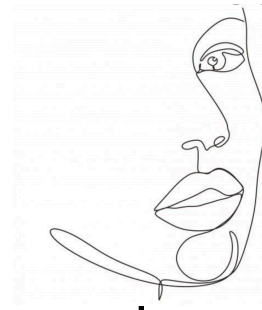


Social identity theory (Tajfel & Turner, 1979)
In-group favoritism (Brewer, 1999)
Social dominance theory (Sidanius & Pratto, 1999)
System justification theory (Jost & Banaji, 1994)

Error-prone heuristics (Tversky & Kahneman, 1974)
Cognitive miser (Fiske & Taylor, 1984)
Illusory correlation (Hamilton & Gifford, 1976)
Attention in stereotype formation (Sherman et al., 2009)

Where do **stereotypes** come from?

Existing explanations:



Motivational biases

- Identity
- Dominance

Cognitive biases

- Limited memory
- Selective attention



Sample biases

- Unequal group size

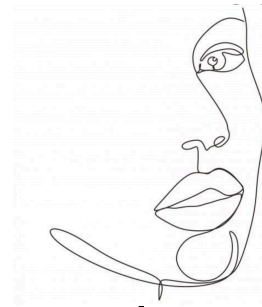
Social identity theory (Tajfel & Turner, 1979)
In-group favoritism (Brewer, 1999)
Social dominance theory (Sidanius & Pratto, 1999)
System justification theory (Jost & Banaji, 1994)

Error-prone heuristics (Tversky & Kahneman, 1974)
Cognitive miser (Fiske & Taylor, 1984)
Illusory correlation (Hamilton & Gifford, 1976)
Attention in stereotype formation (Sherman et al., 2009)

Beware of samples (Fiedler, 2000)
Hot-stove effect (Denrell, 2005)

Where do **stereotypes** come from?

Existing explanations:



Motivational biases

- Identity
- Dominance

Cognitive biases

- Limited memory
- Selective attention



Sample biases

- Unequal group size

Group differences

- Gender

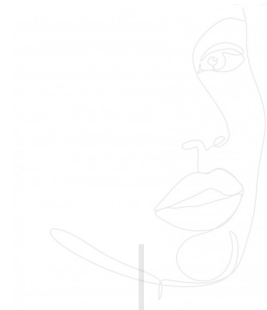
Social identity theory (Tajfel & Turner, 1979)
In-group favoritism (Brewer, 1999)
Social dominance theory (Sidanius & Pratto, 1999)
System justification theory (Jost & Banaji, 1994)

Error-prone heuristics (Tversky & Kahneman, 1974)
Cognitive miser (Fiske & Taylor, 1984)
Illusory correlation (Hamilton & Gifford, 1976)
Attention in stereotype formation (Sherman et al., 2009)

Beware of samples (Fiedler, 2000)
Hot-stove effect (Denrell, 2005)

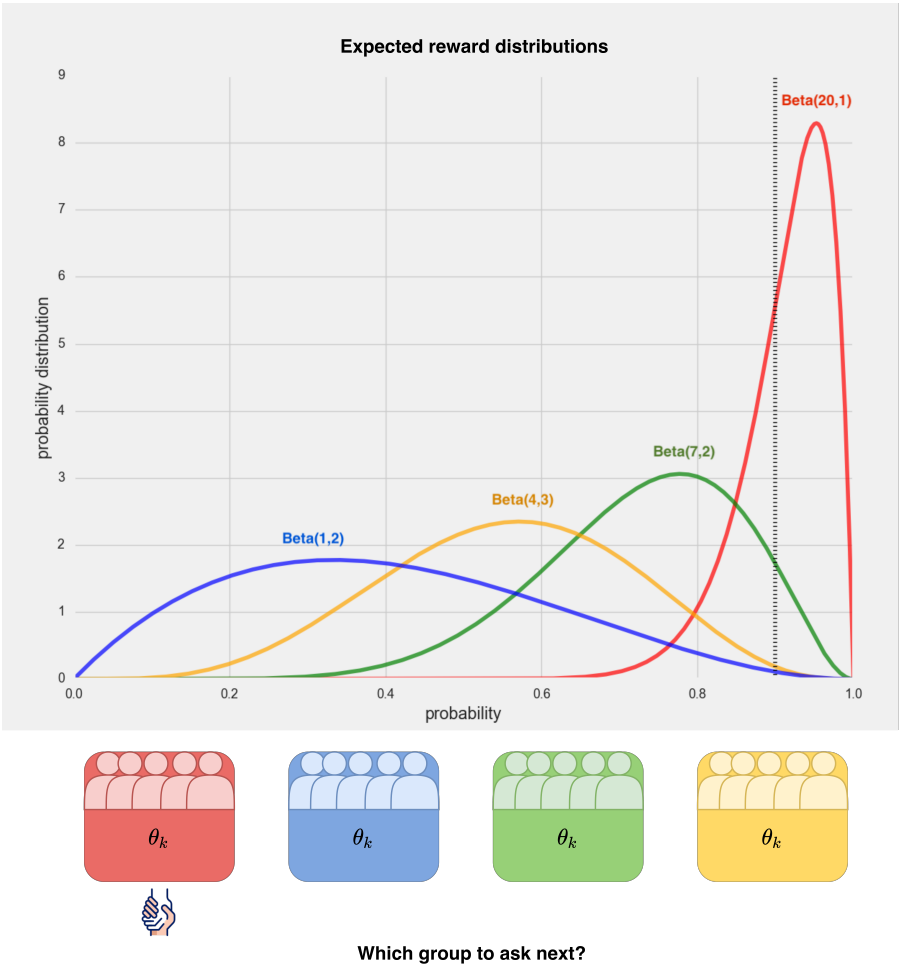
Biosocial constructionism (Wood & Eagly, 2012)
Stereotype accuracy (Jussim, 2017)

Where do **stereotypes** come from?



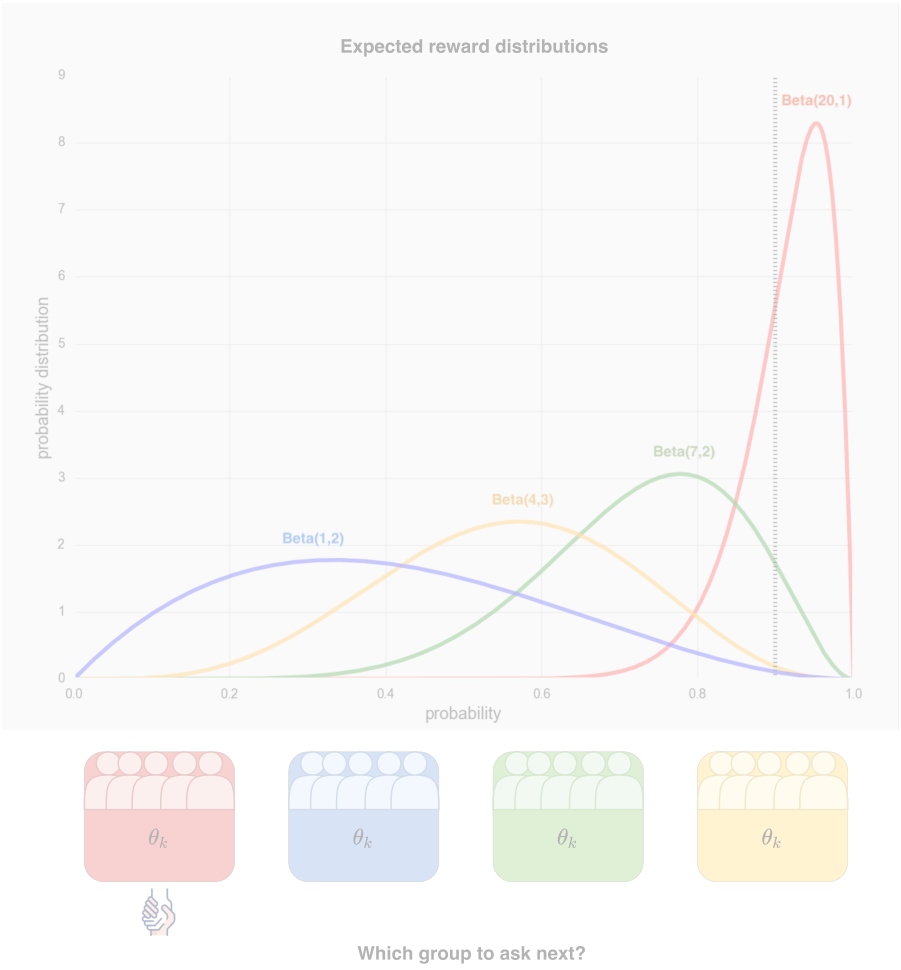
A 5th Mechanism: **Exploration**

A 5th Mechanism: **Exploration**

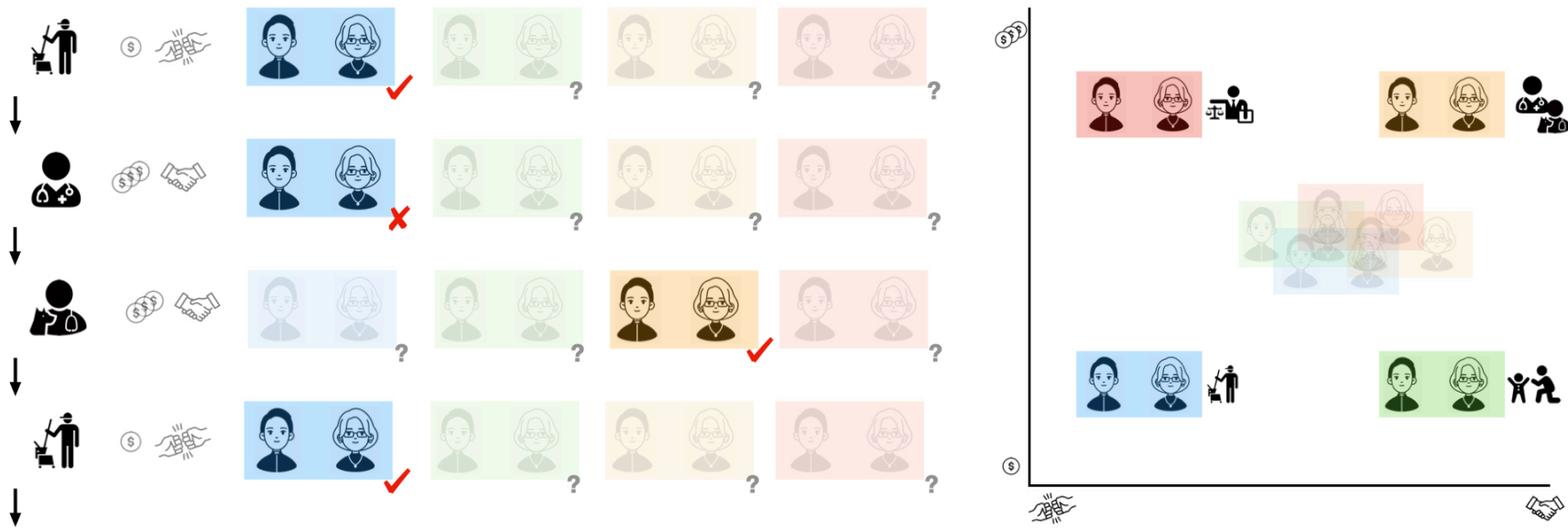


Sketching the Mechanism

A 5th Mechanism: **Exploration**



Sketching the Mechanism



Enriching the Context

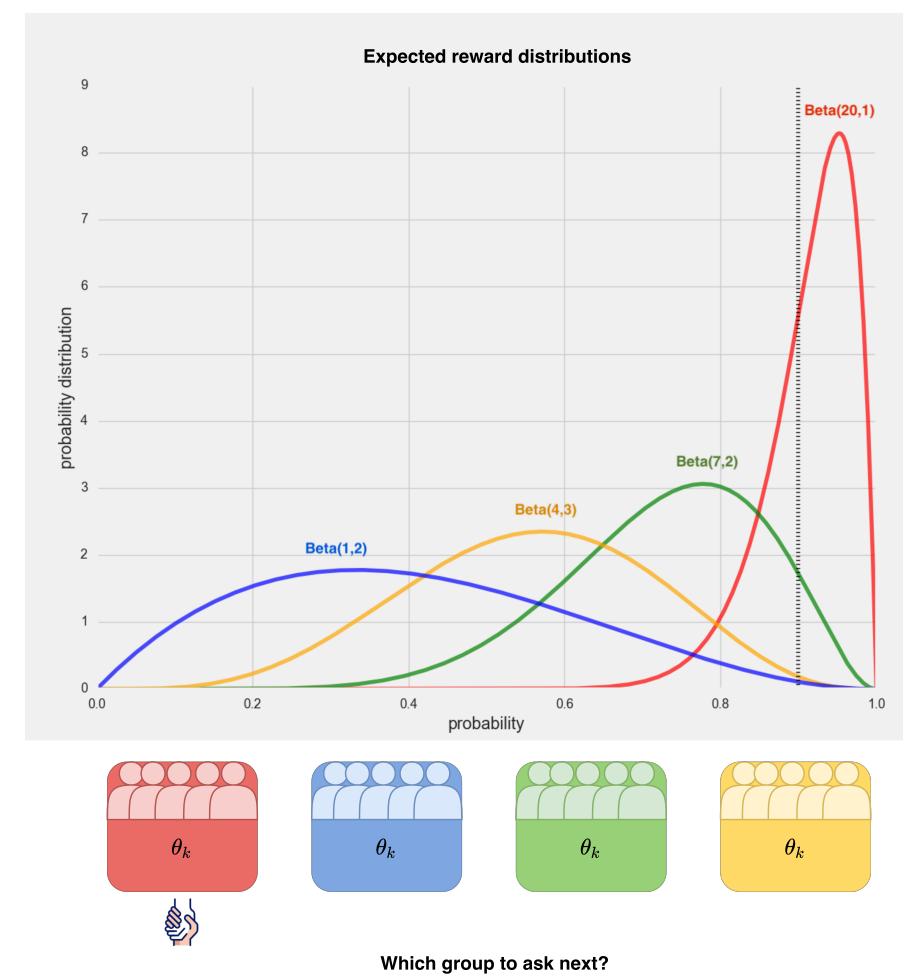
Sketching the Mechanism



Tom Griffiths



Susan Fiske

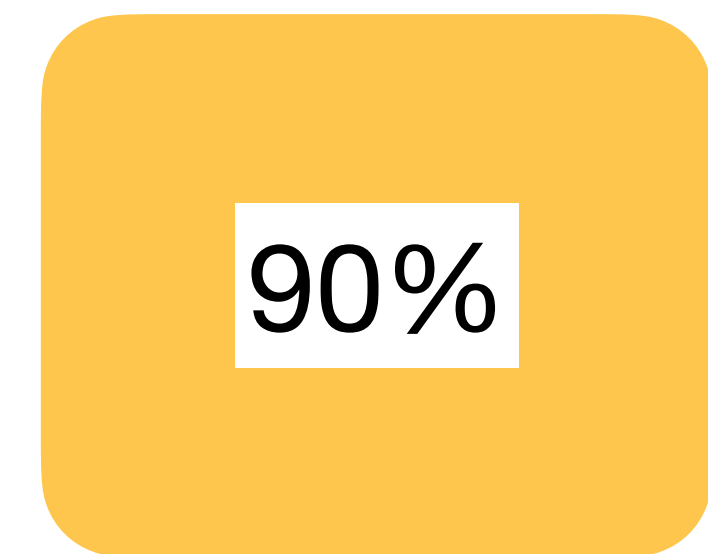
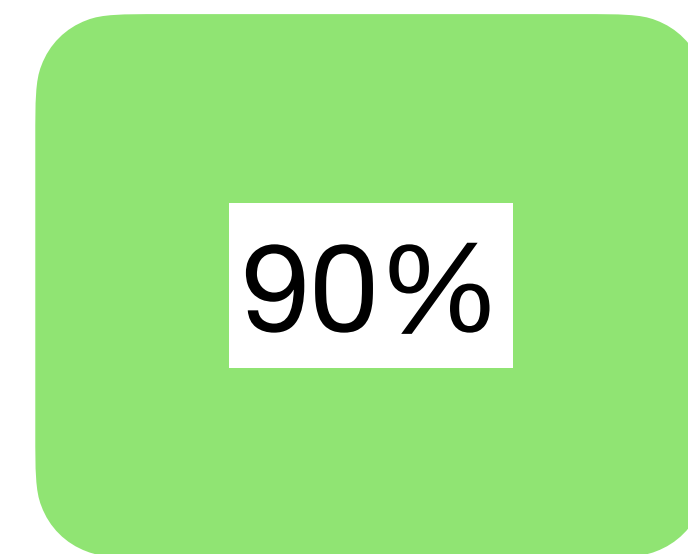
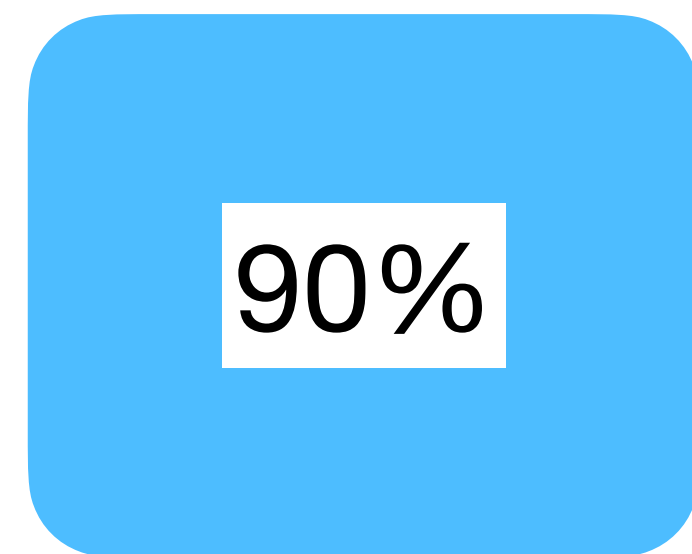
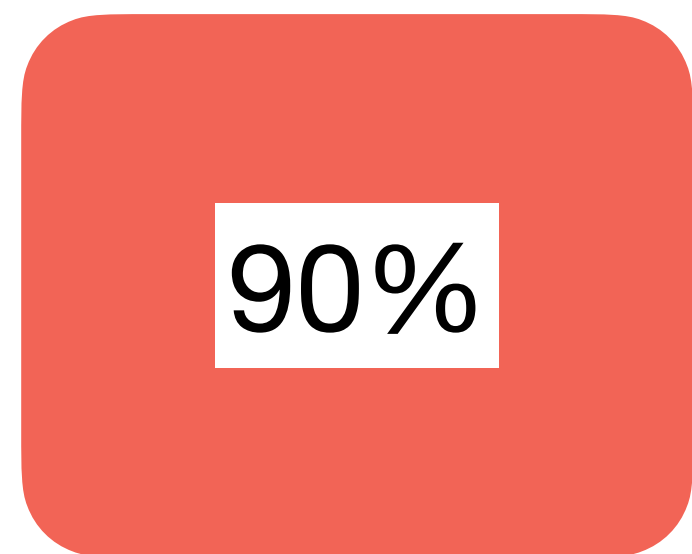
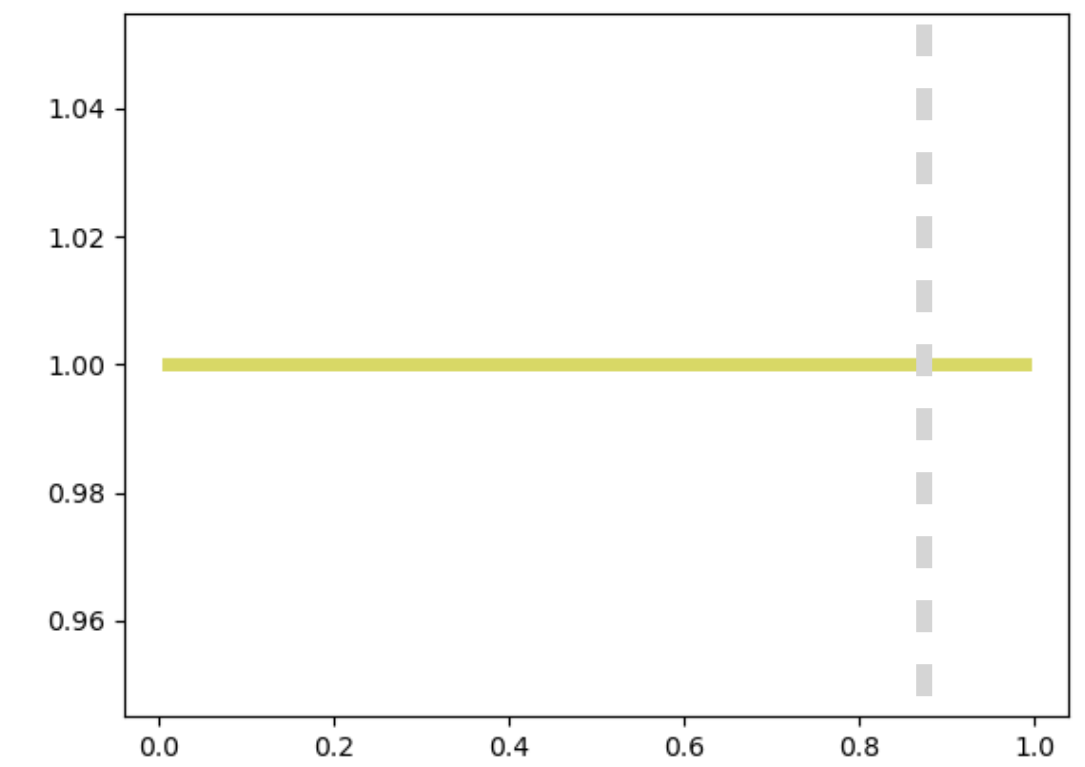
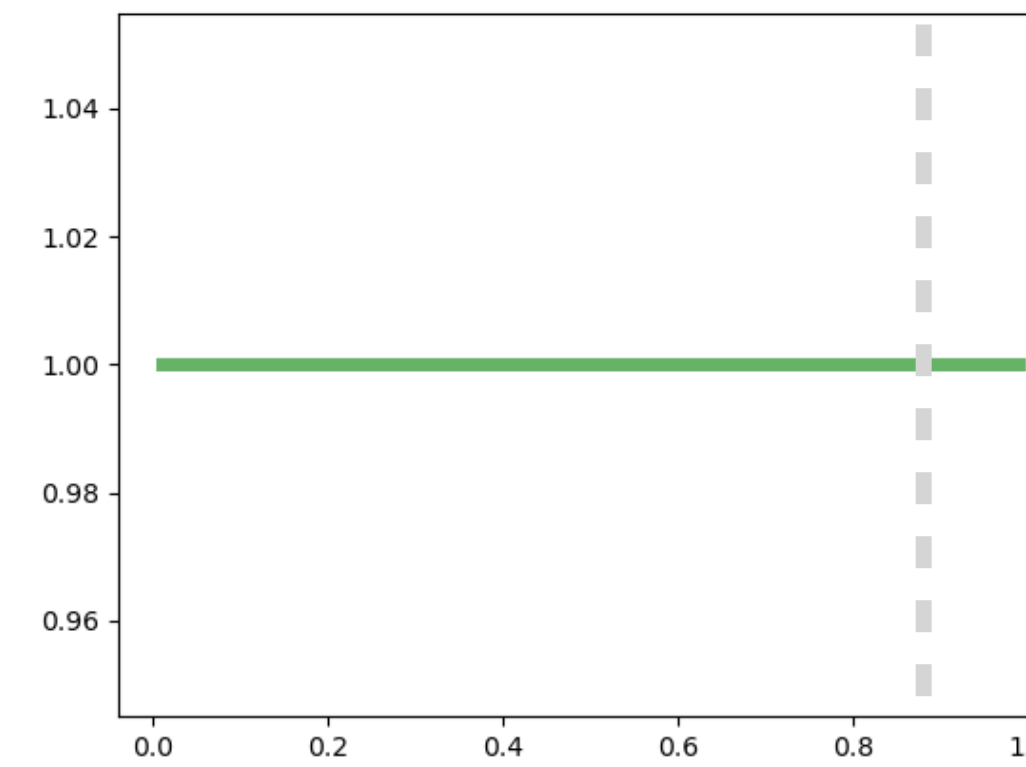
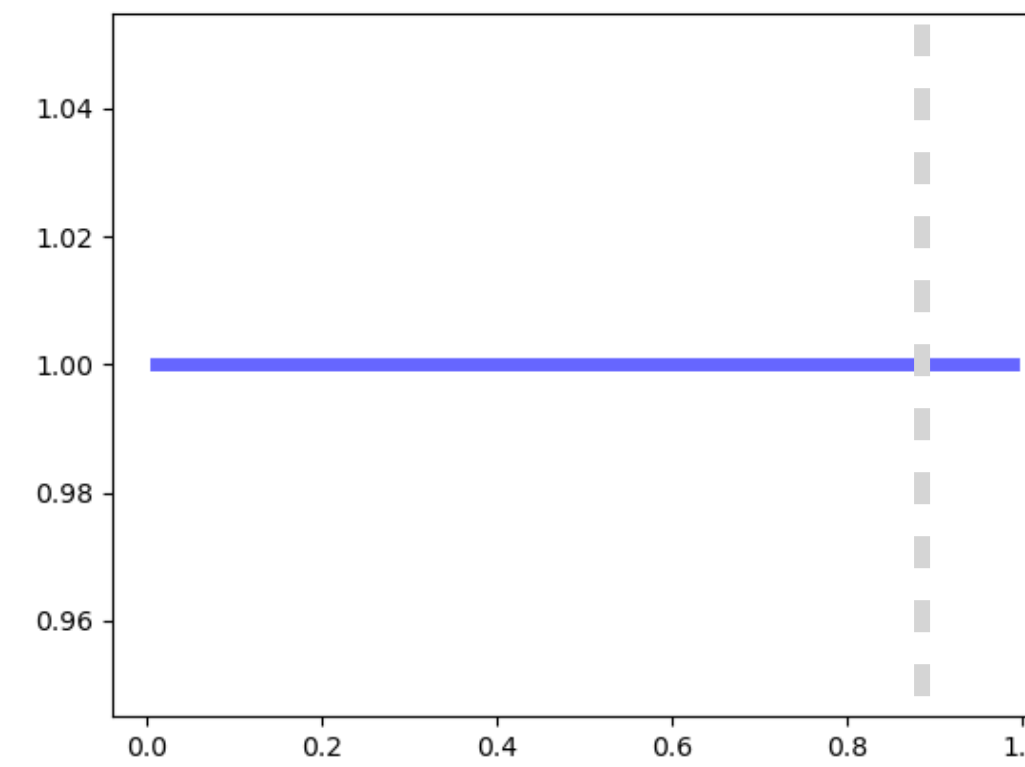
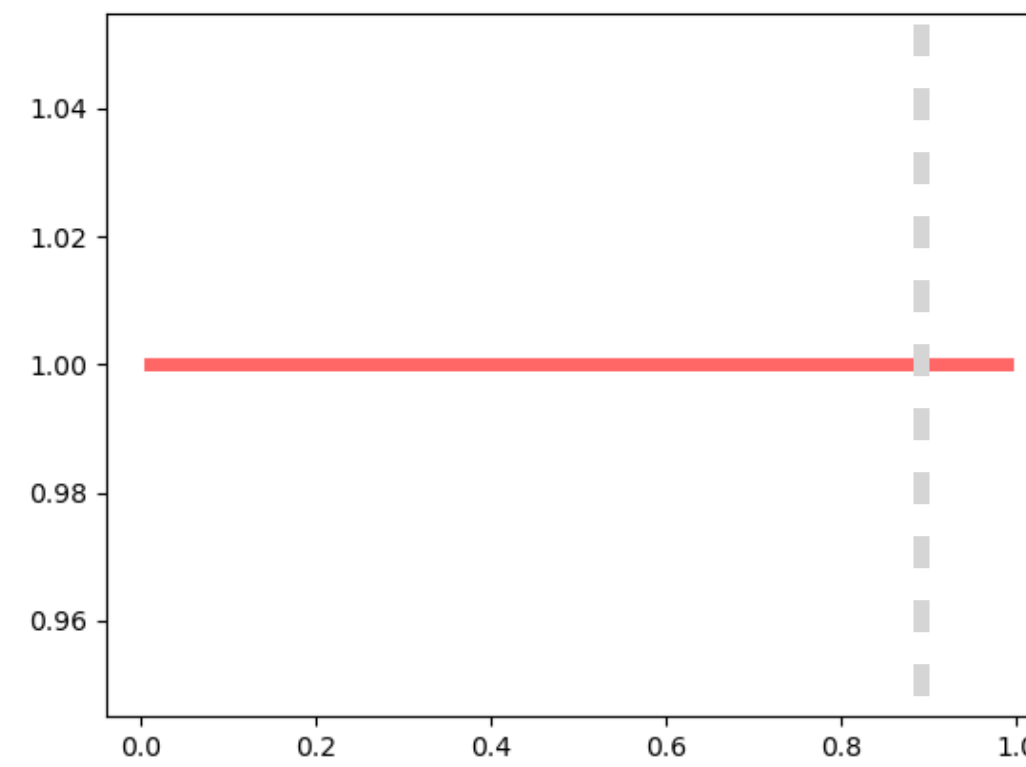


People believe that groups differ from each other even when they do not. Why?

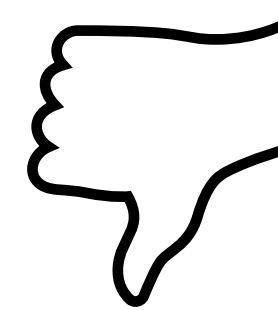
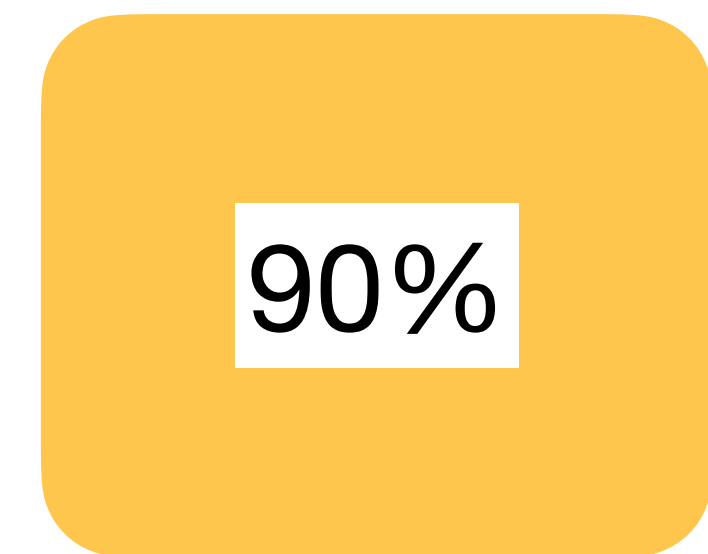
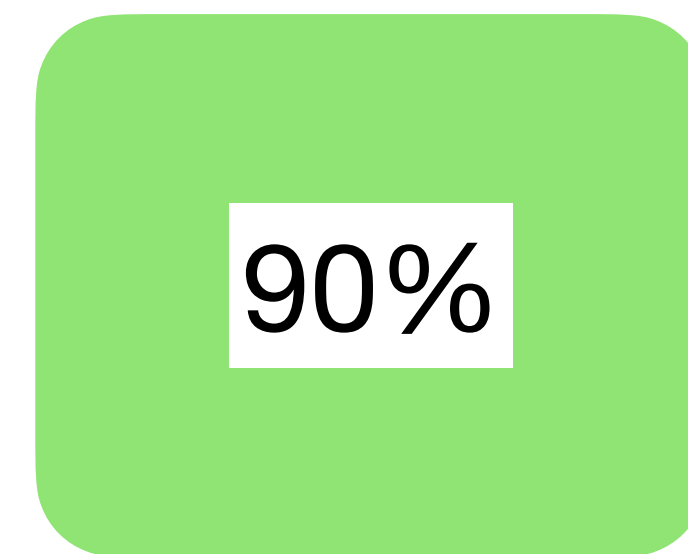
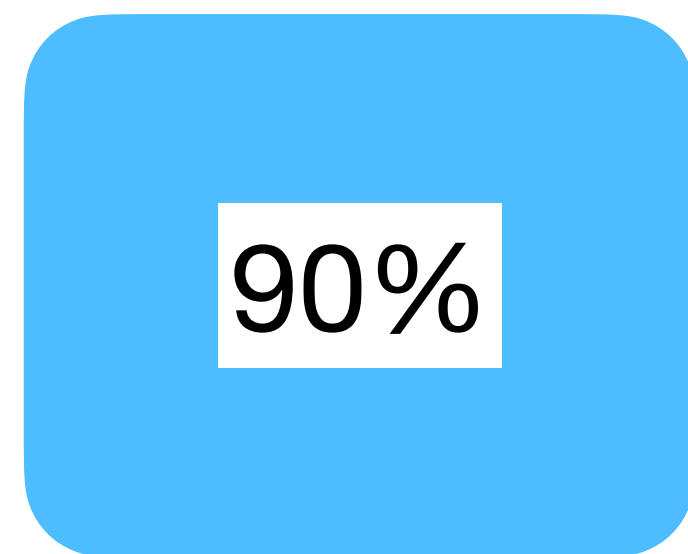
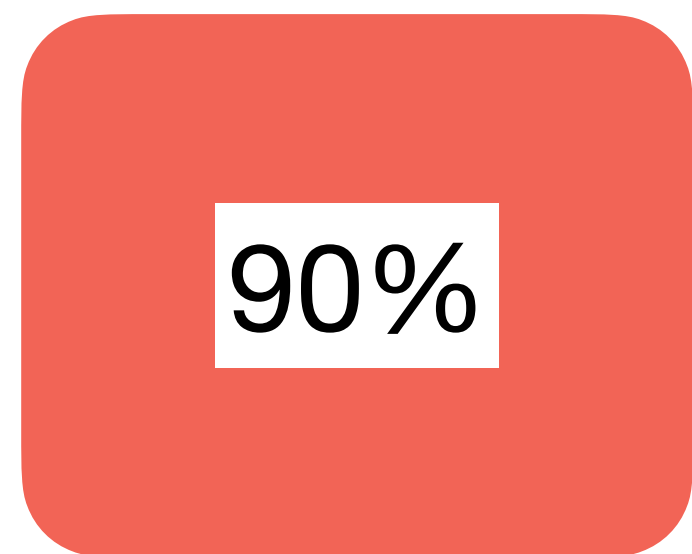
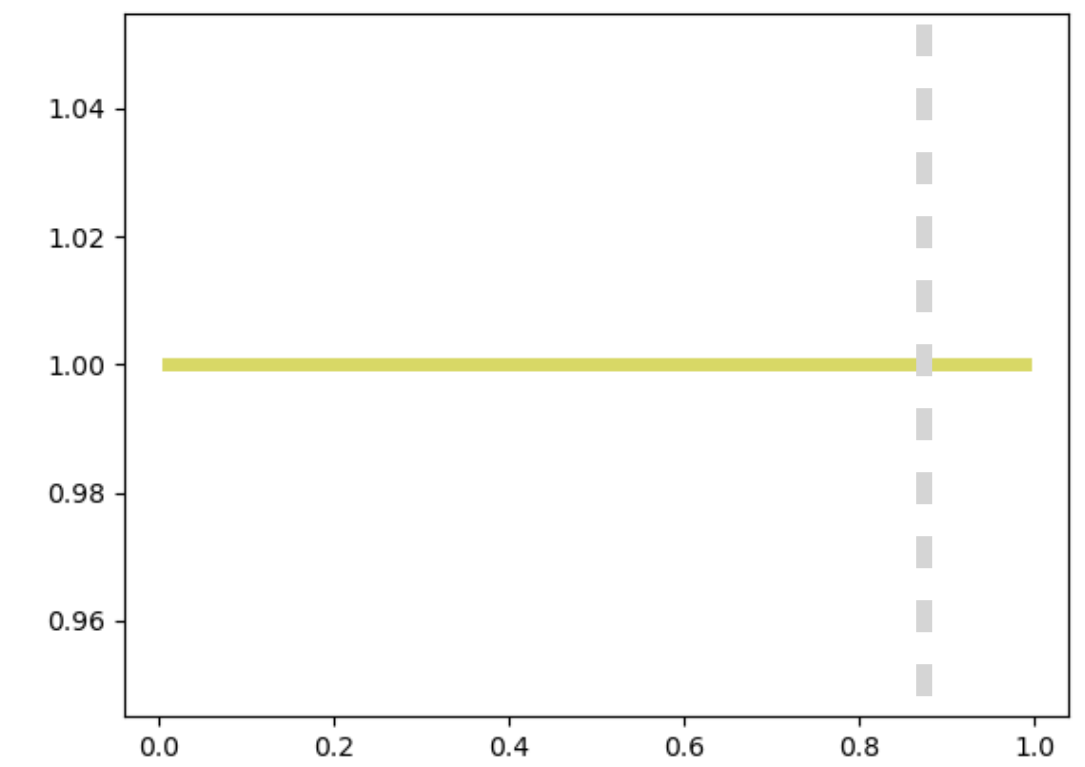
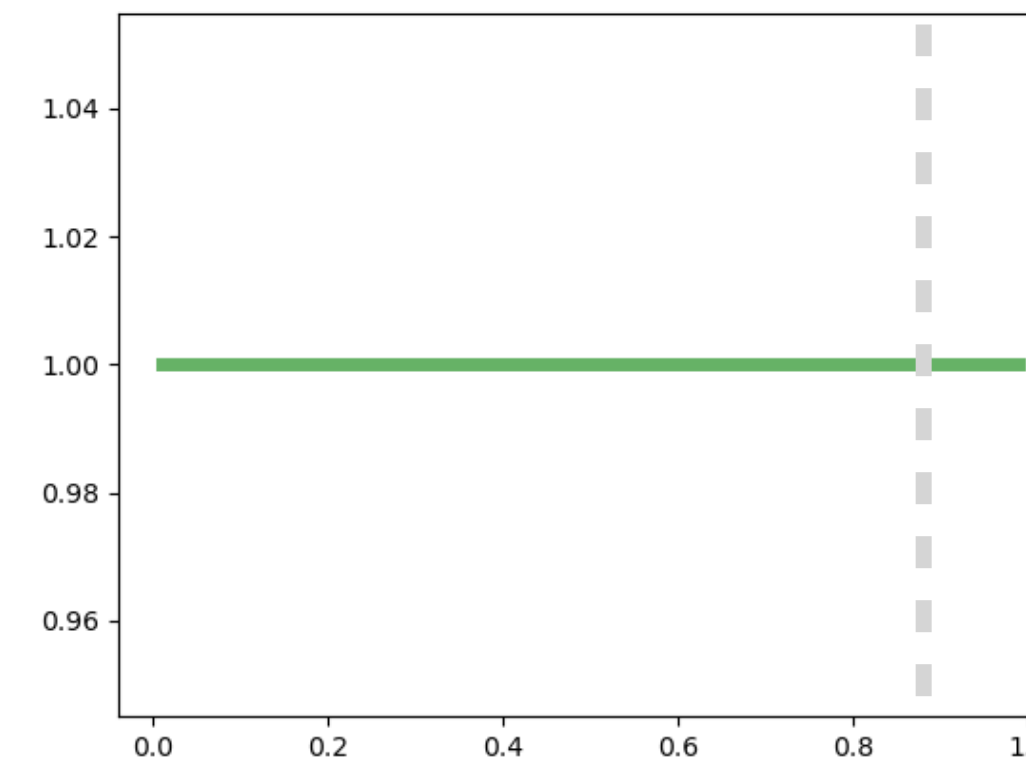
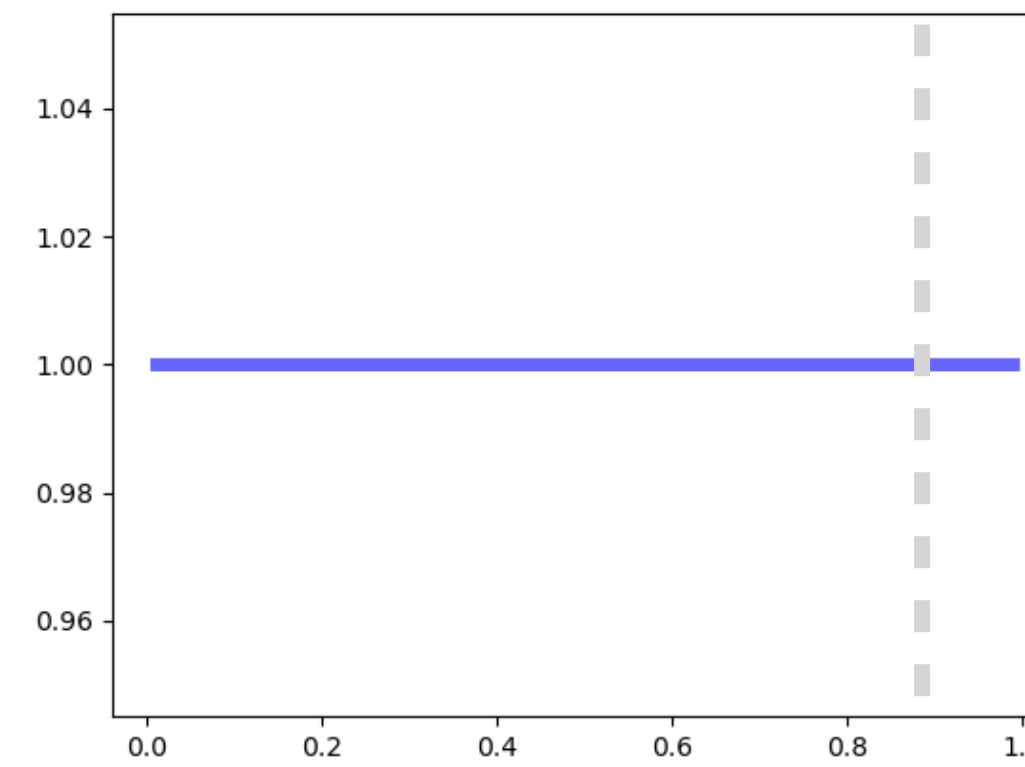
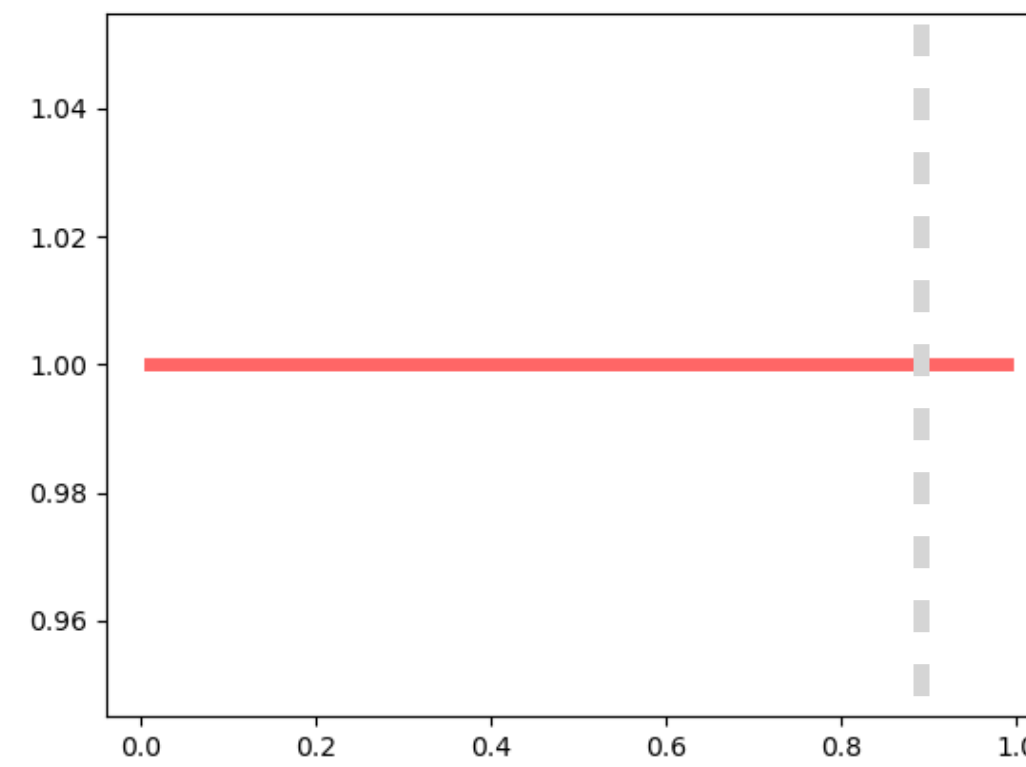
Sketching the Mechanism: **Intuition**



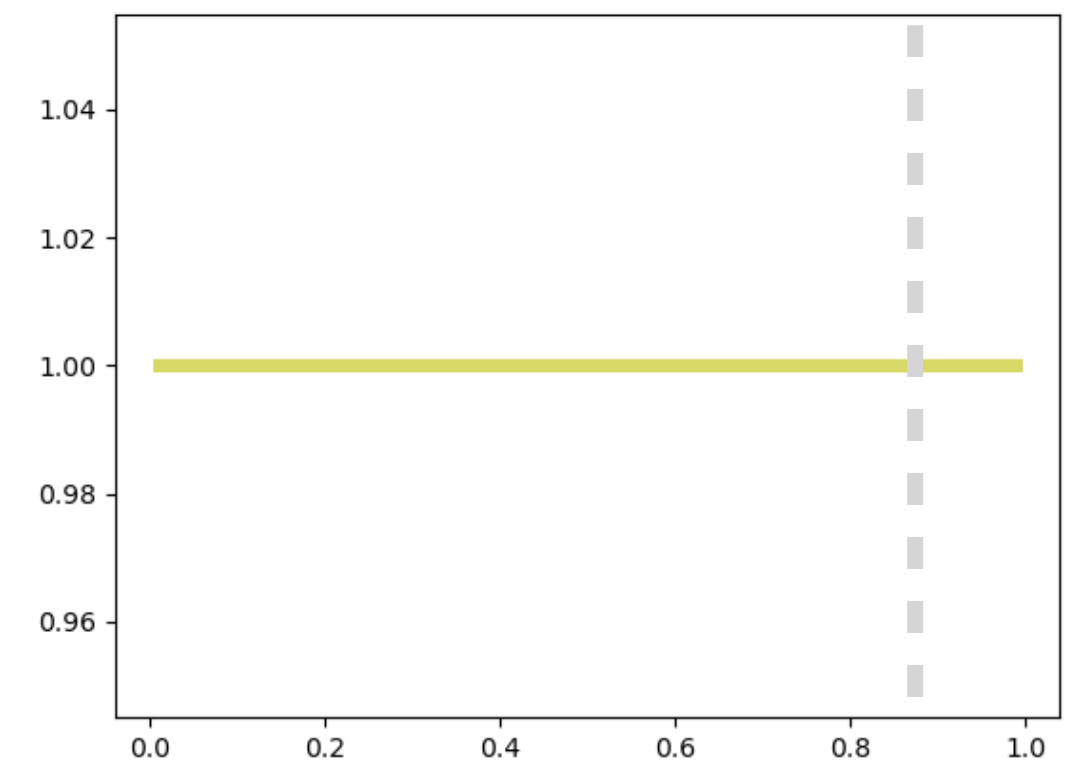
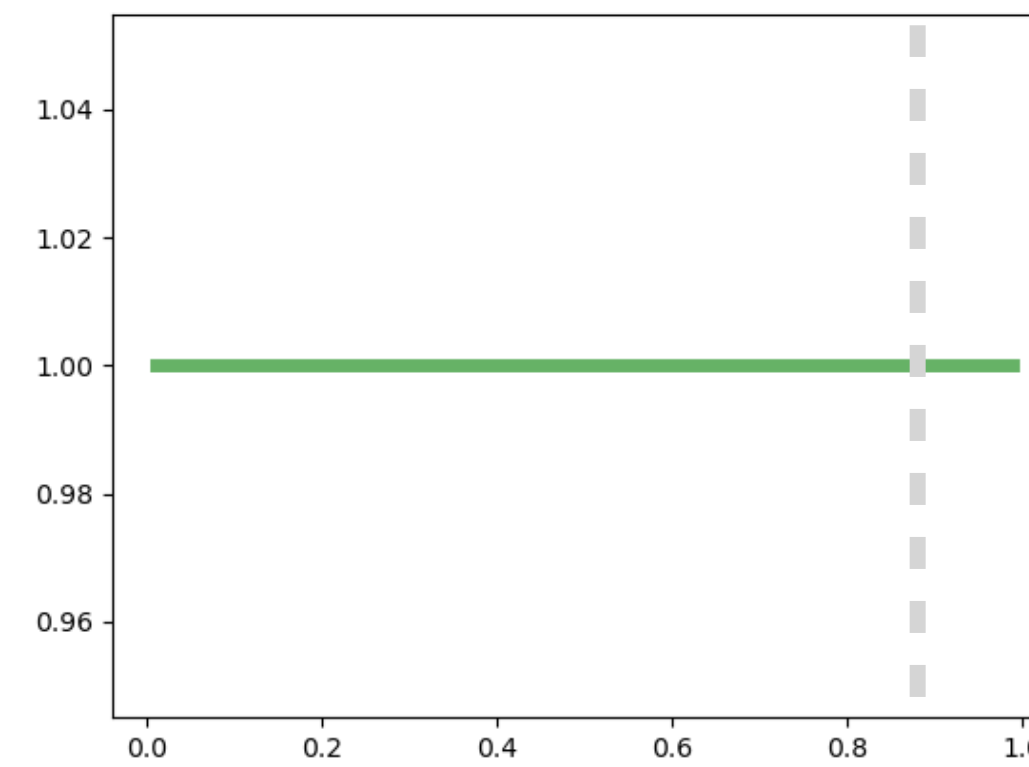
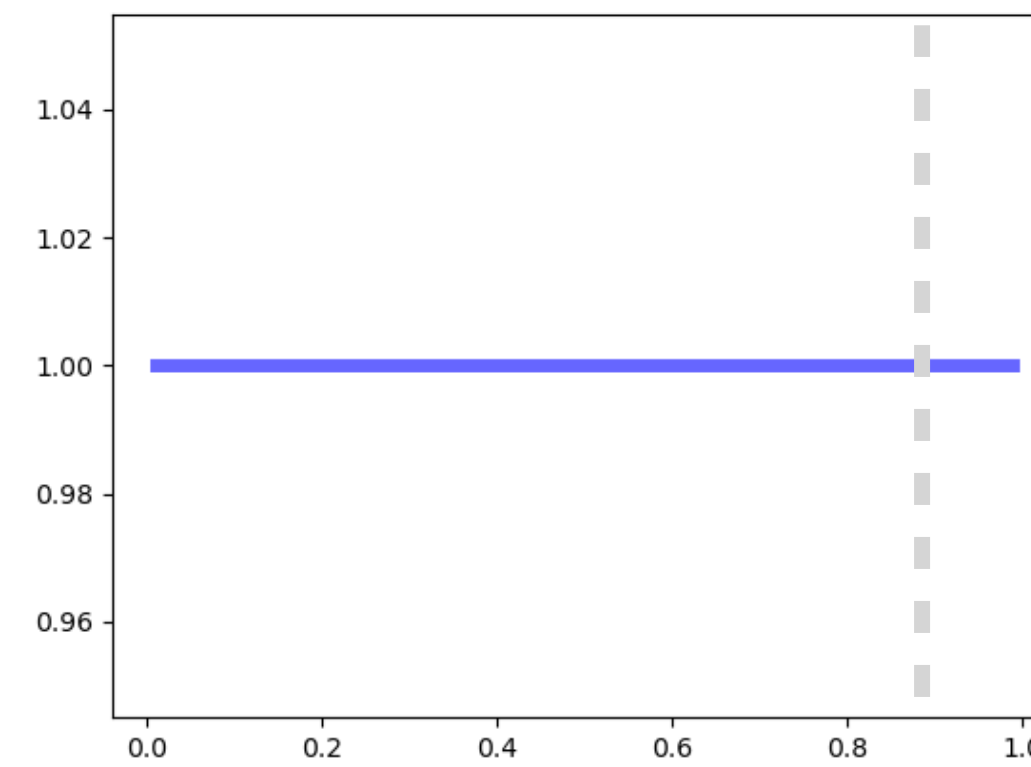
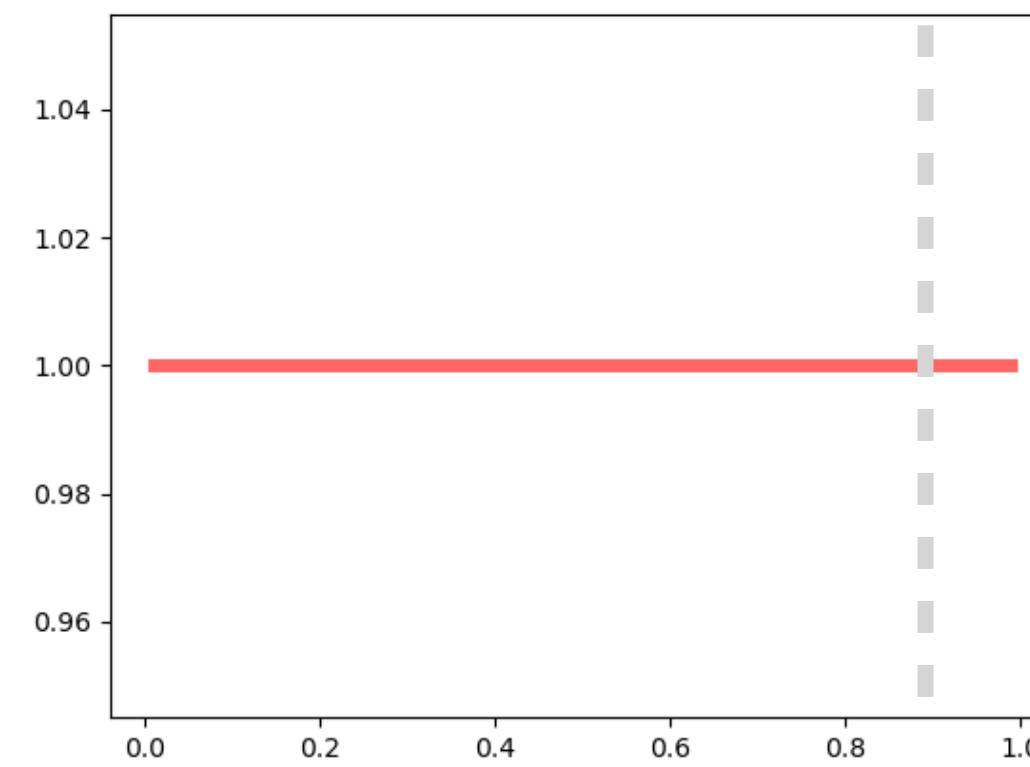
Sketching the Mechanism: Intuition



Sketching the Mechanism: Intuition

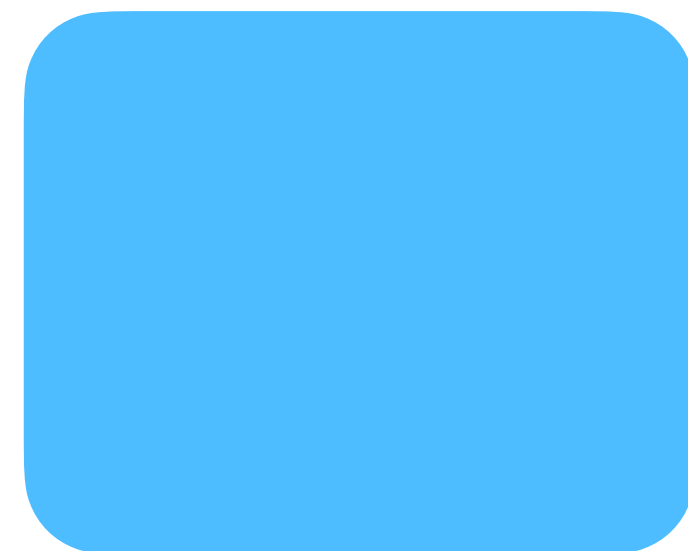
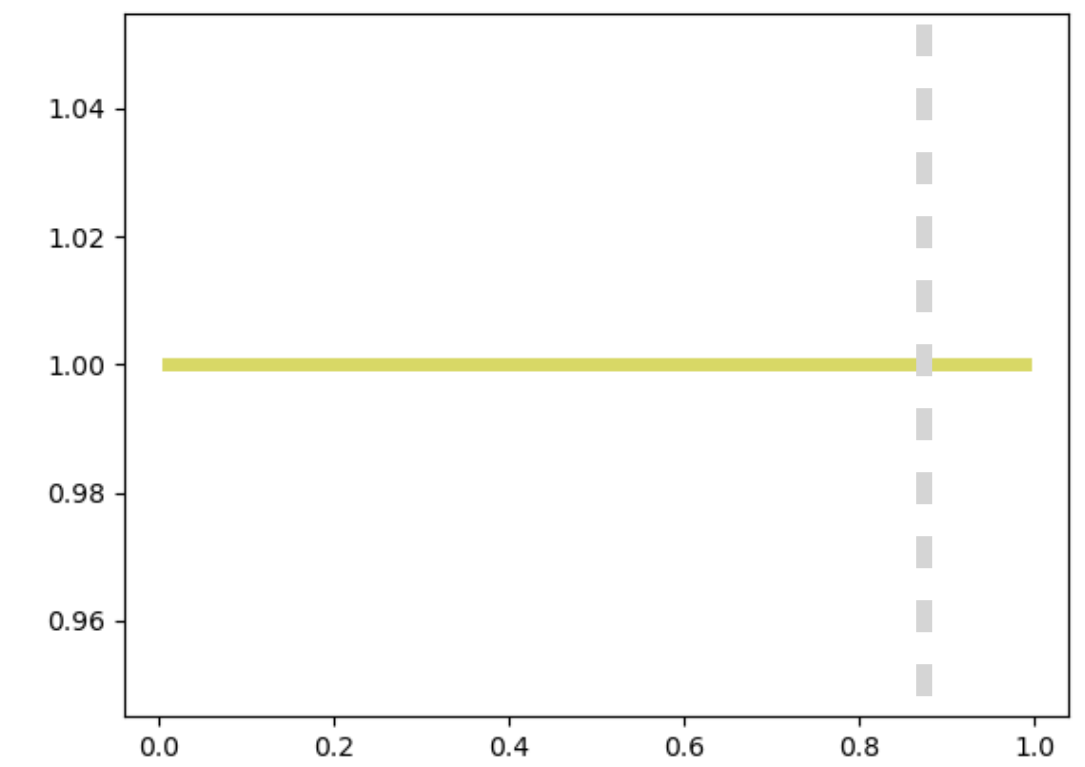
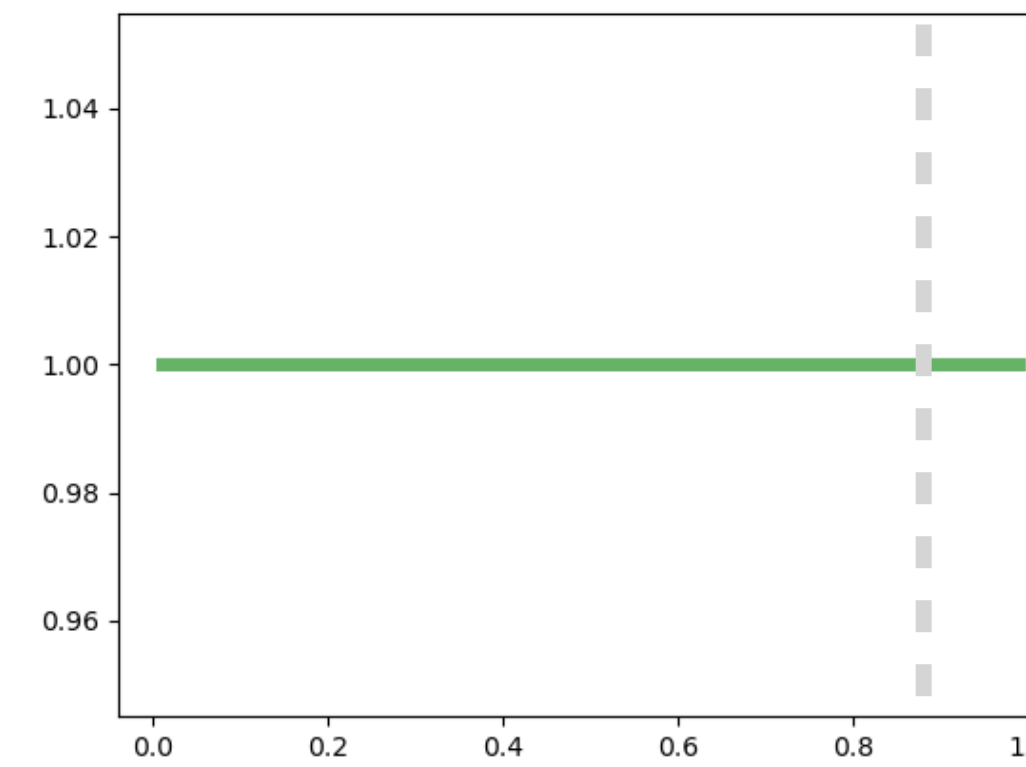
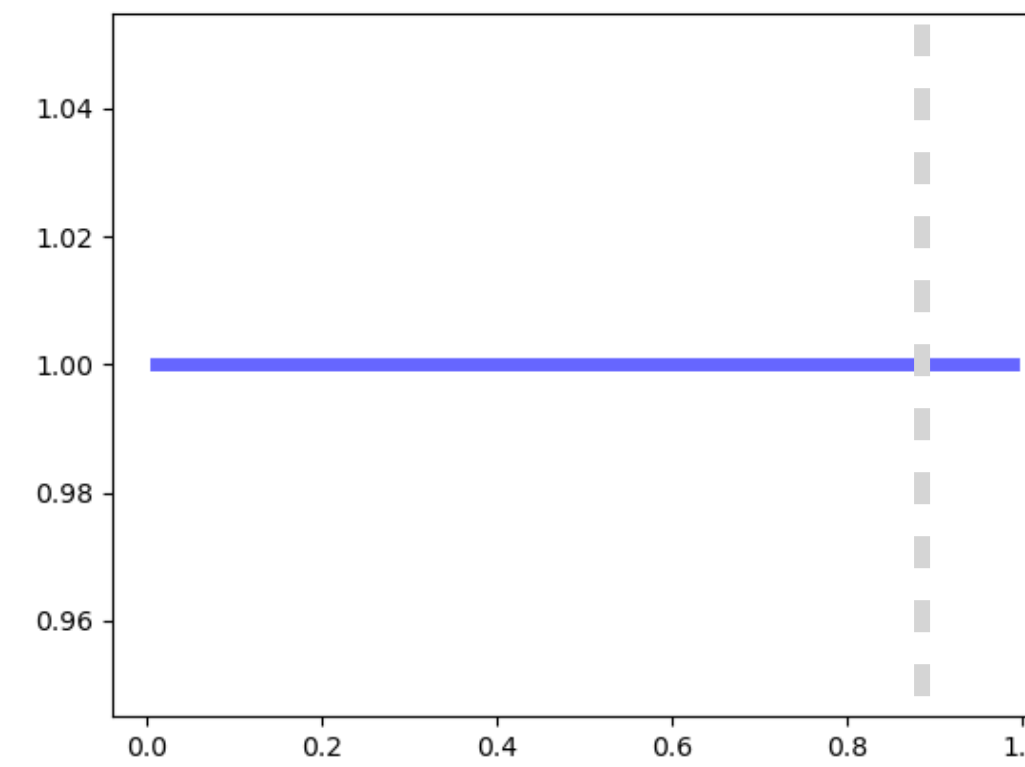
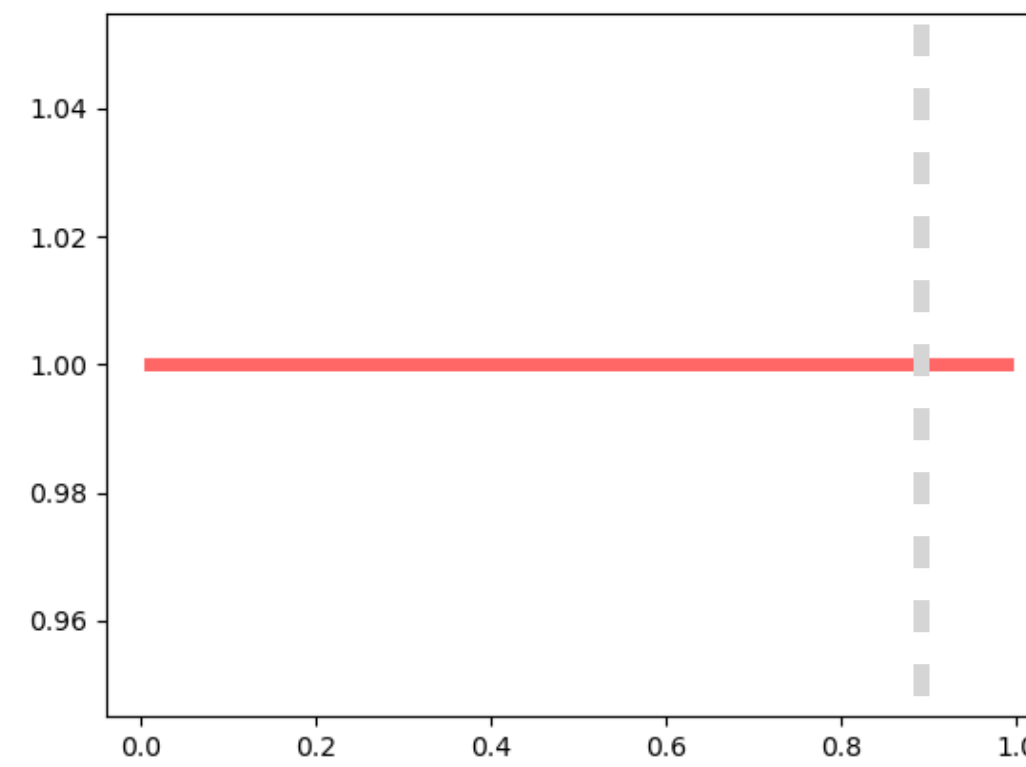


Sketching the Mechanism: Intuition

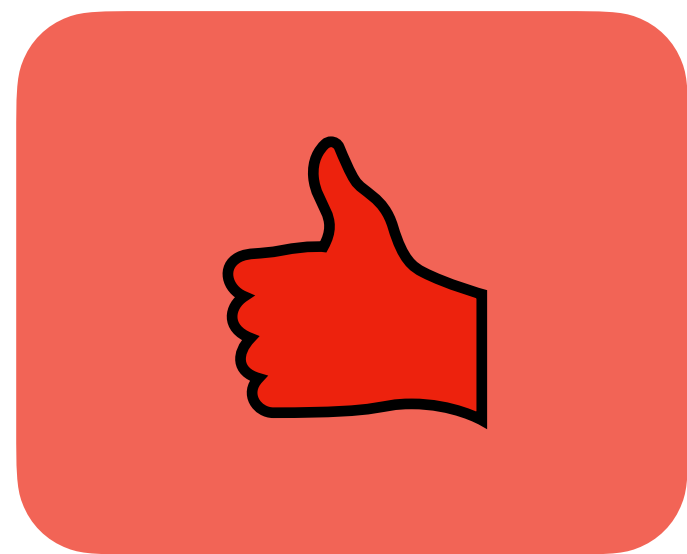
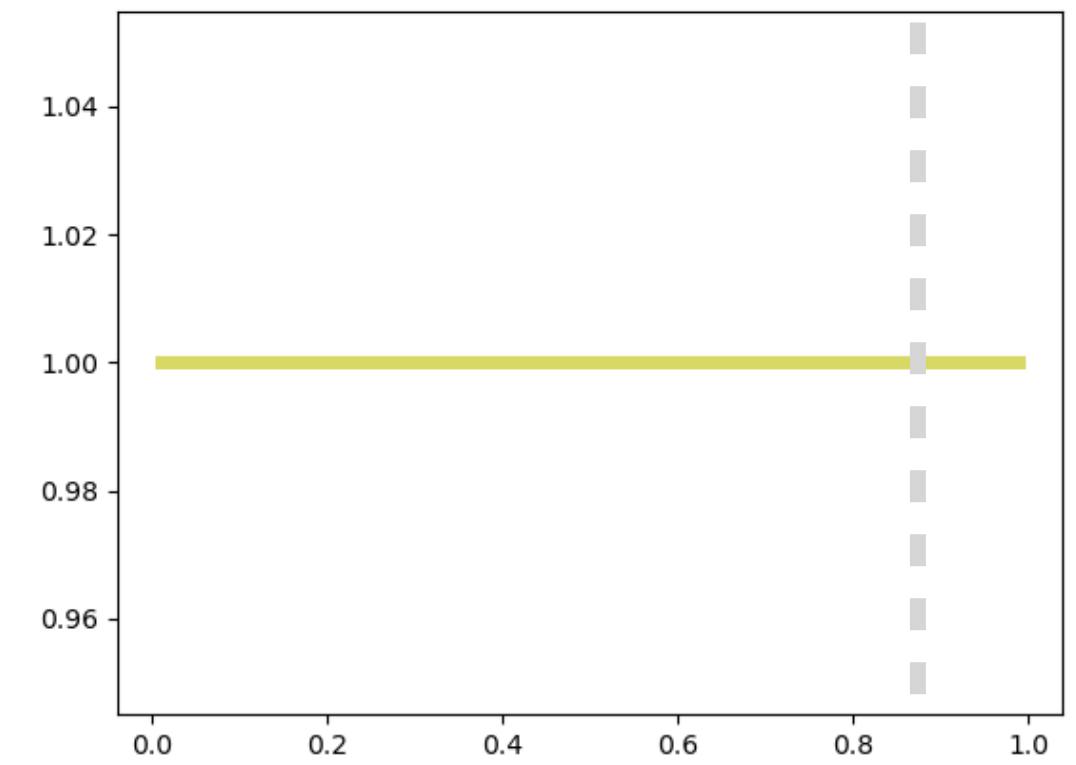
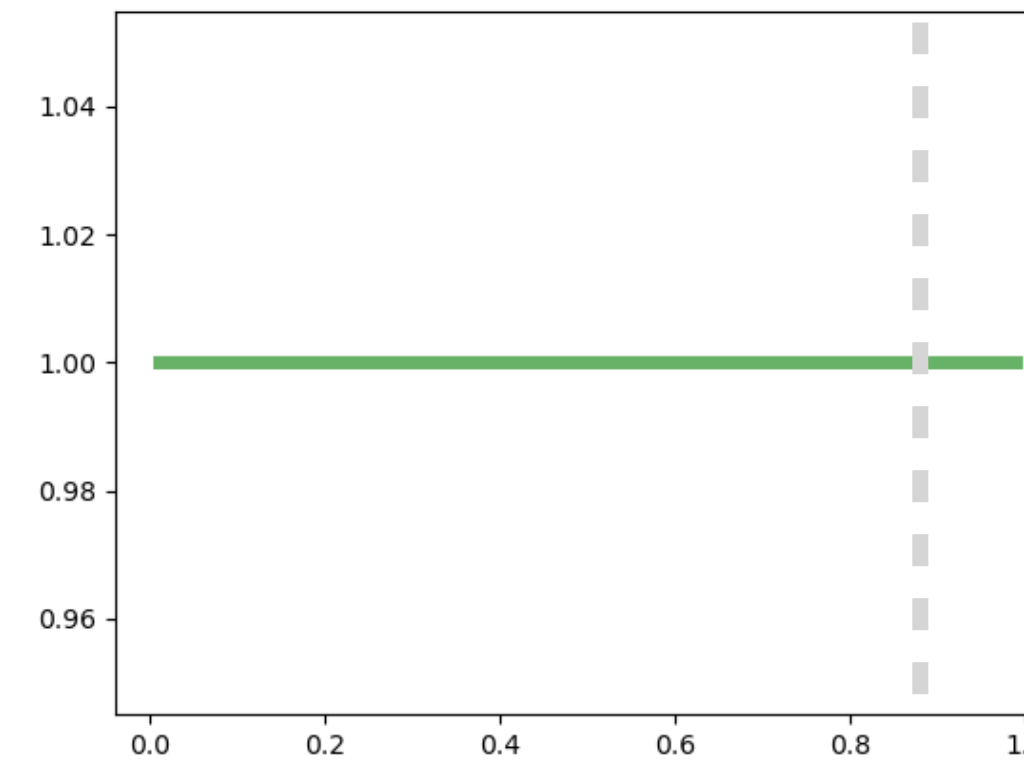
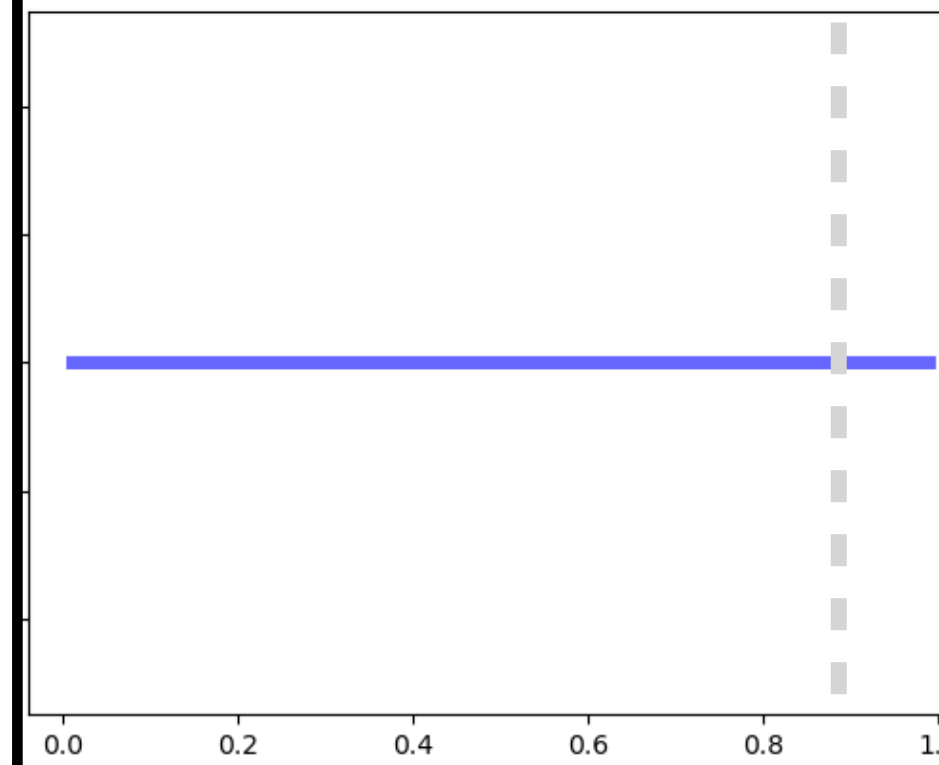
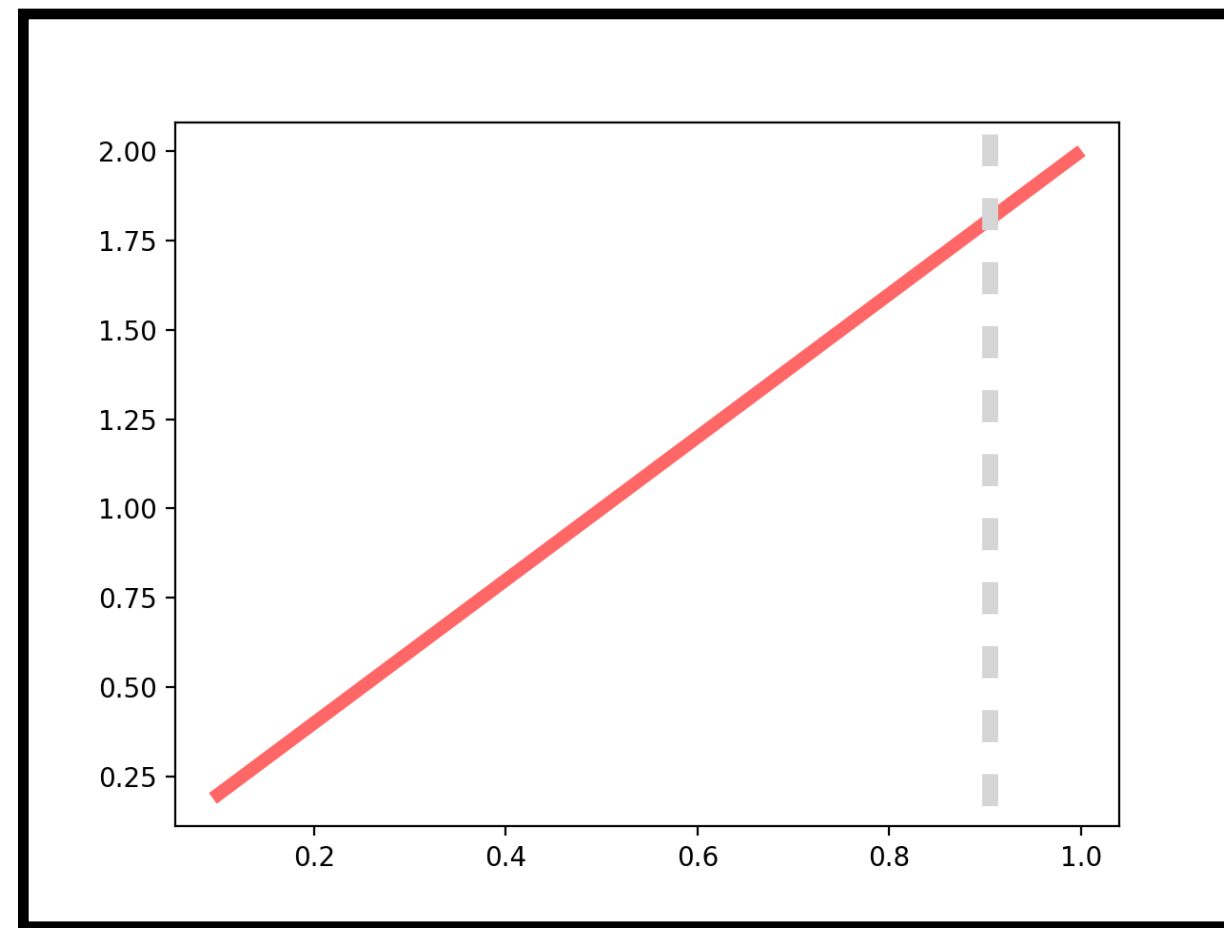


1

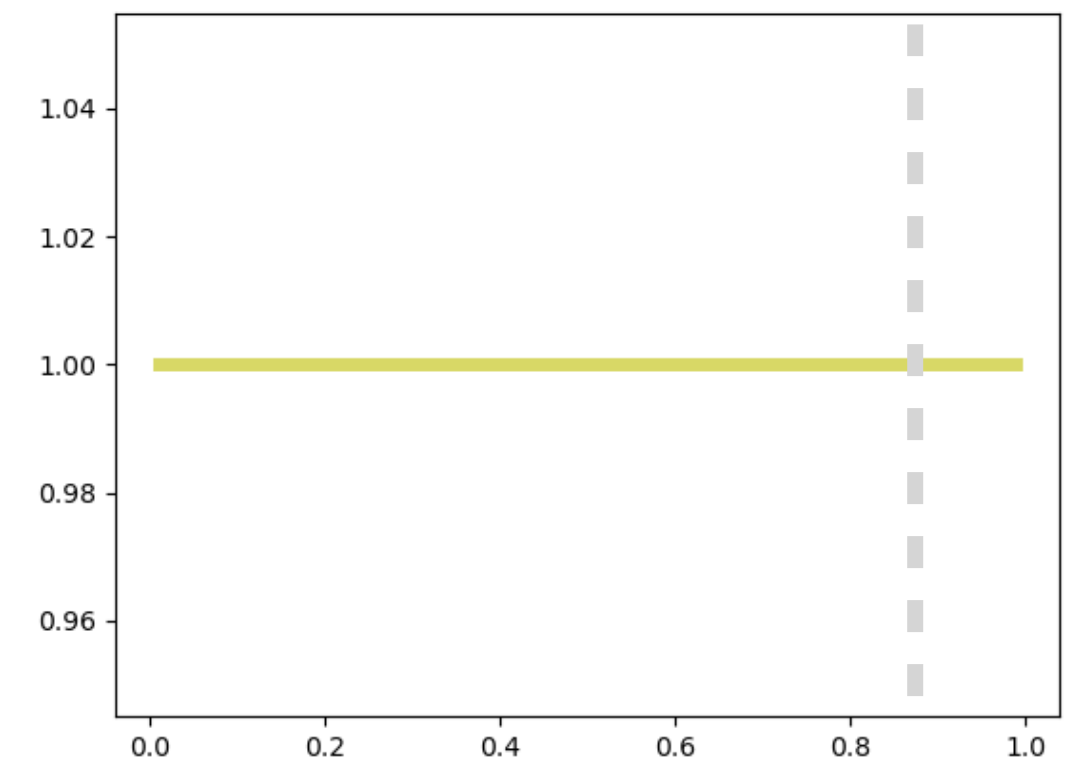
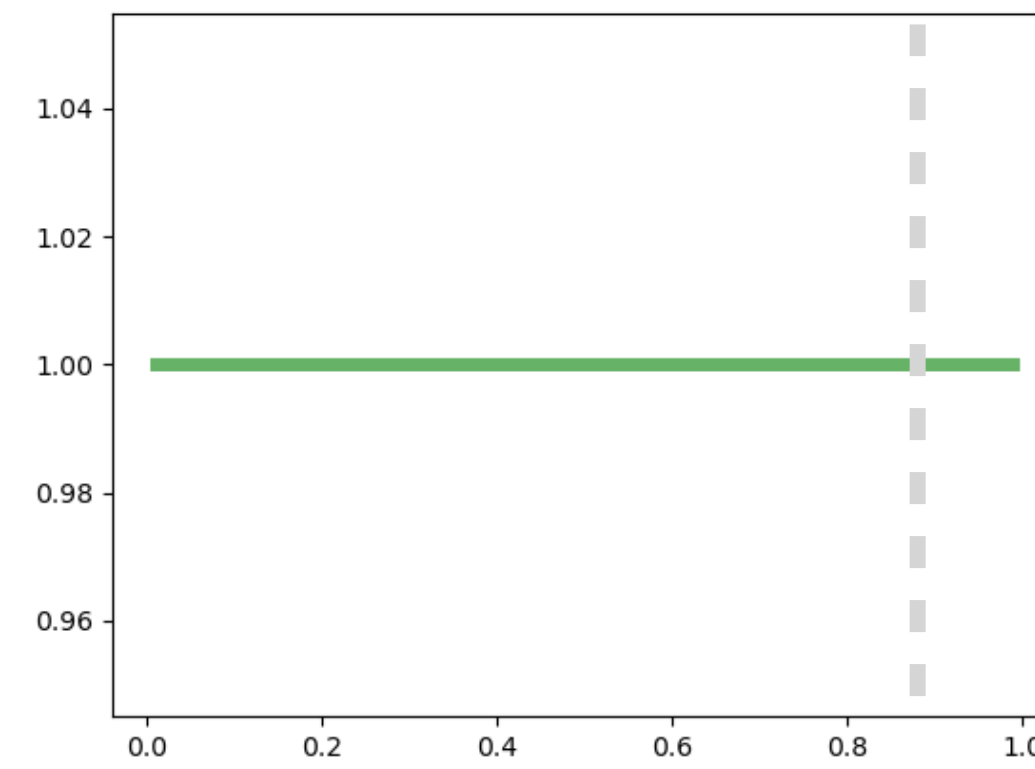
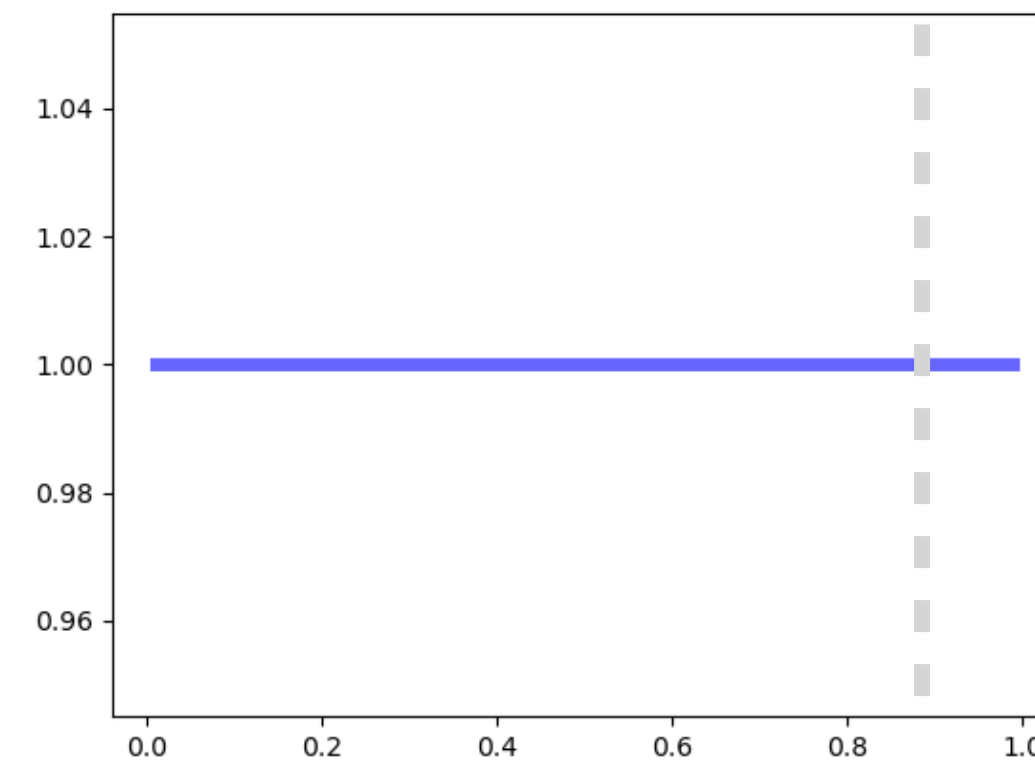
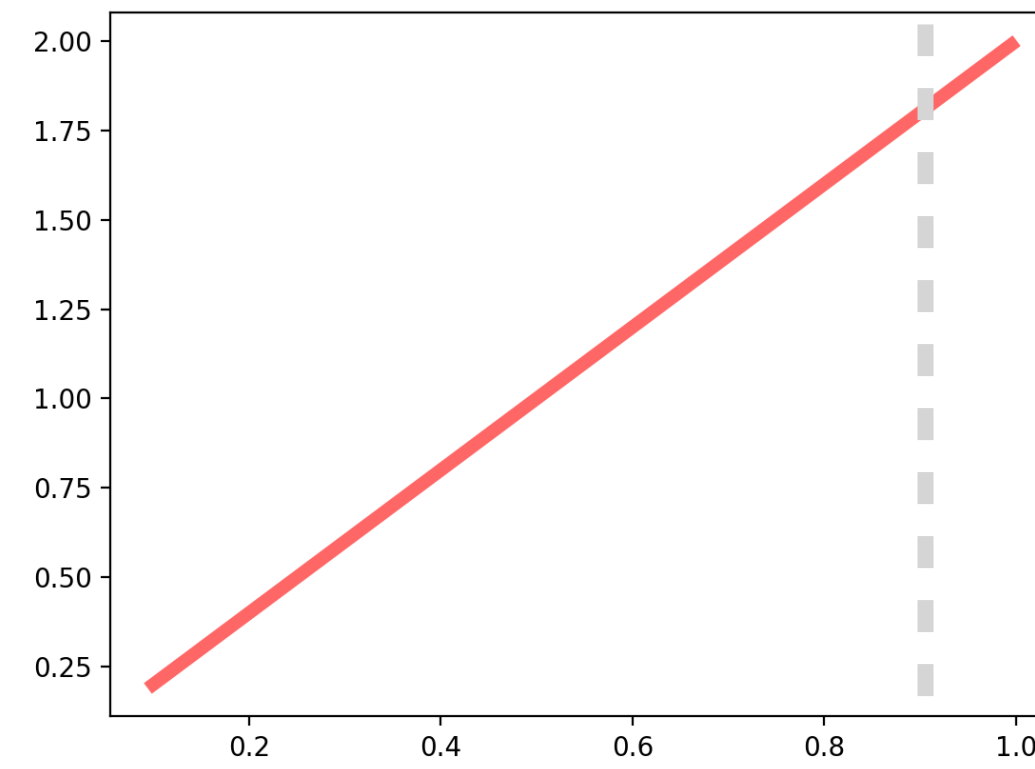
Sketching the Mechanism: Intuition



Sketching the Mechanism: Intuition



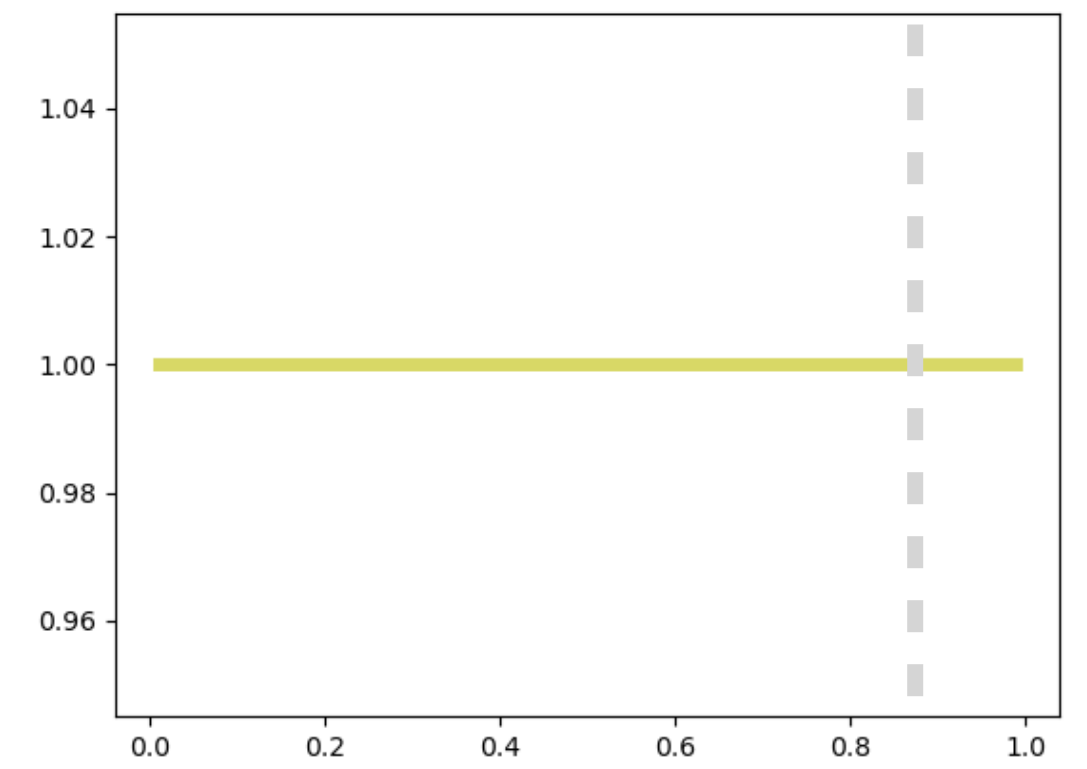
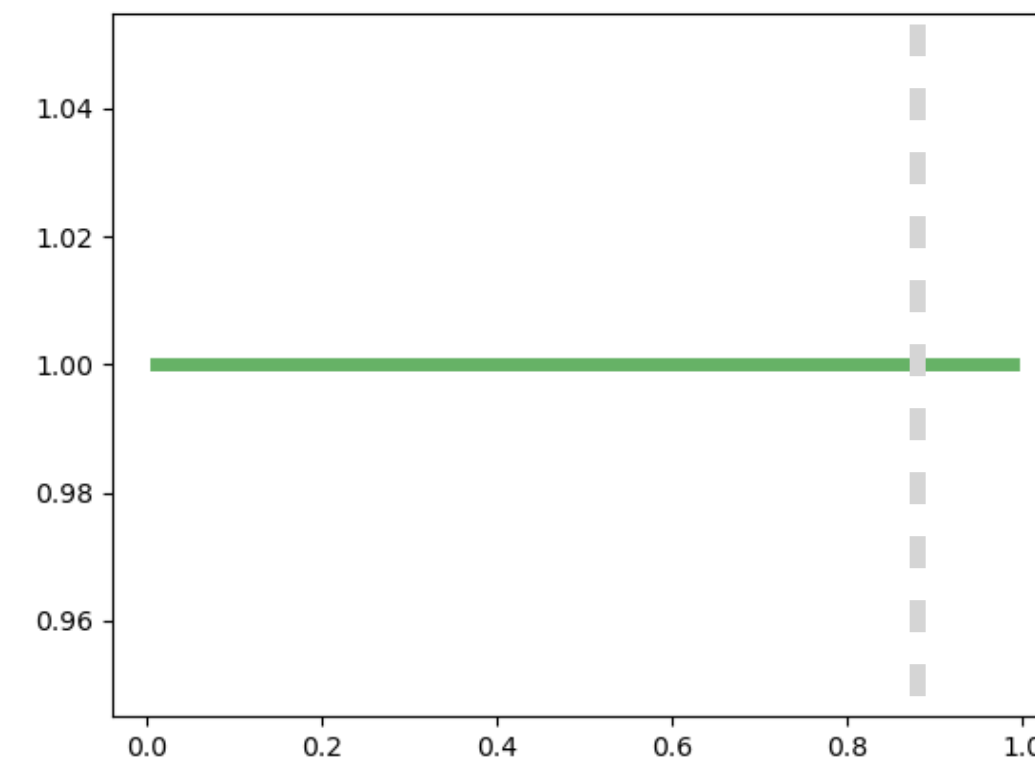
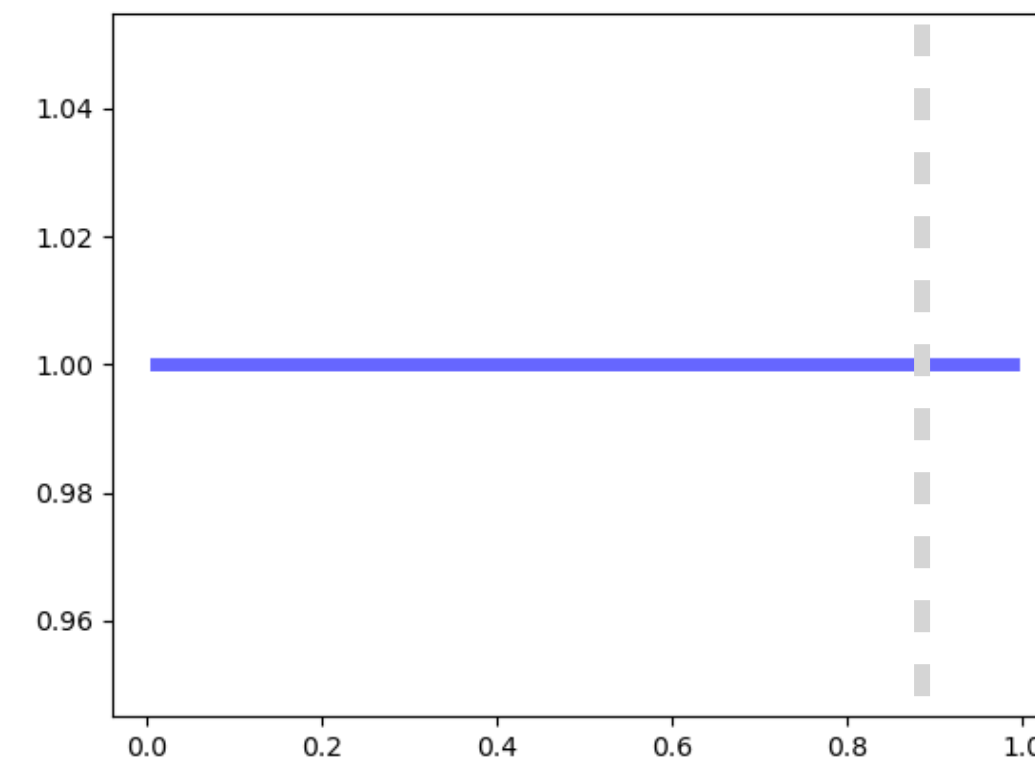
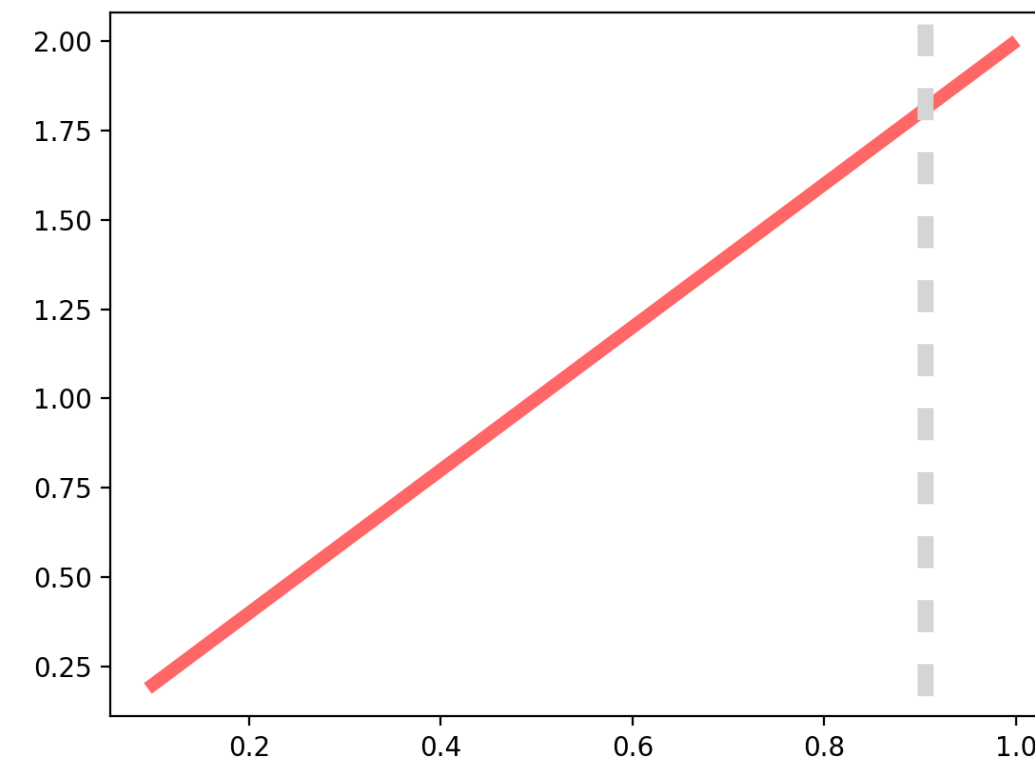
Sketching the Mechanism: Intuition



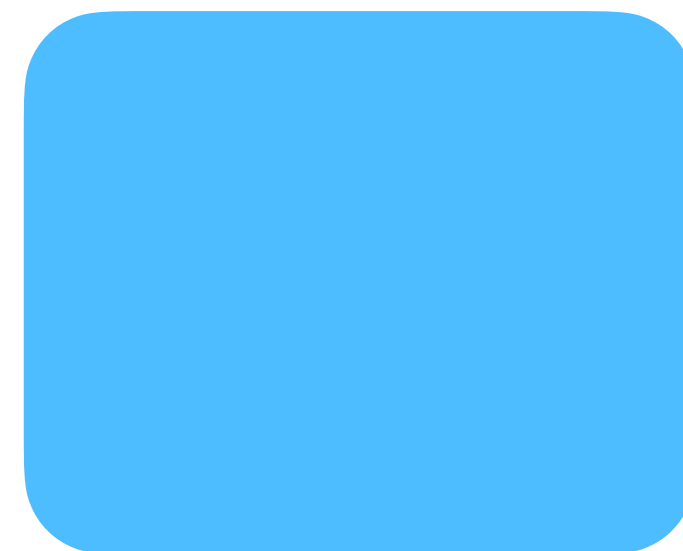
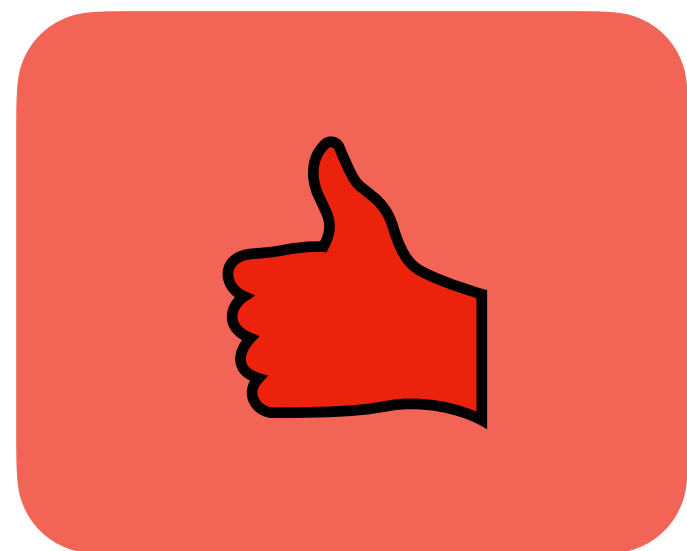
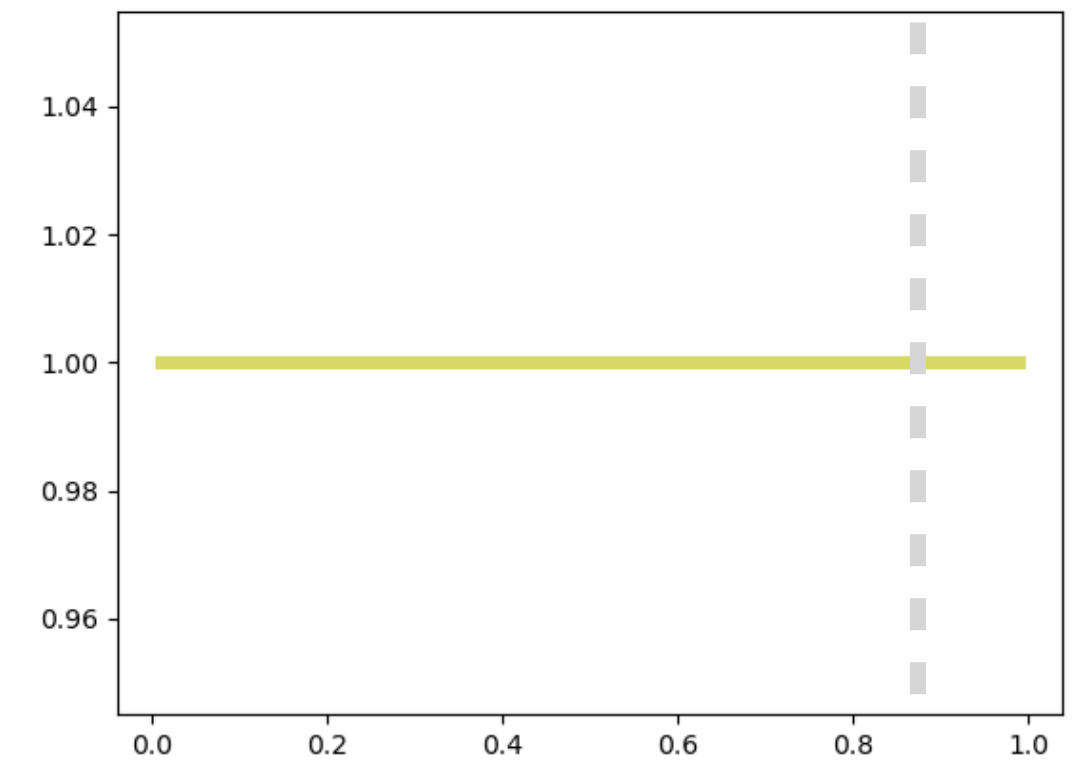
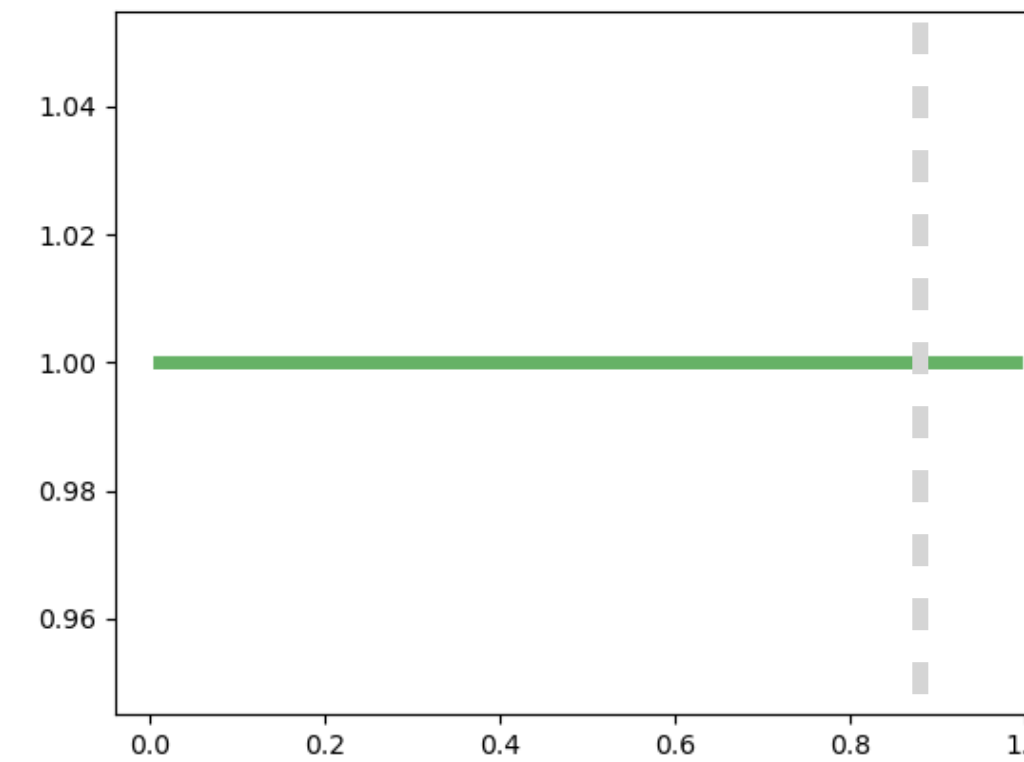
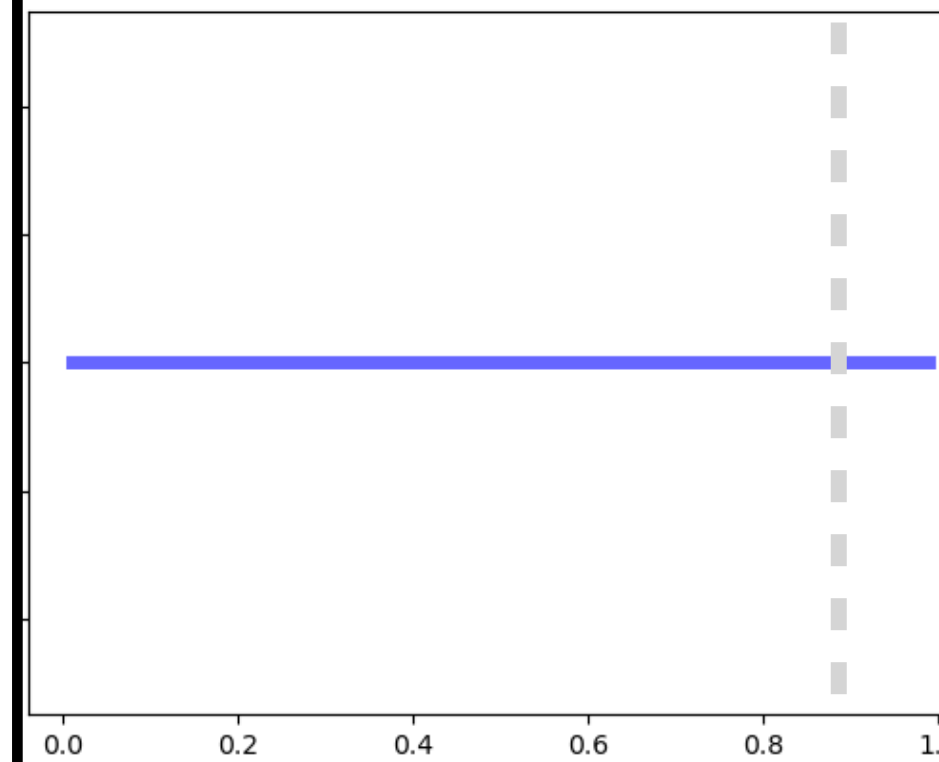
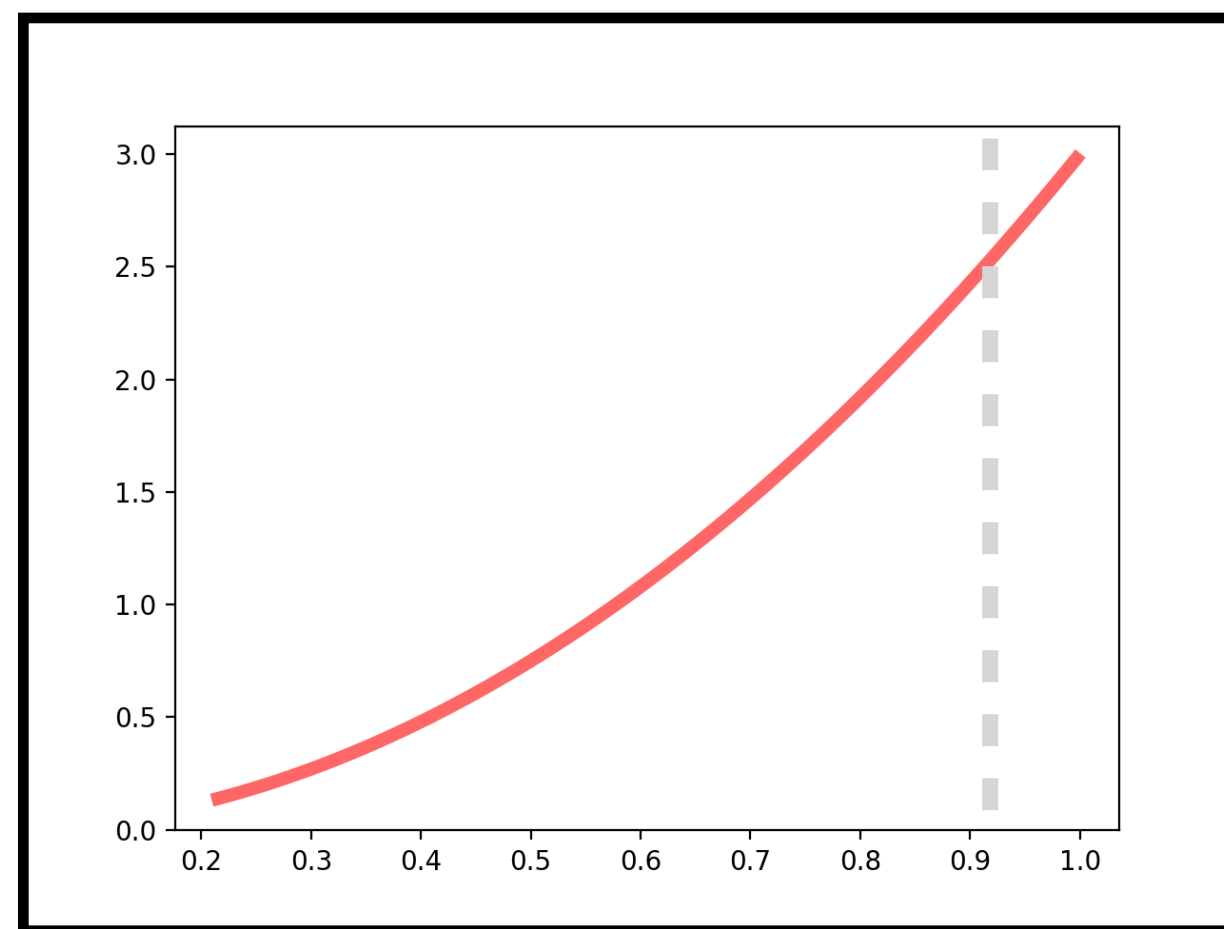
2



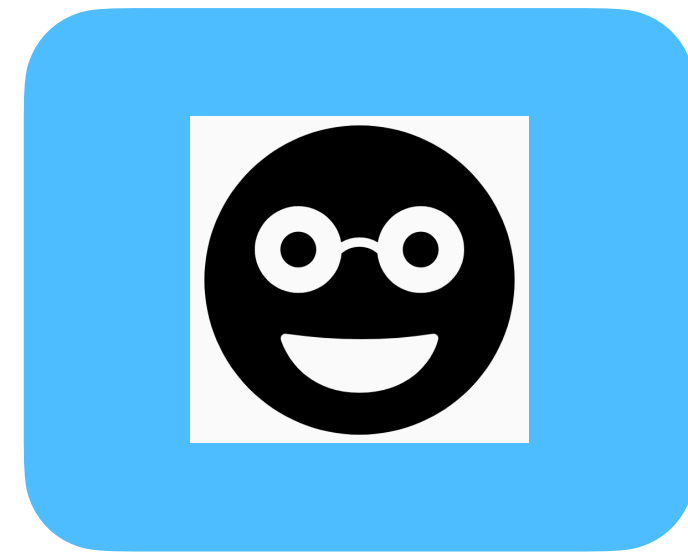
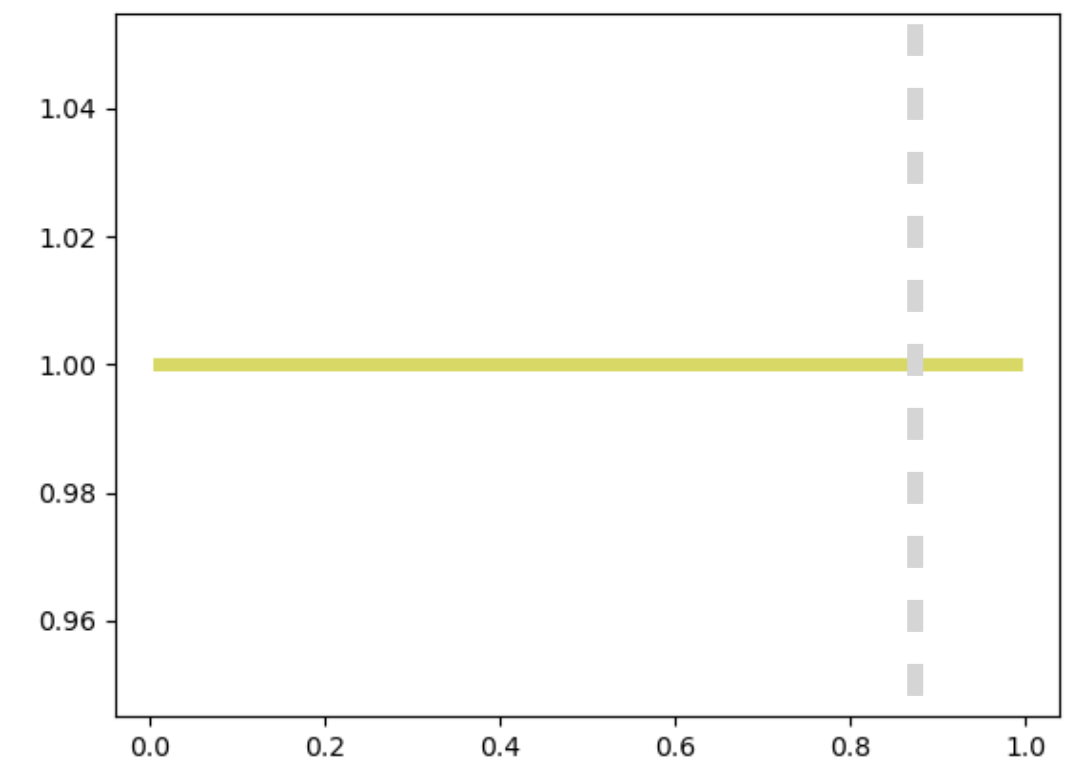
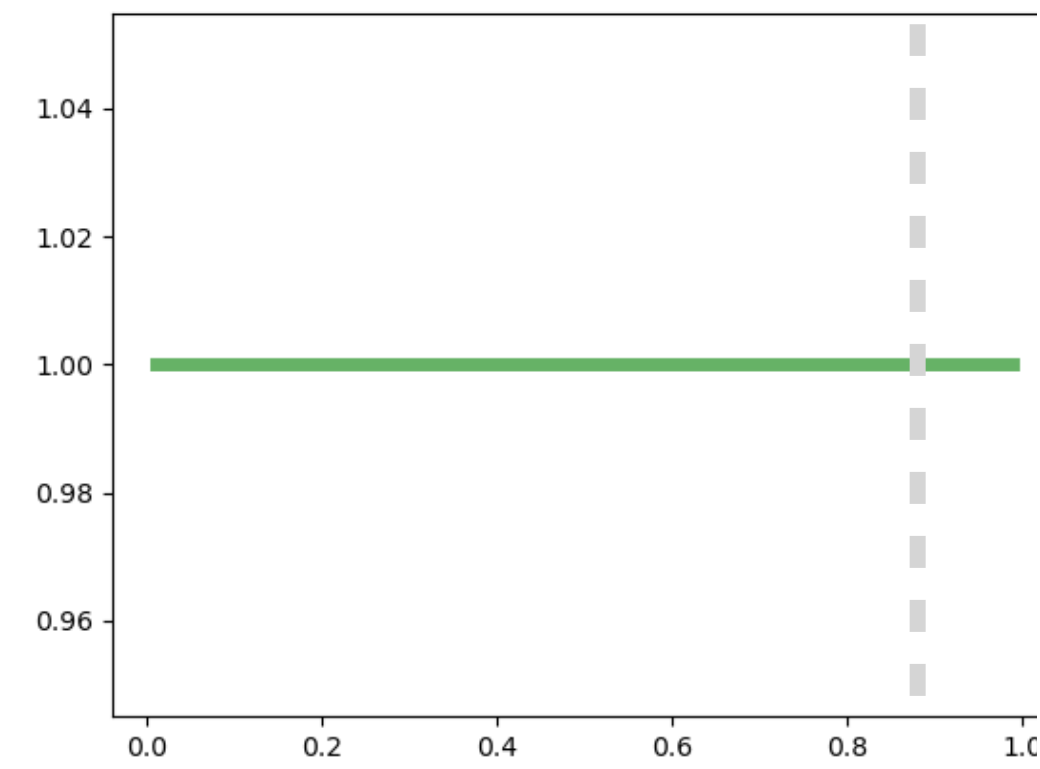
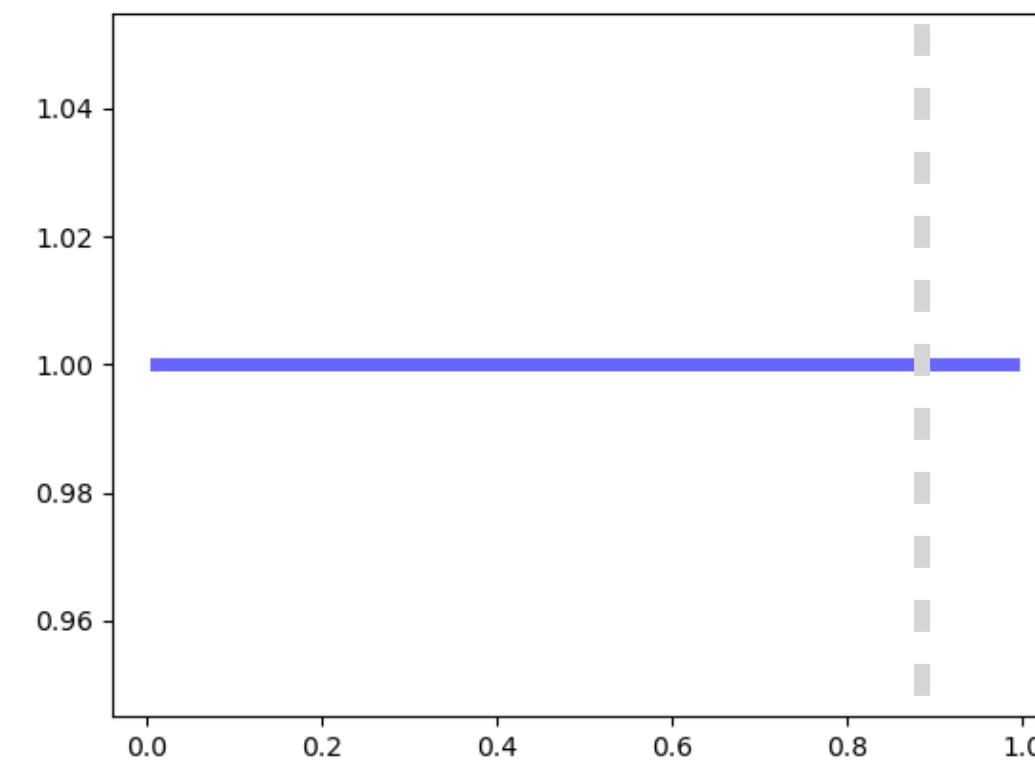
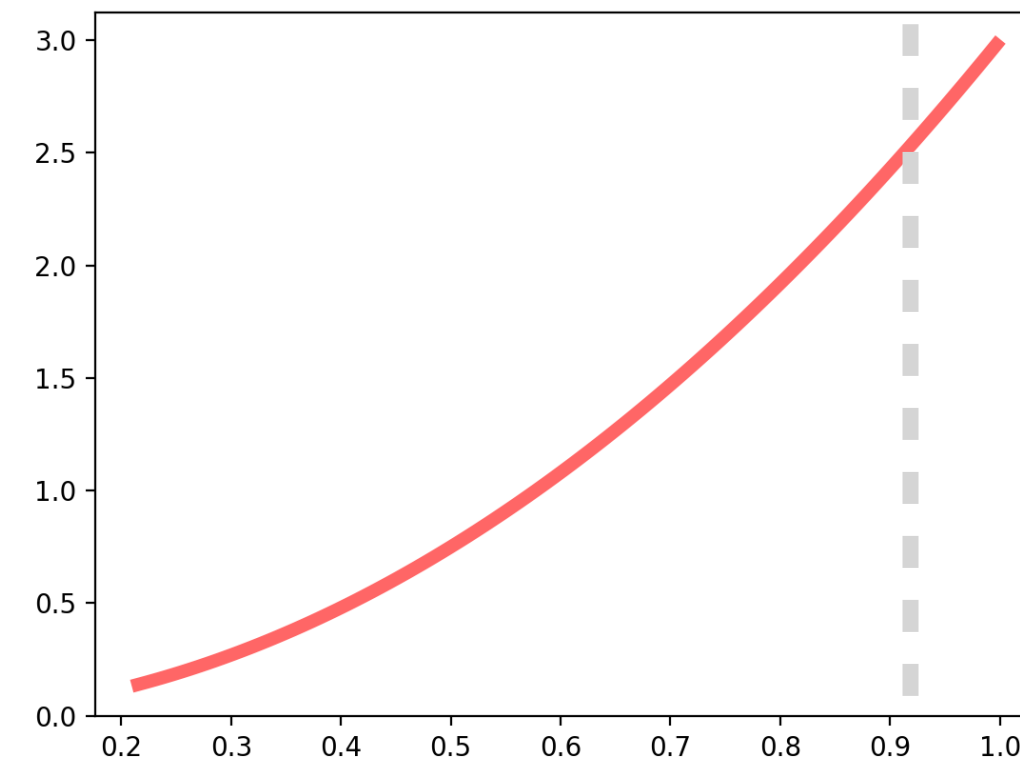
Sketching the Mechanism: Intuition



Sketching the Mechanism: Intuition



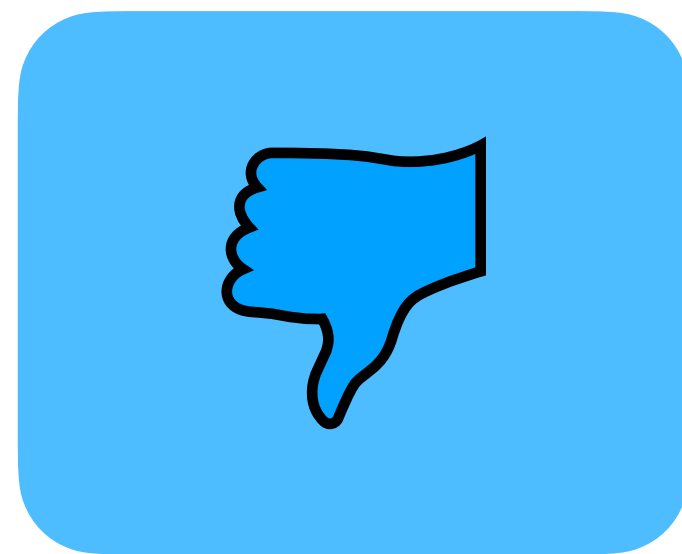
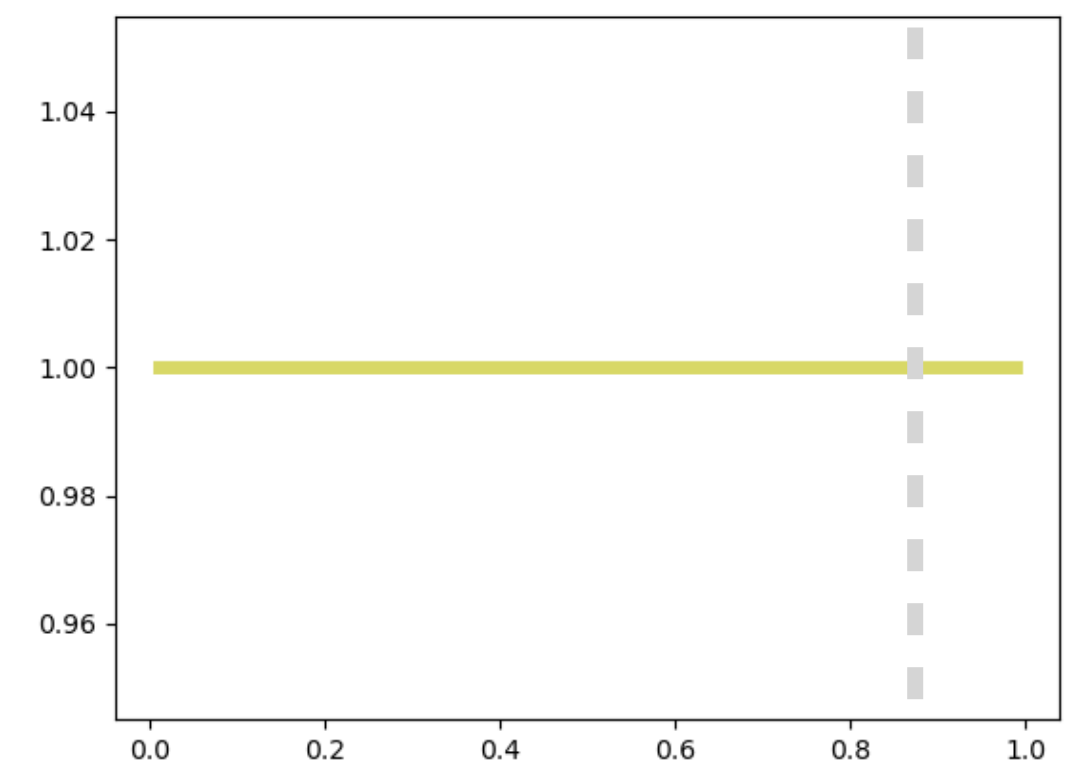
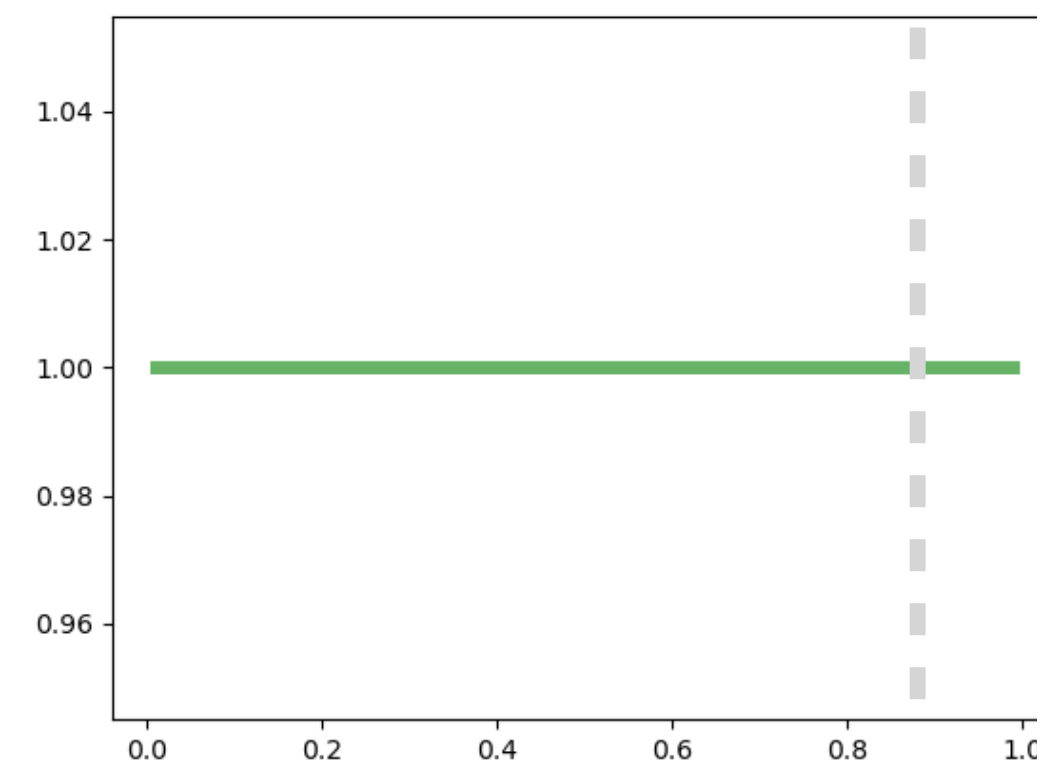
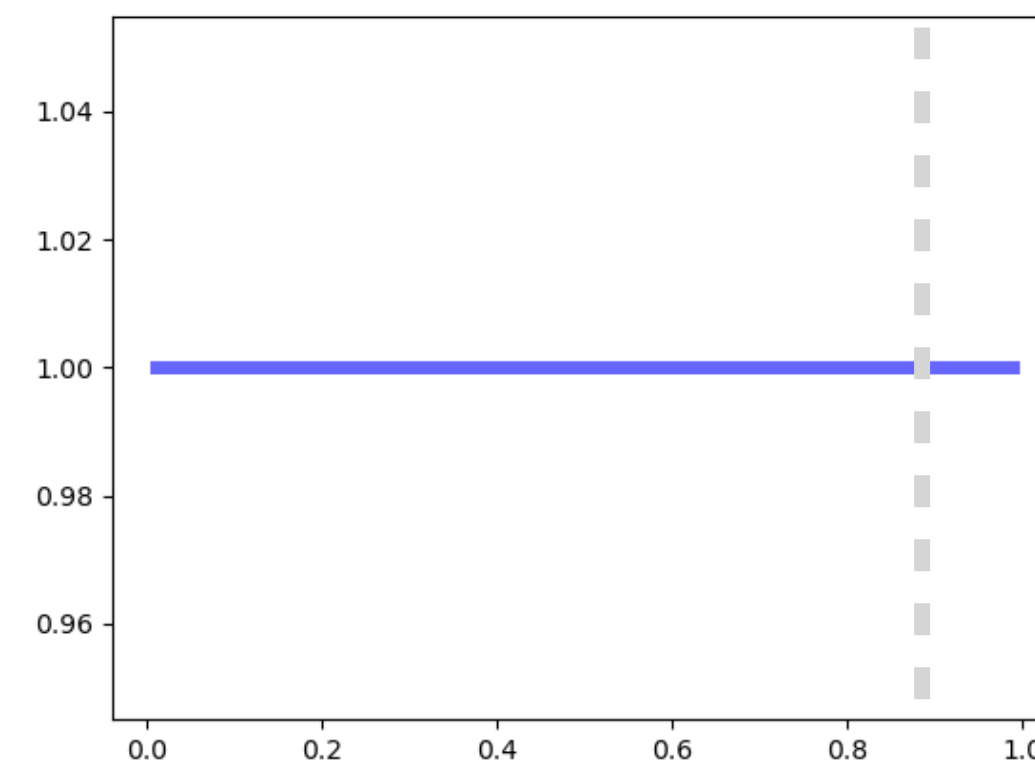
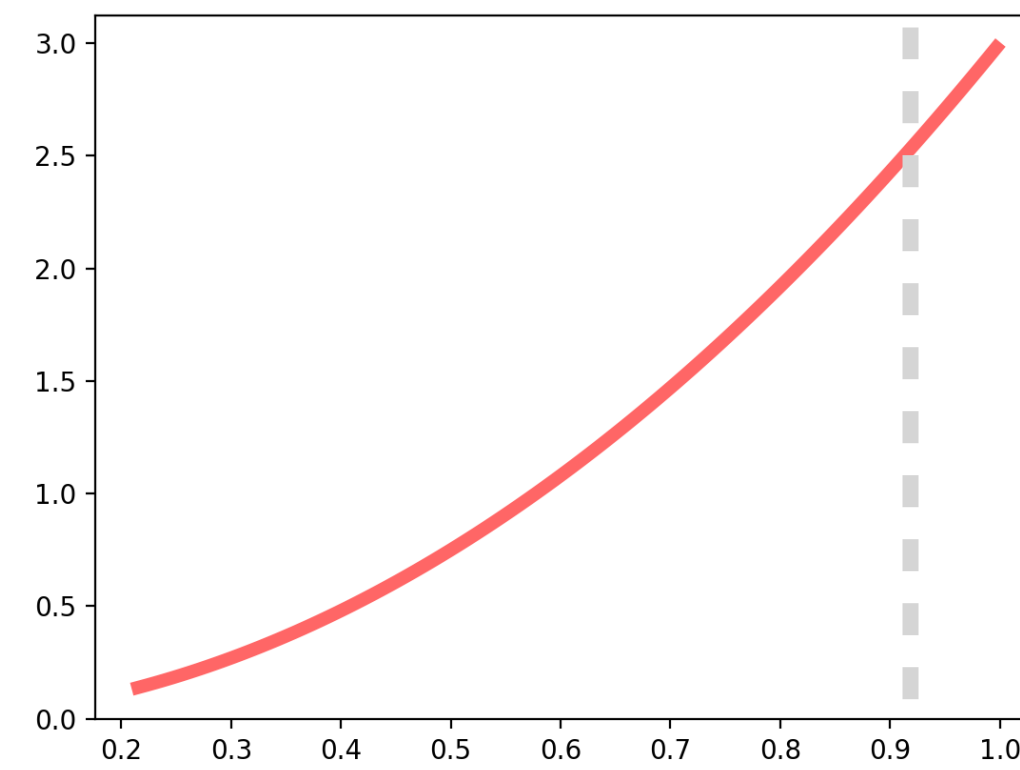
Sketching the Mechanism: Intuition



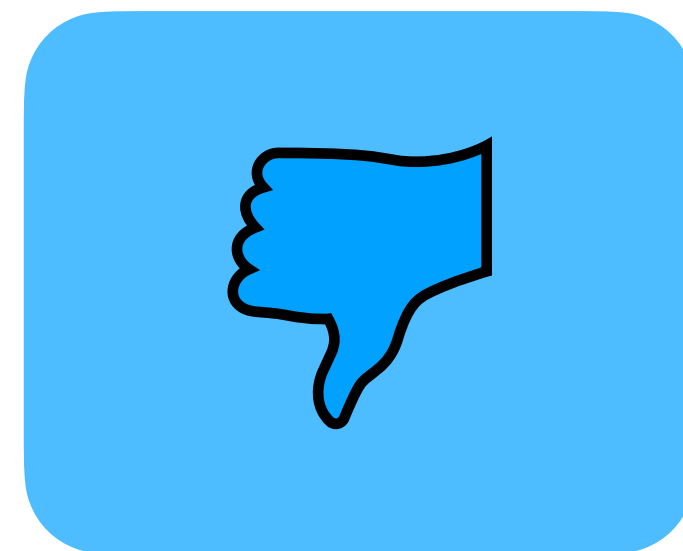
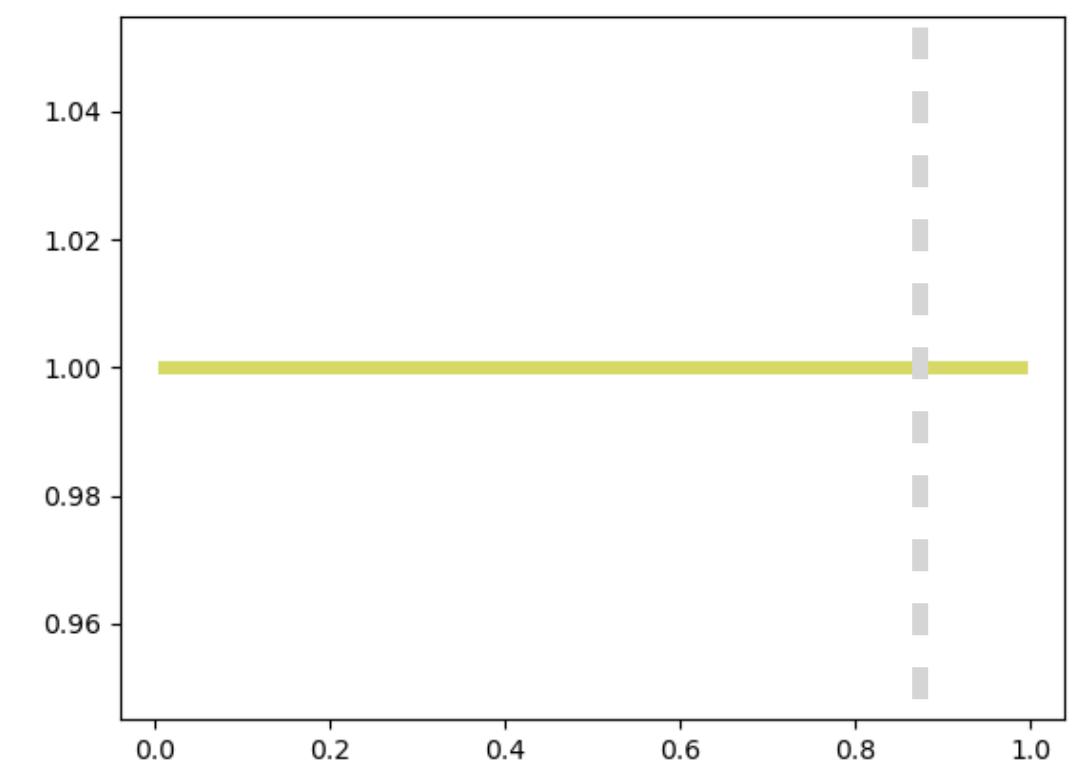
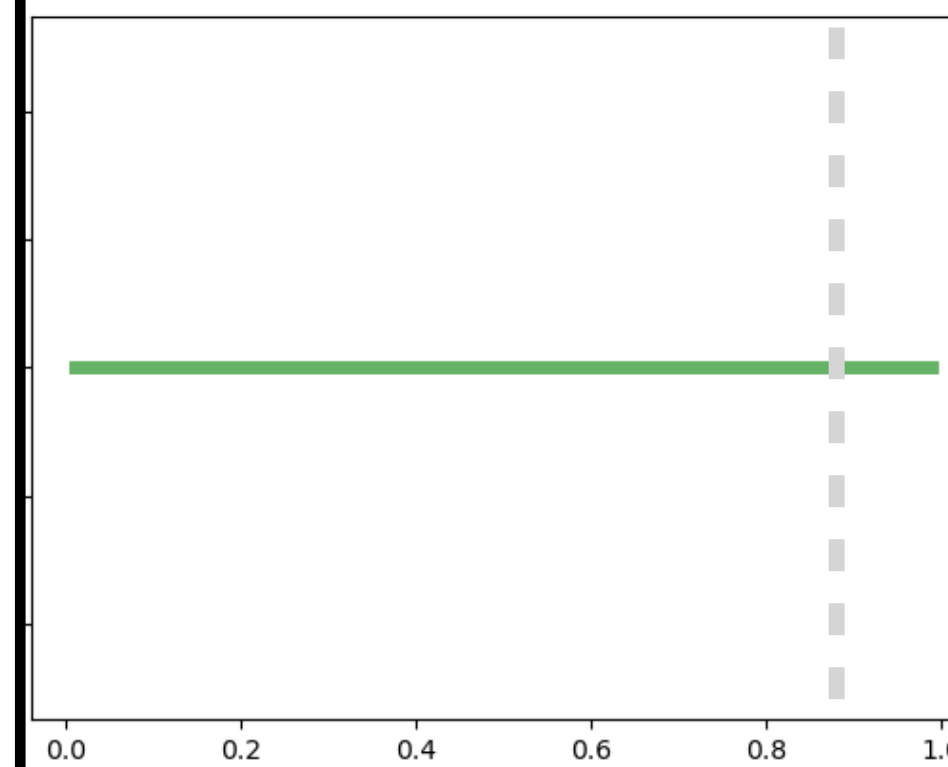
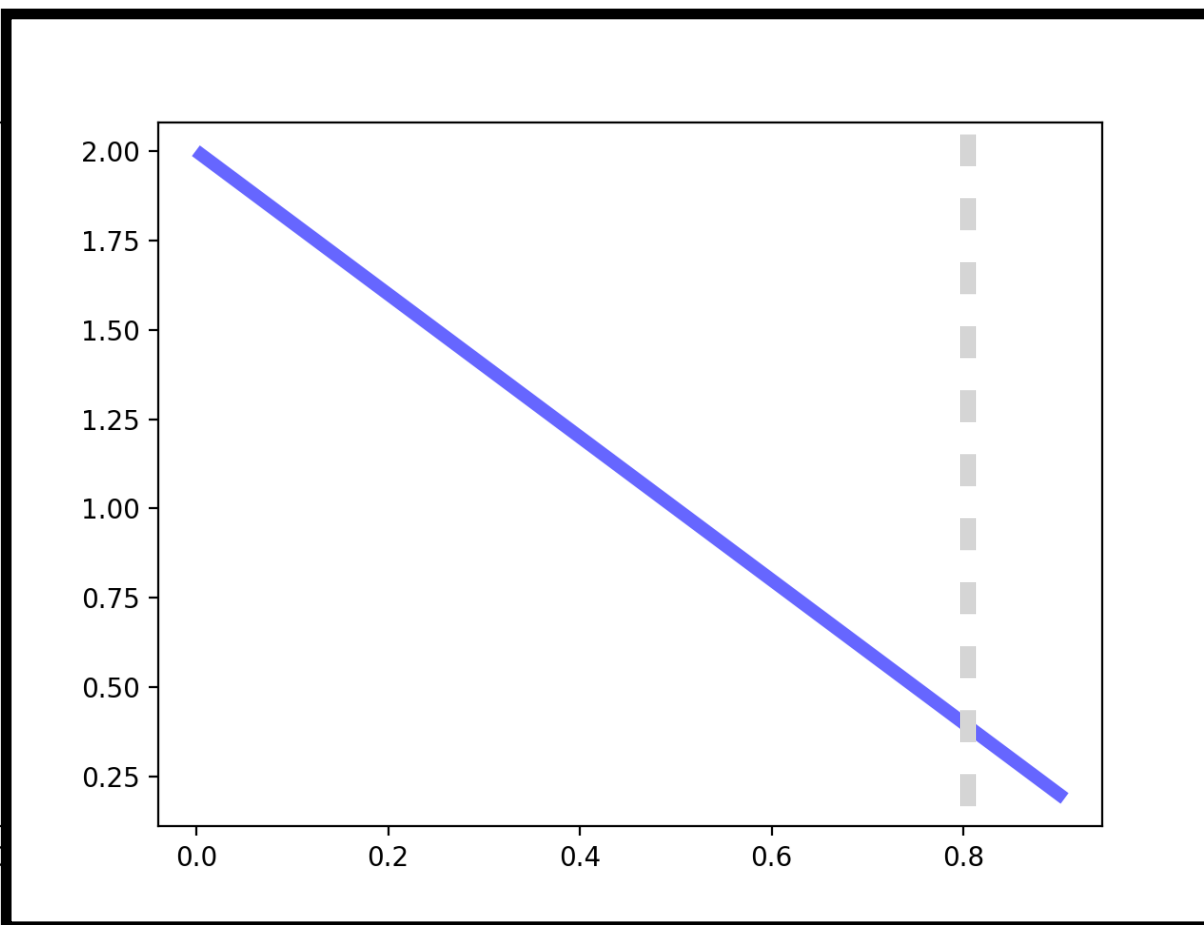
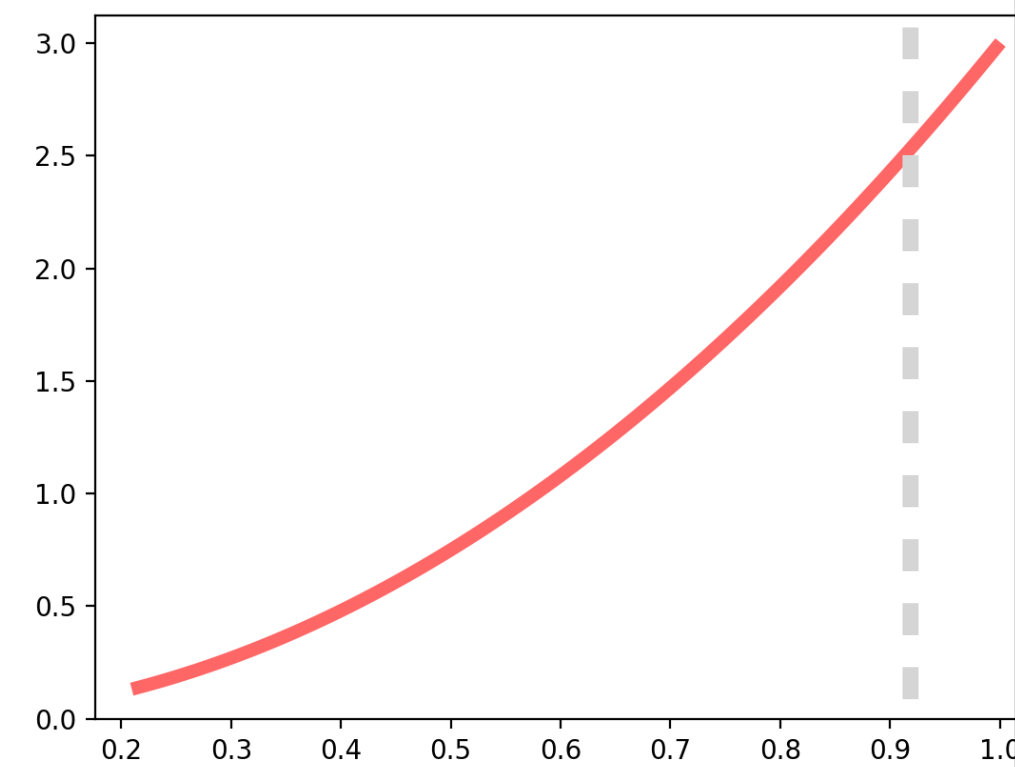
1



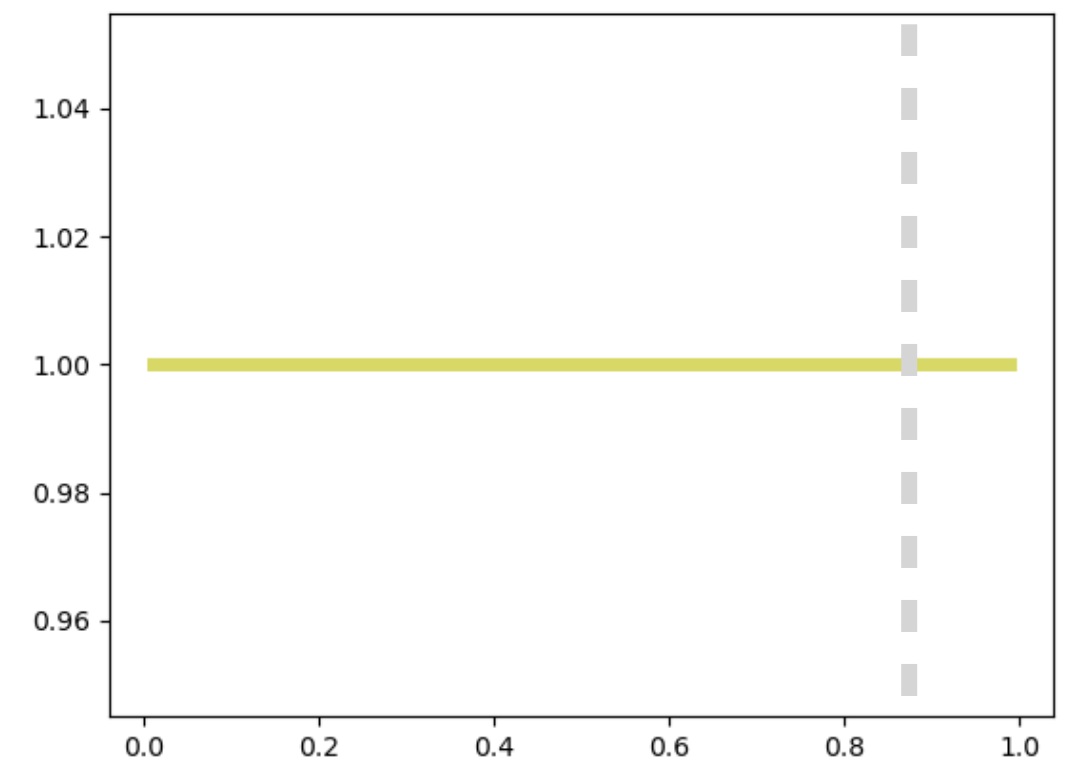
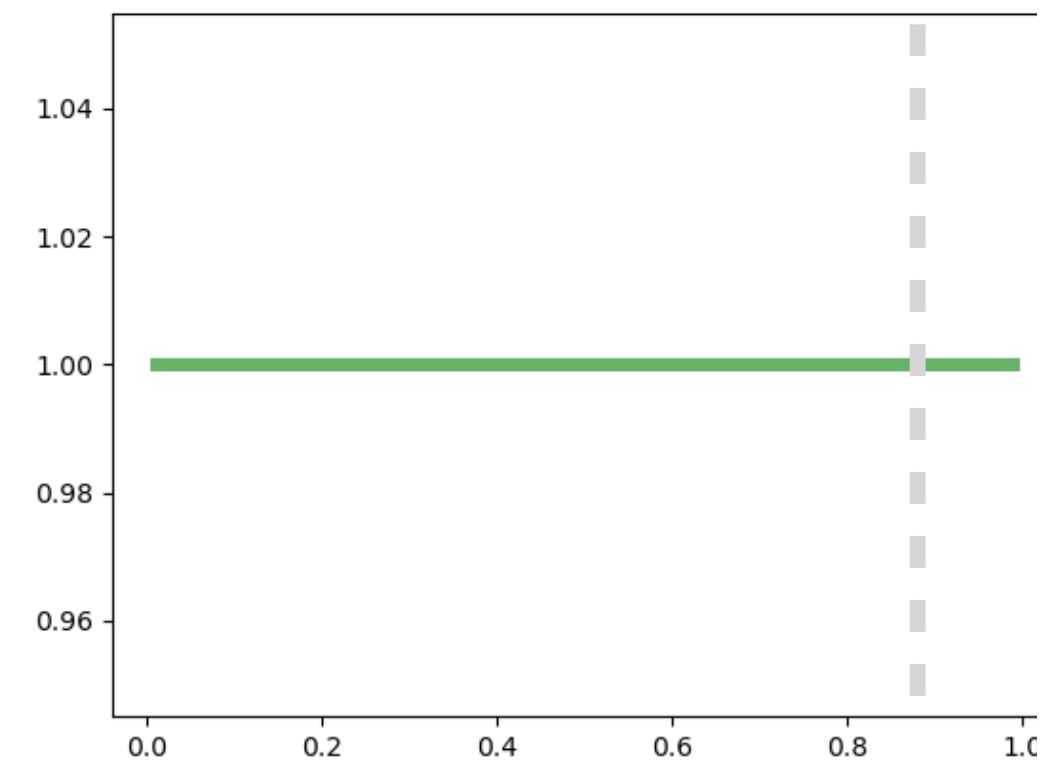
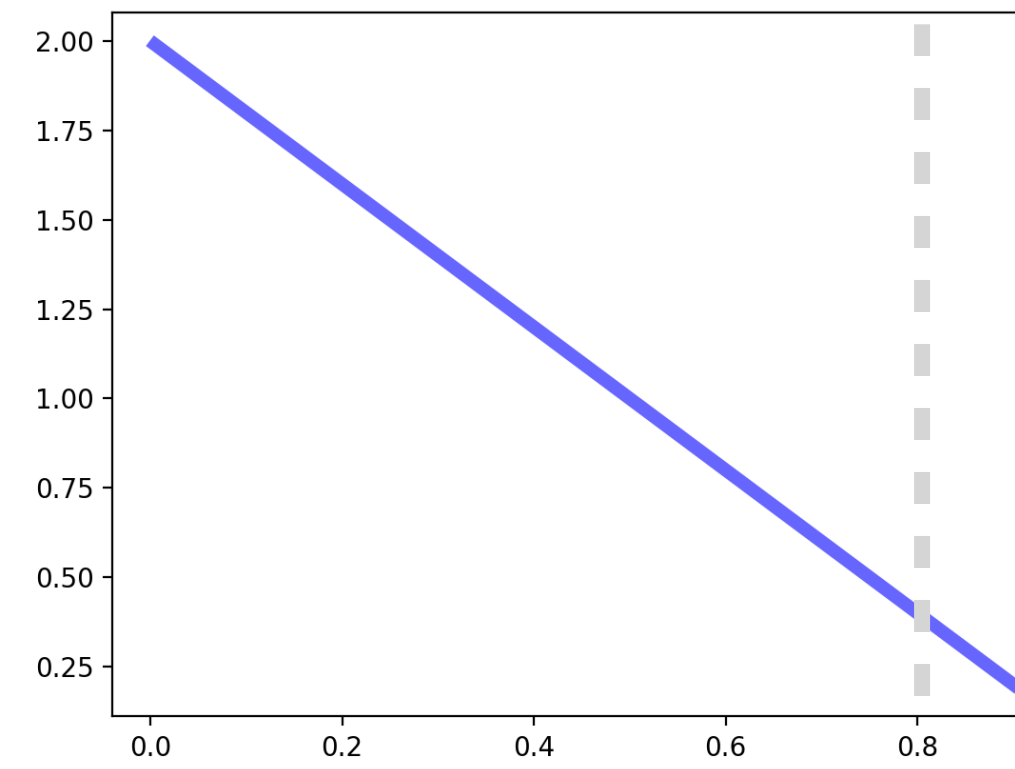
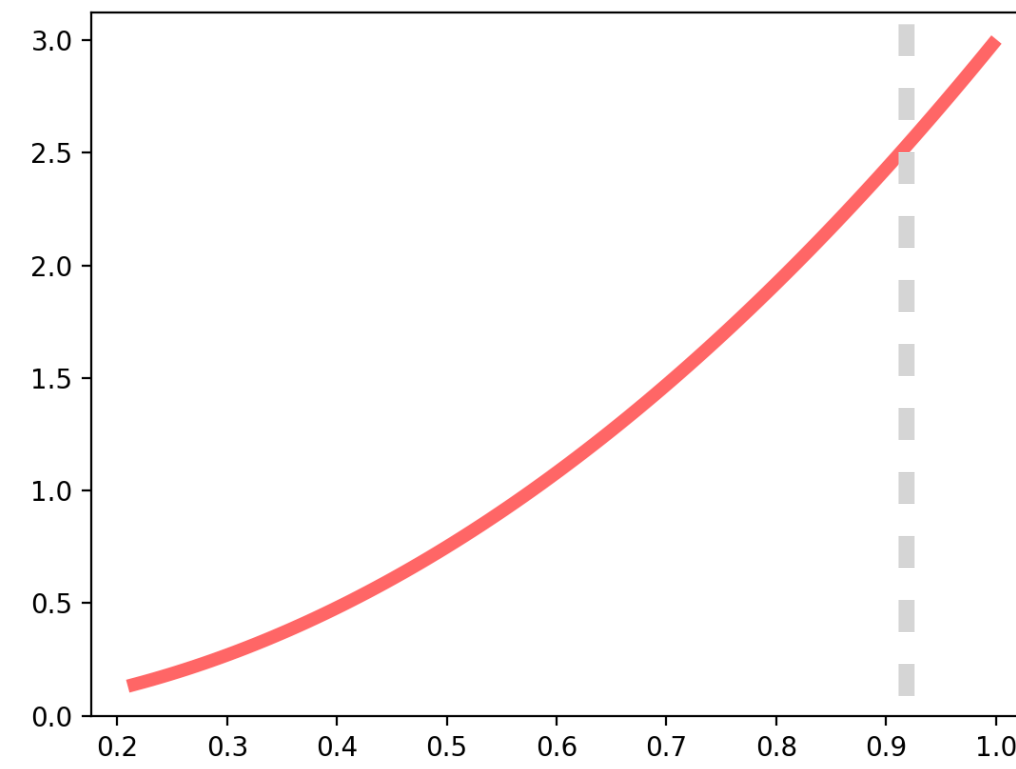
Sketching the Mechanism: Intuition



Sketching the Mechanism: Intuition

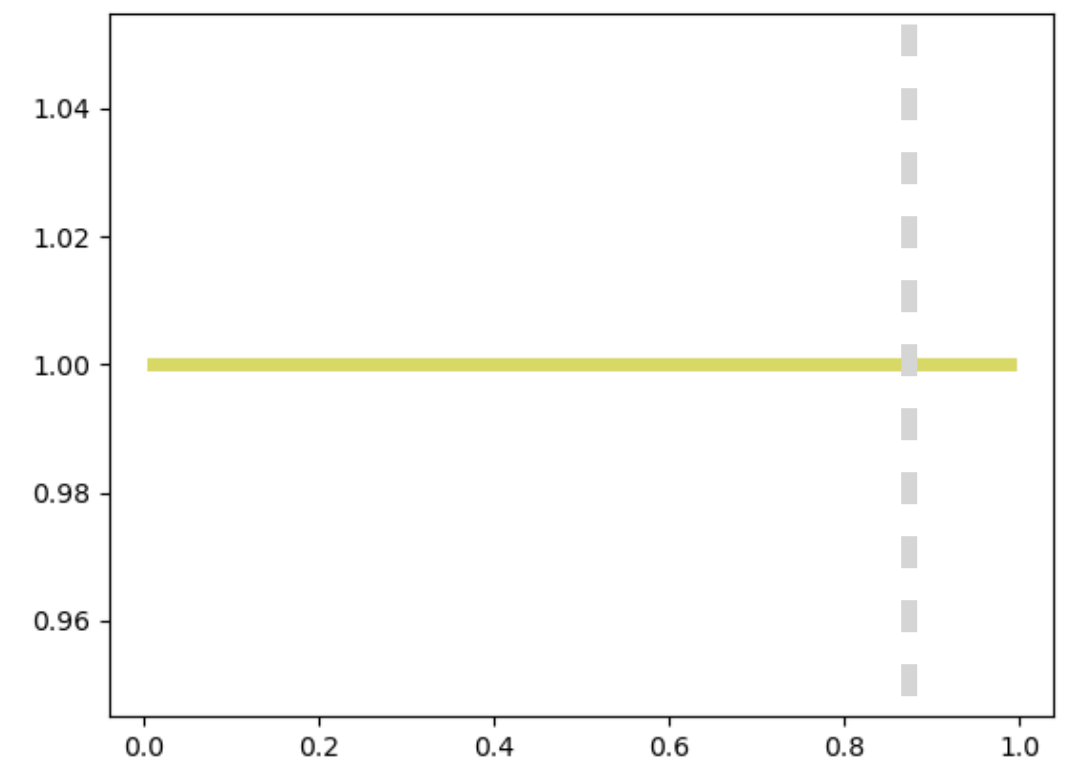
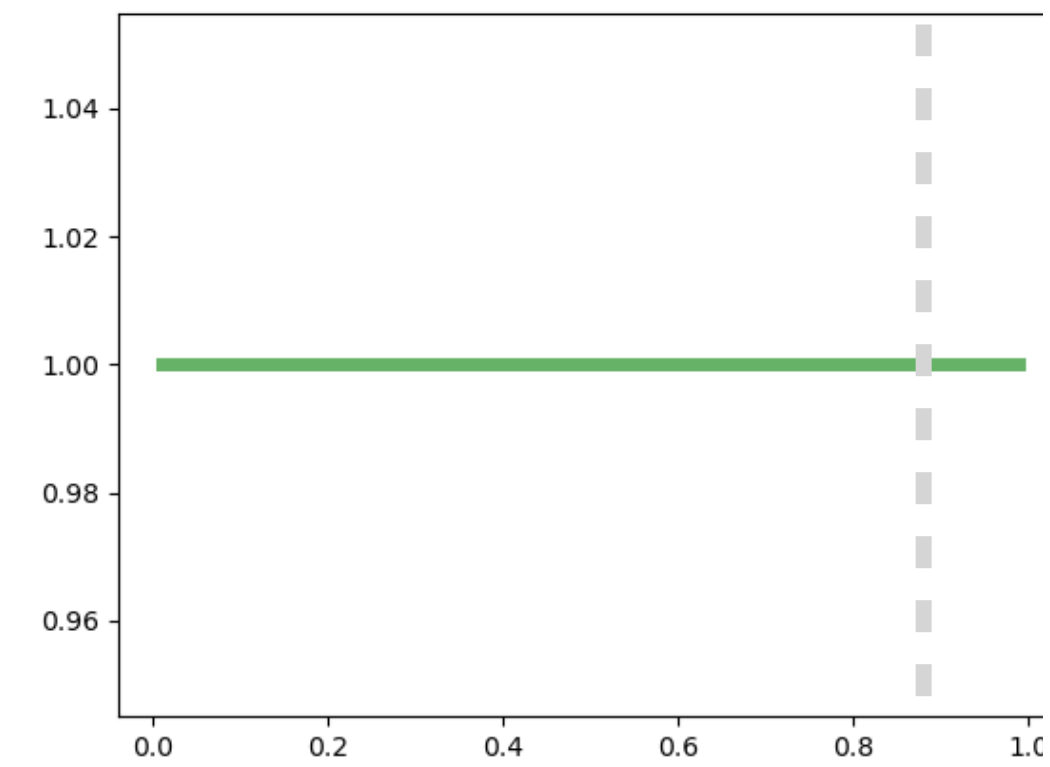
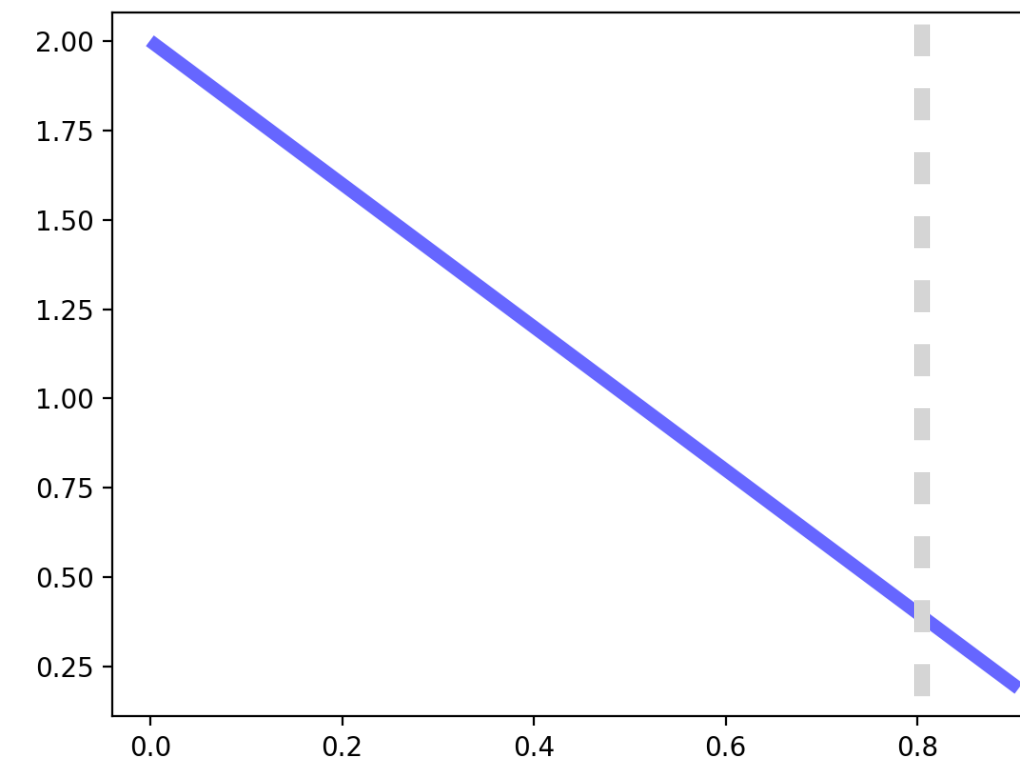
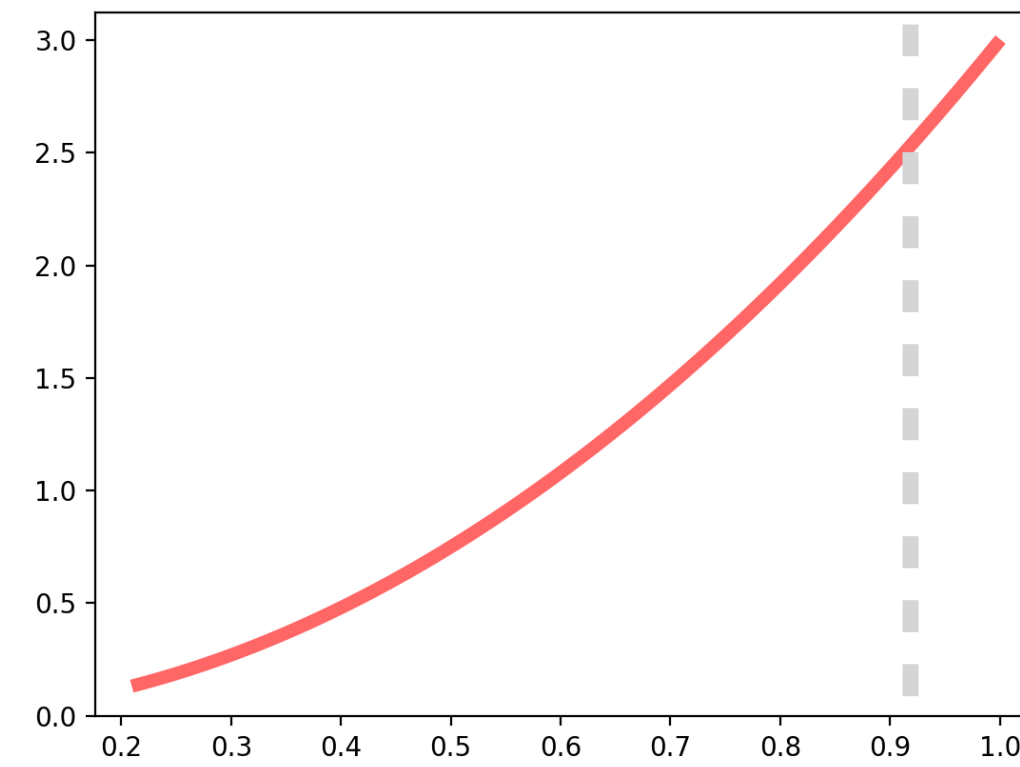


Sketching the Mechanism: Intuition

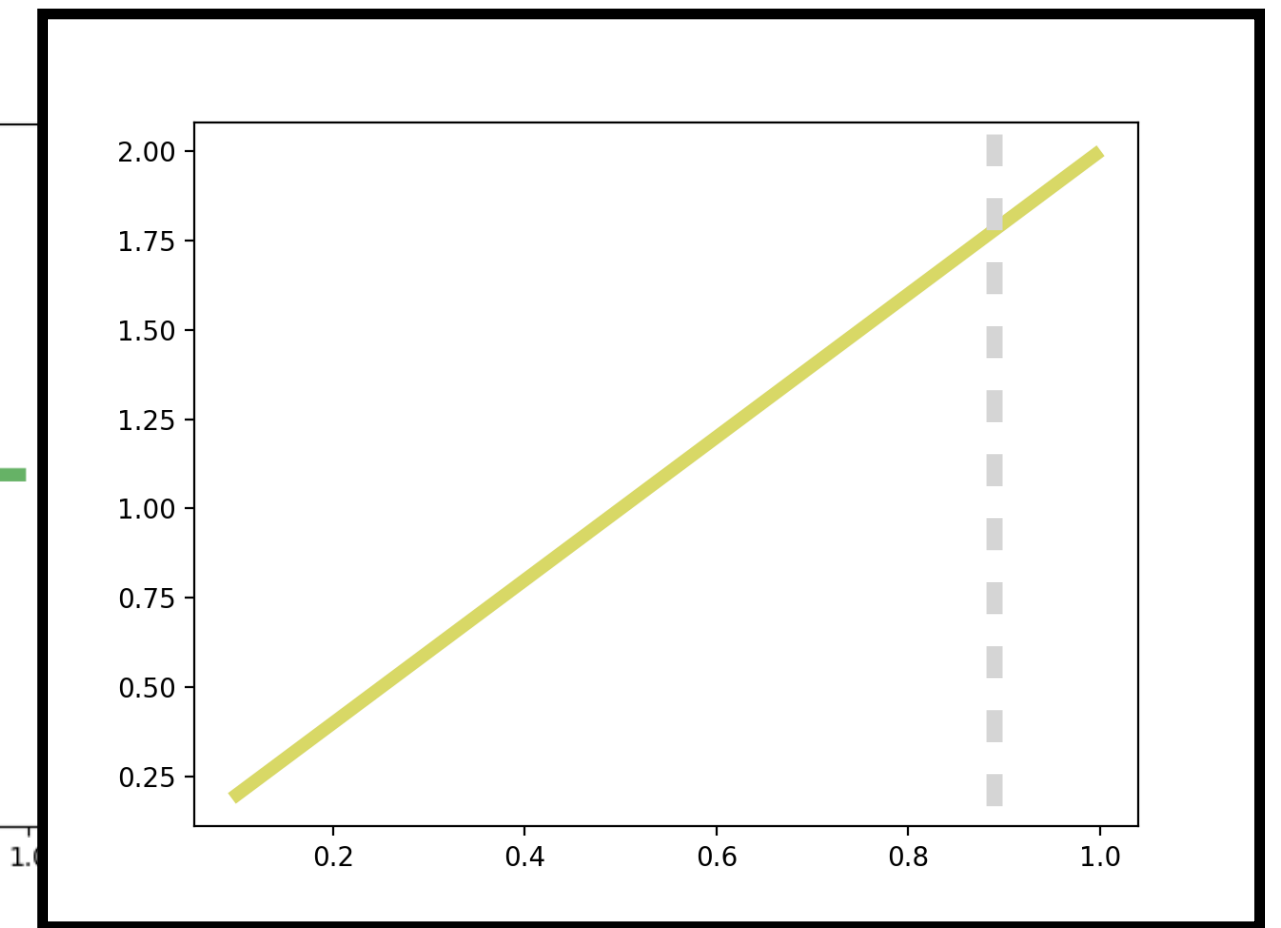
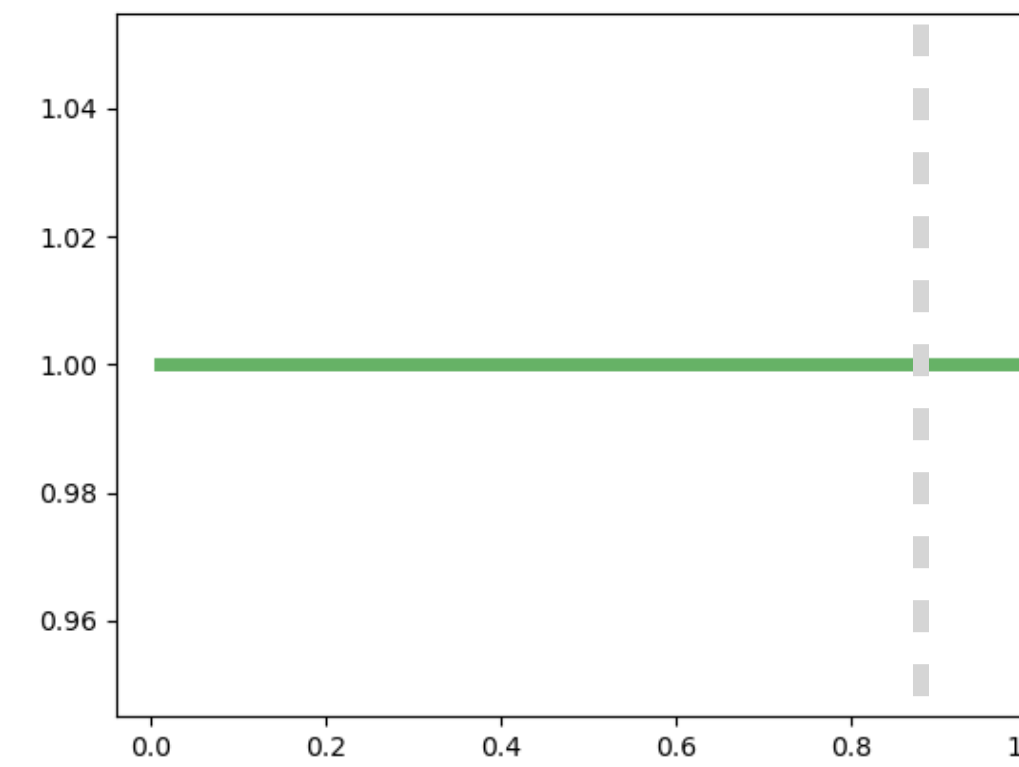
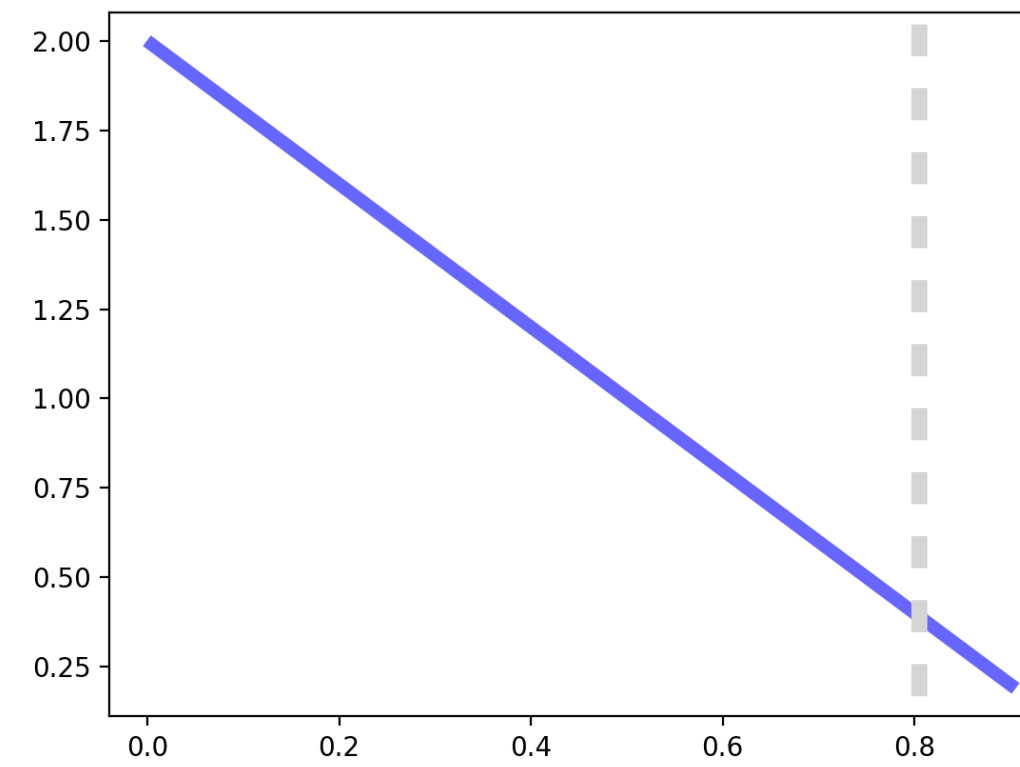
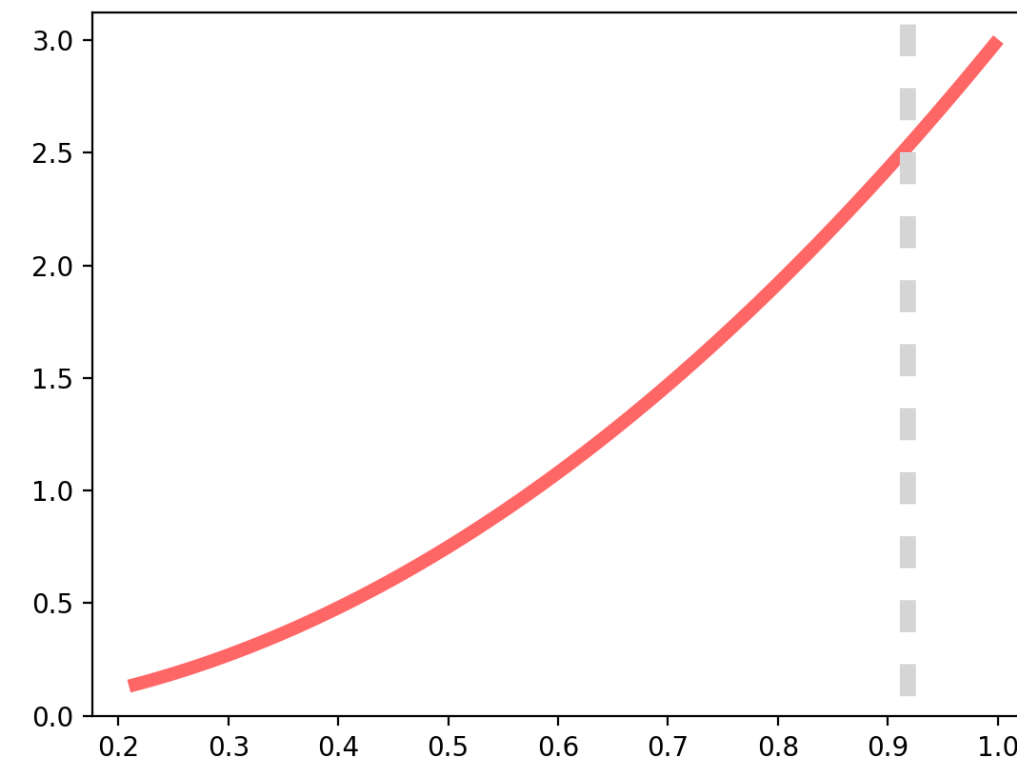


1

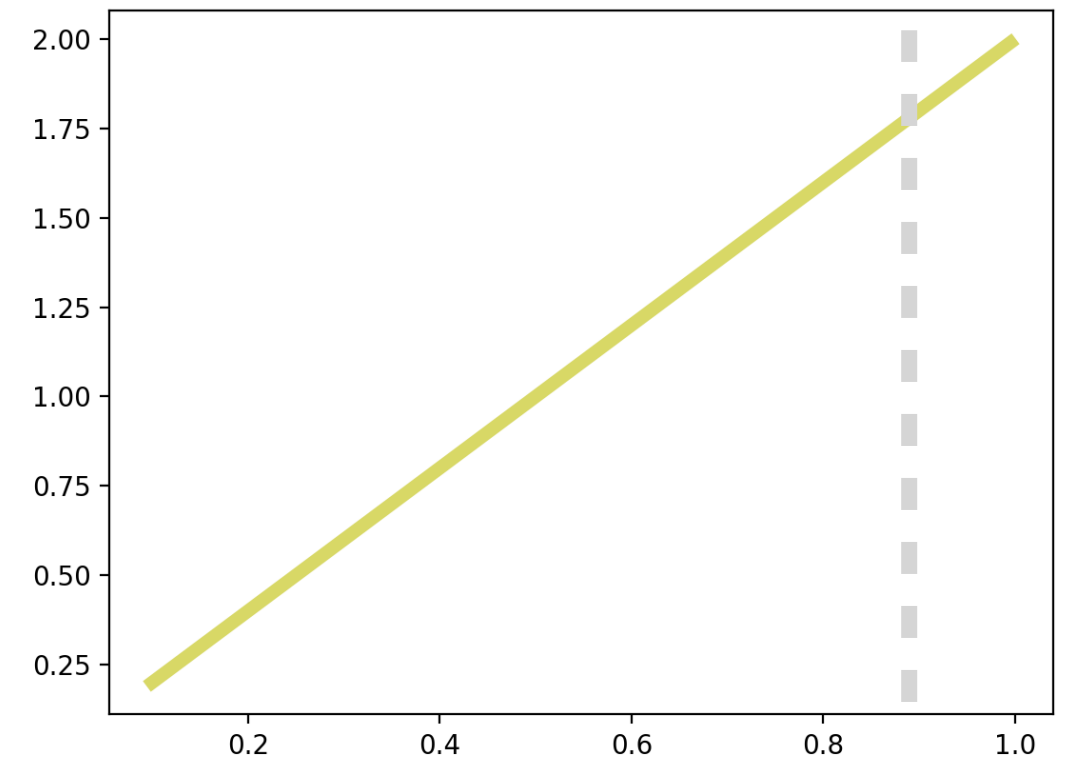
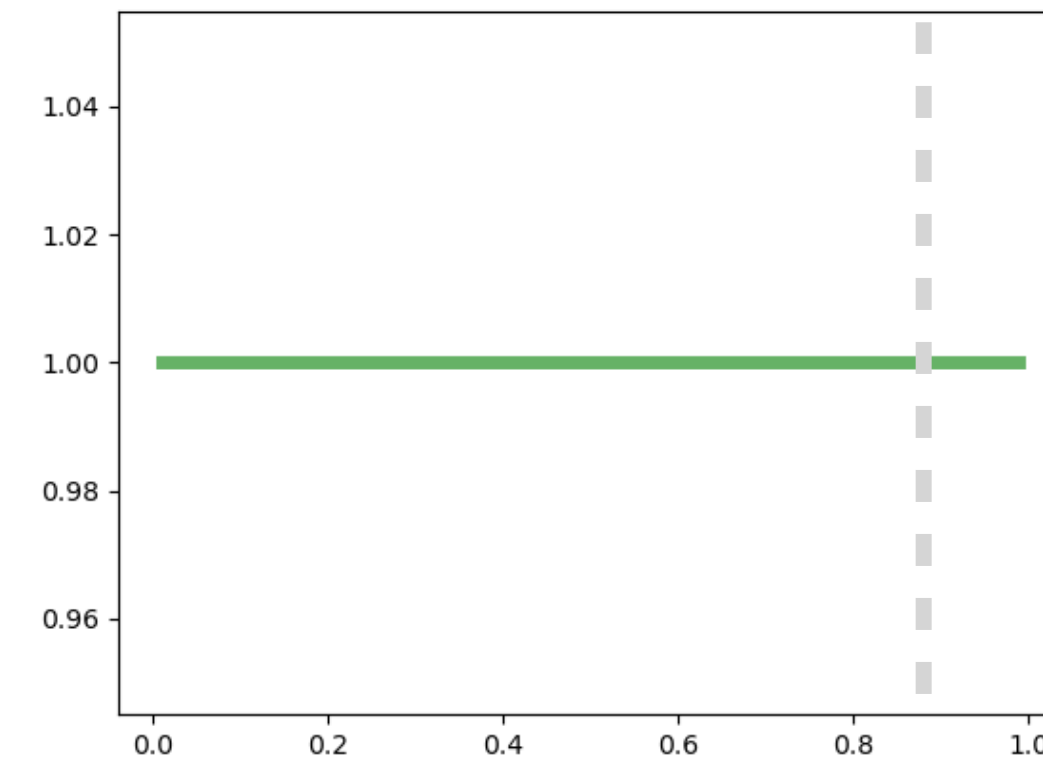
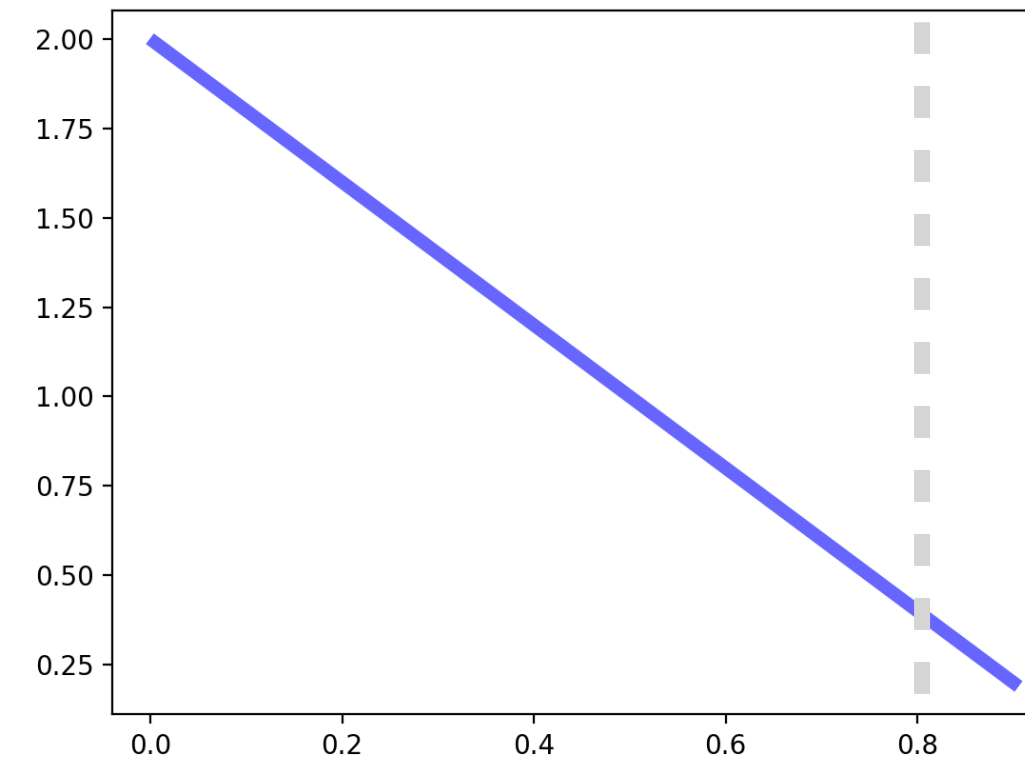
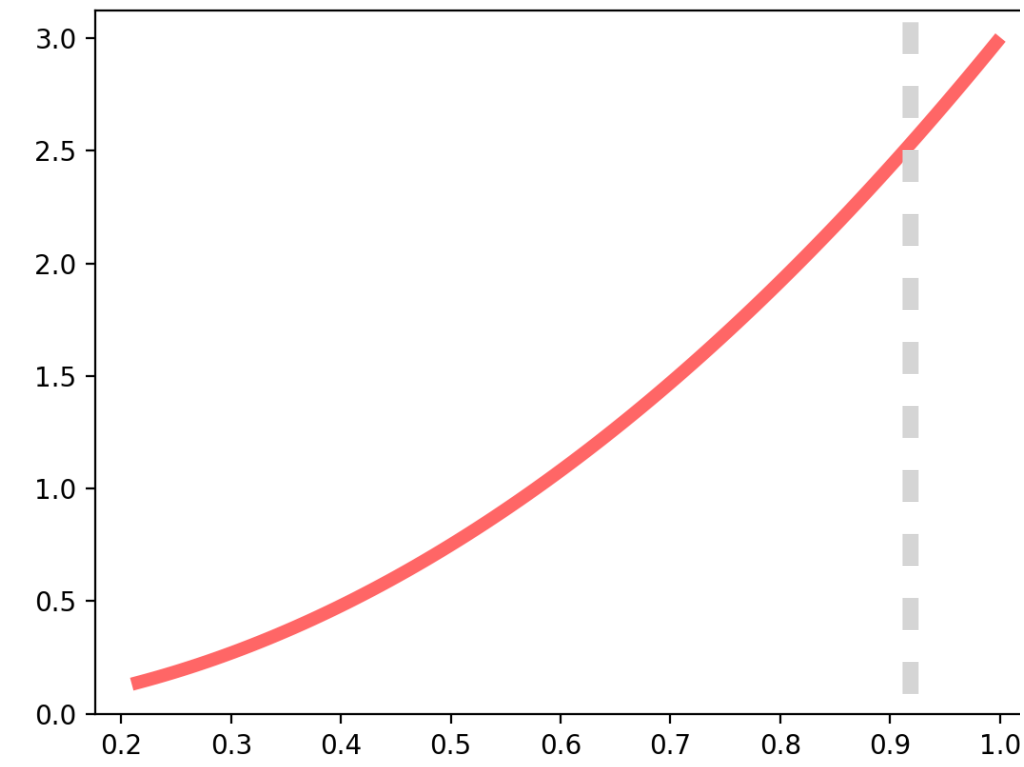
Sketching the Mechanism: Intuition



Sketching the Mechanism: Intuition

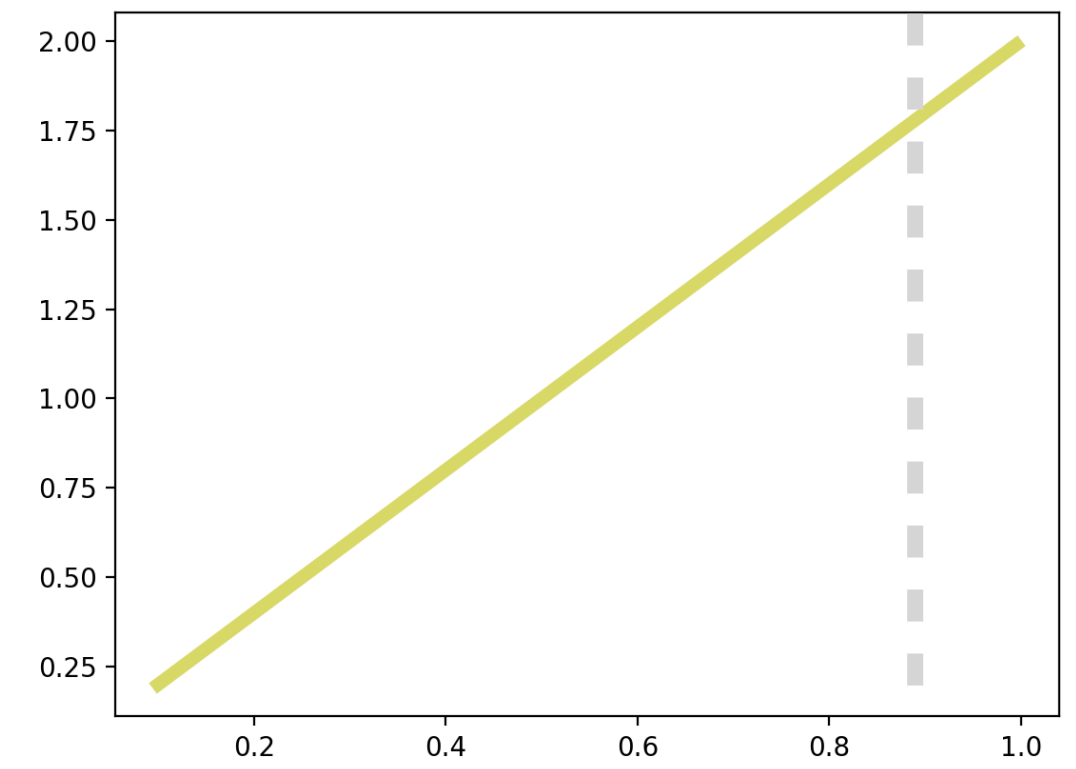
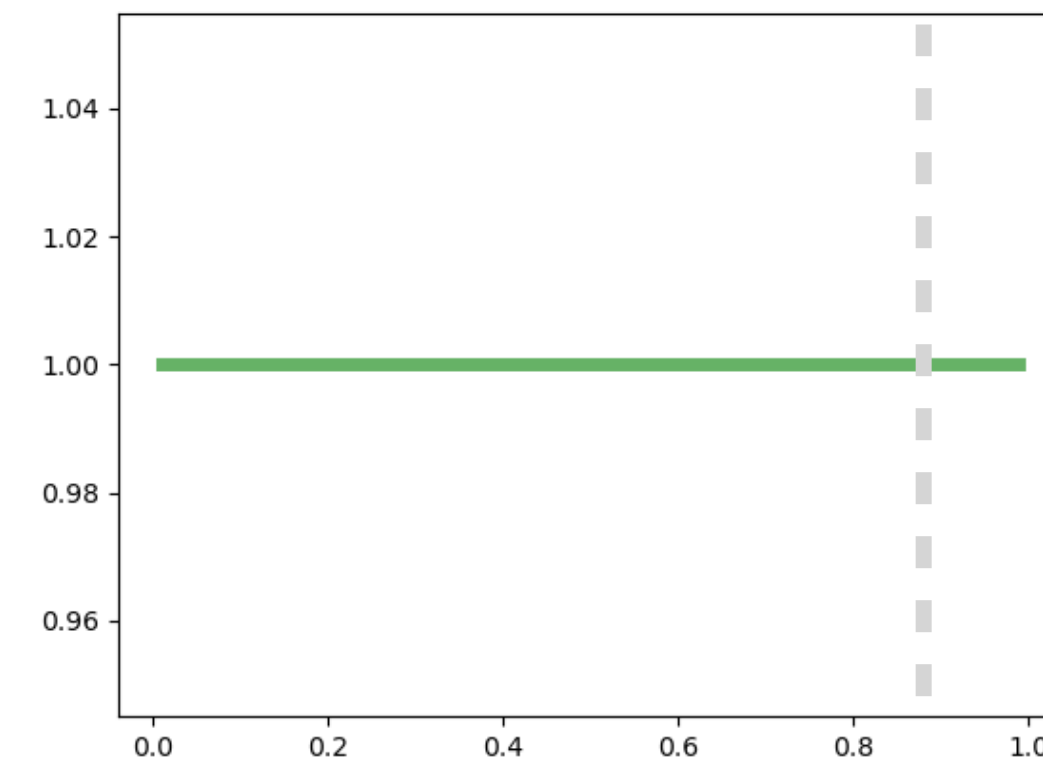
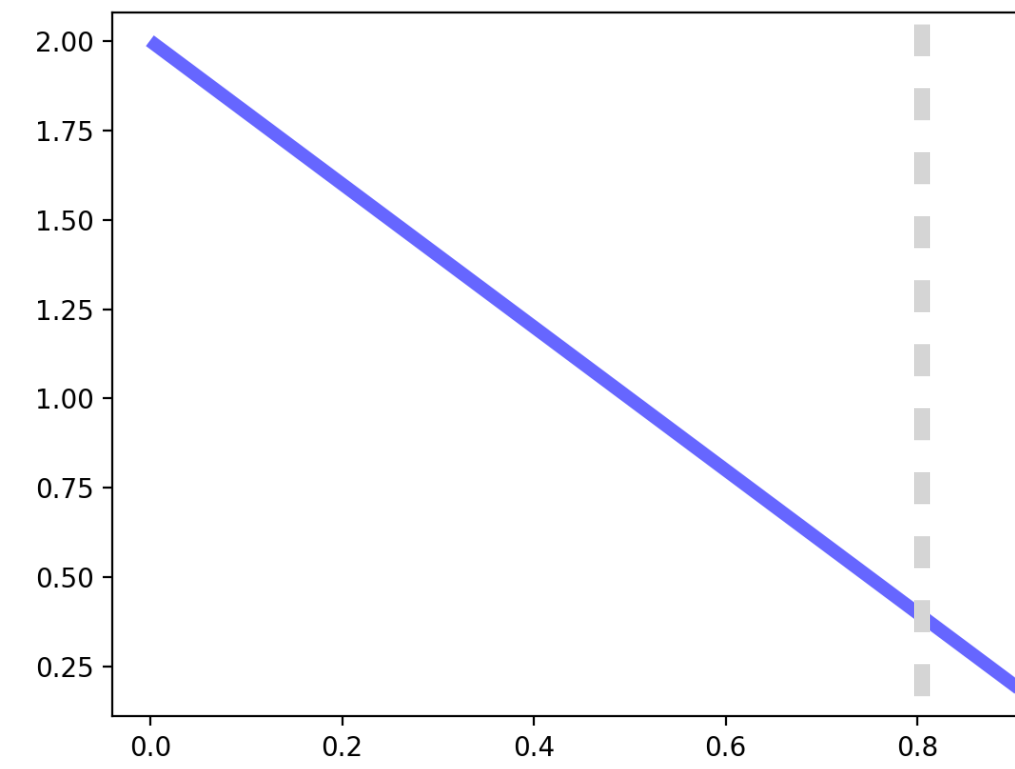
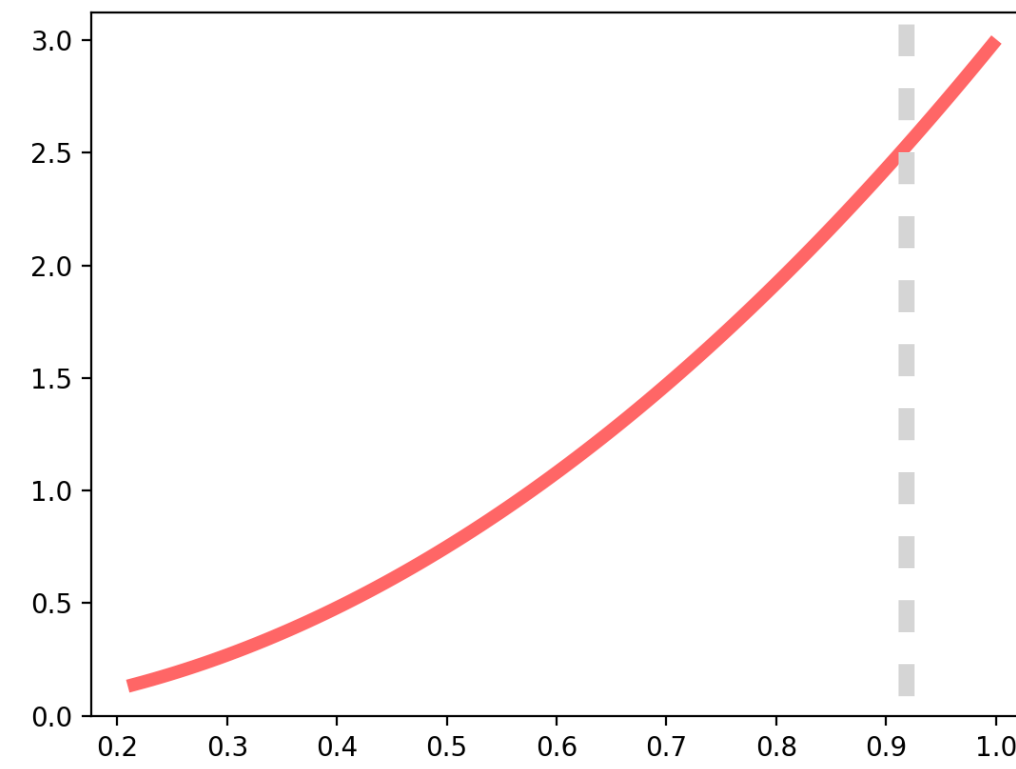


Sketching the Mechanism: Intuition

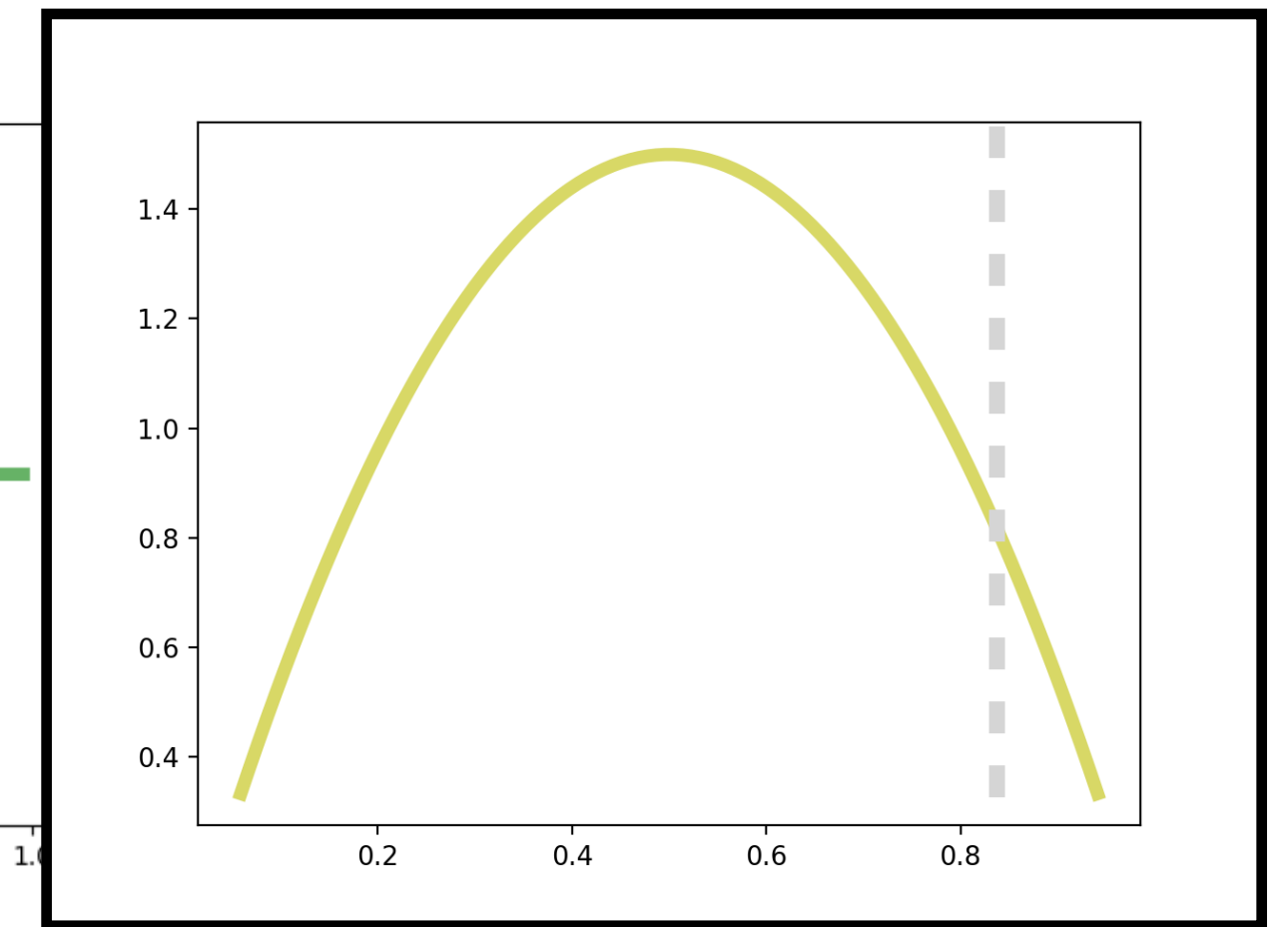
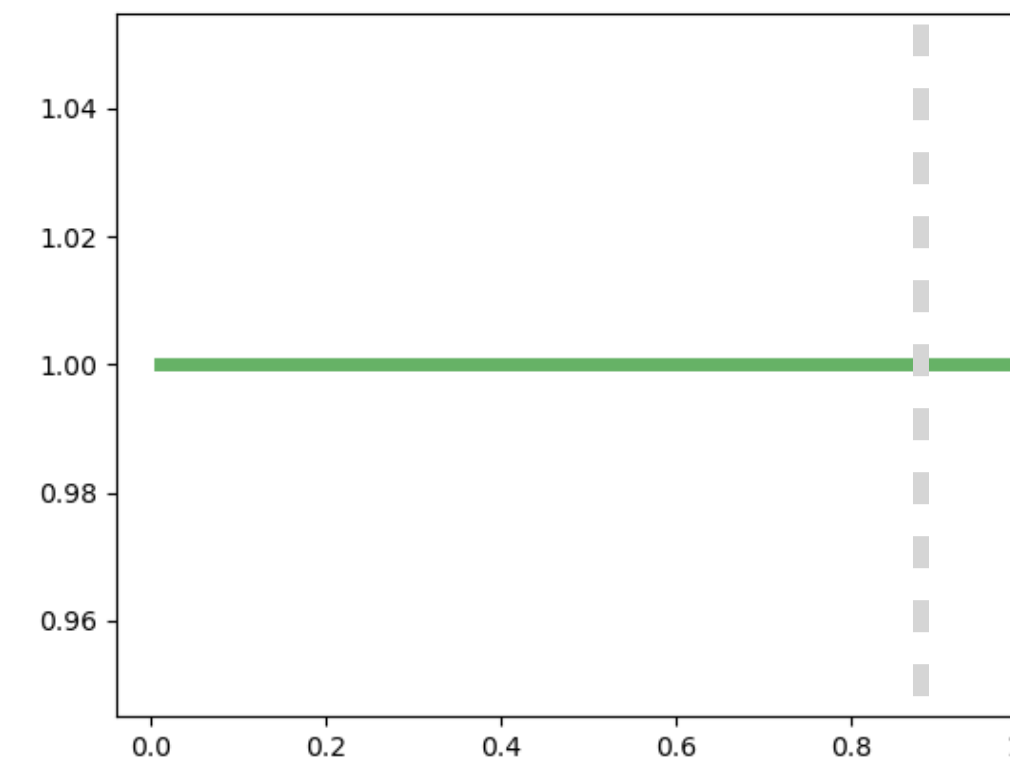
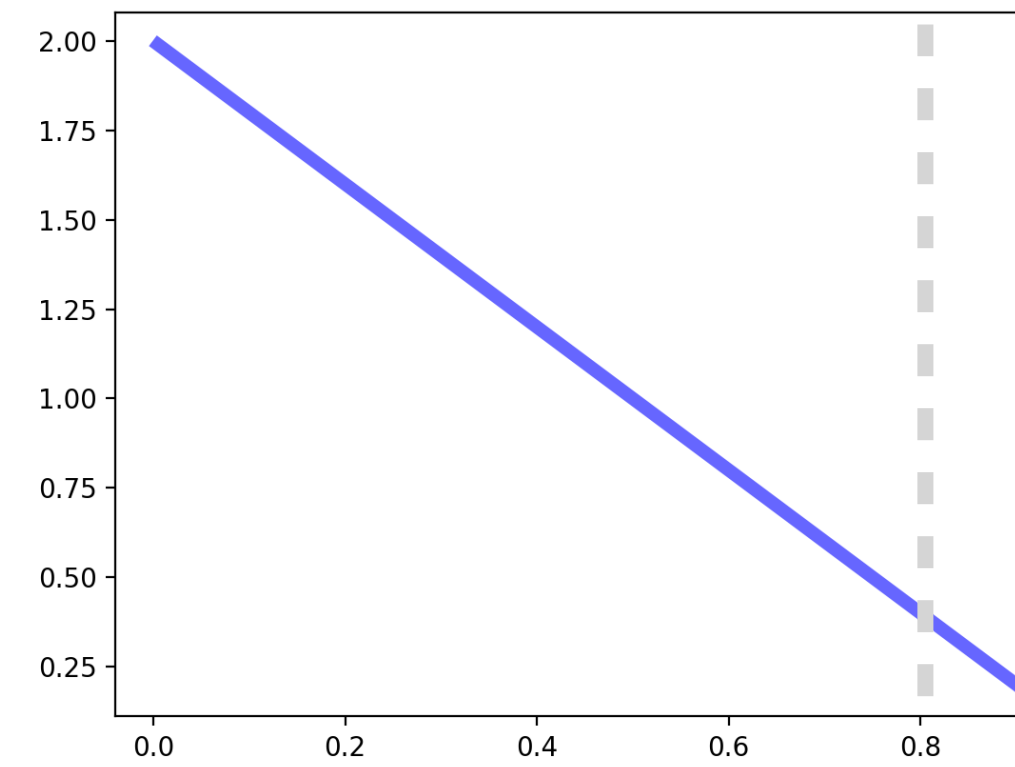
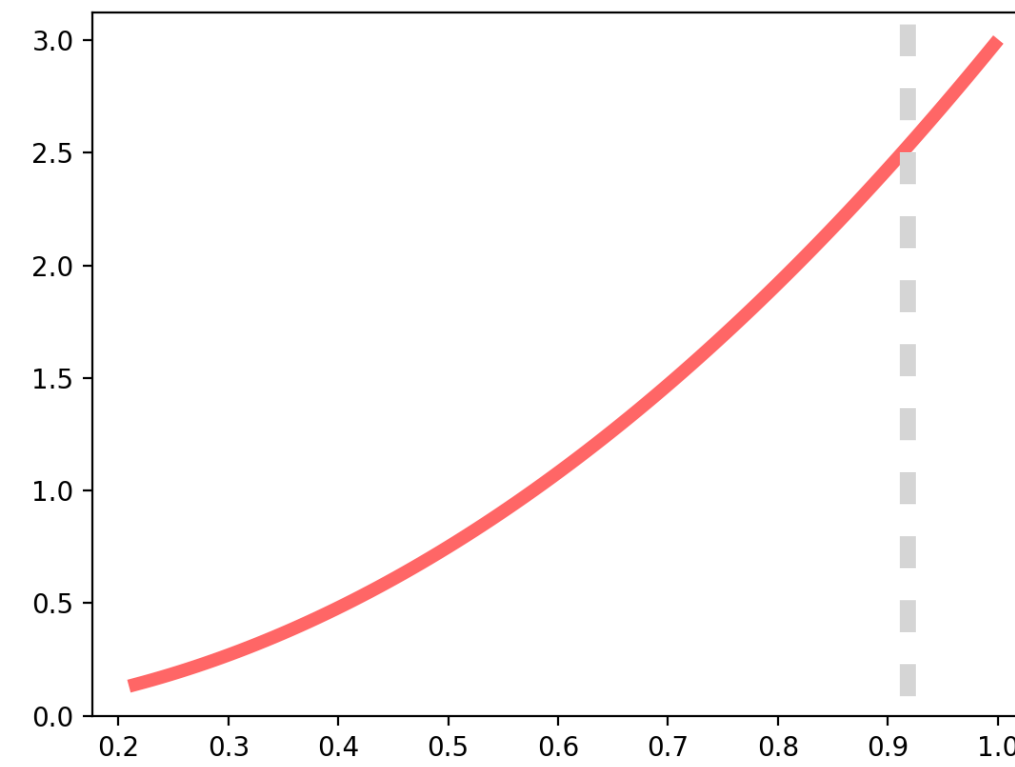


2

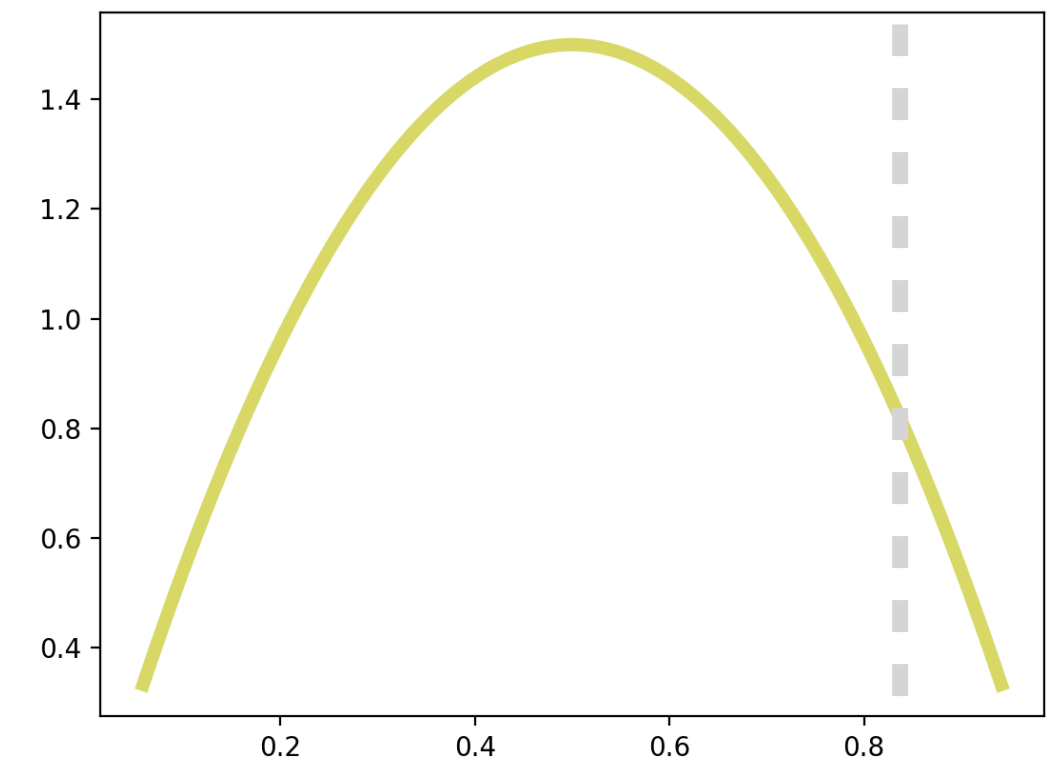
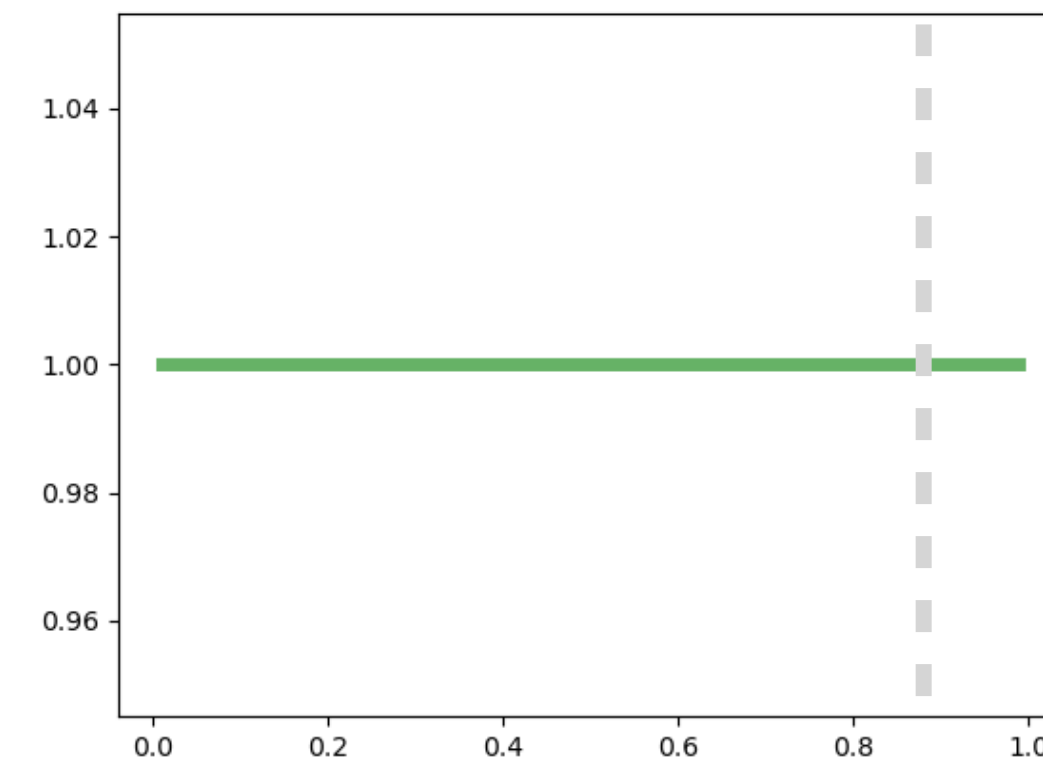
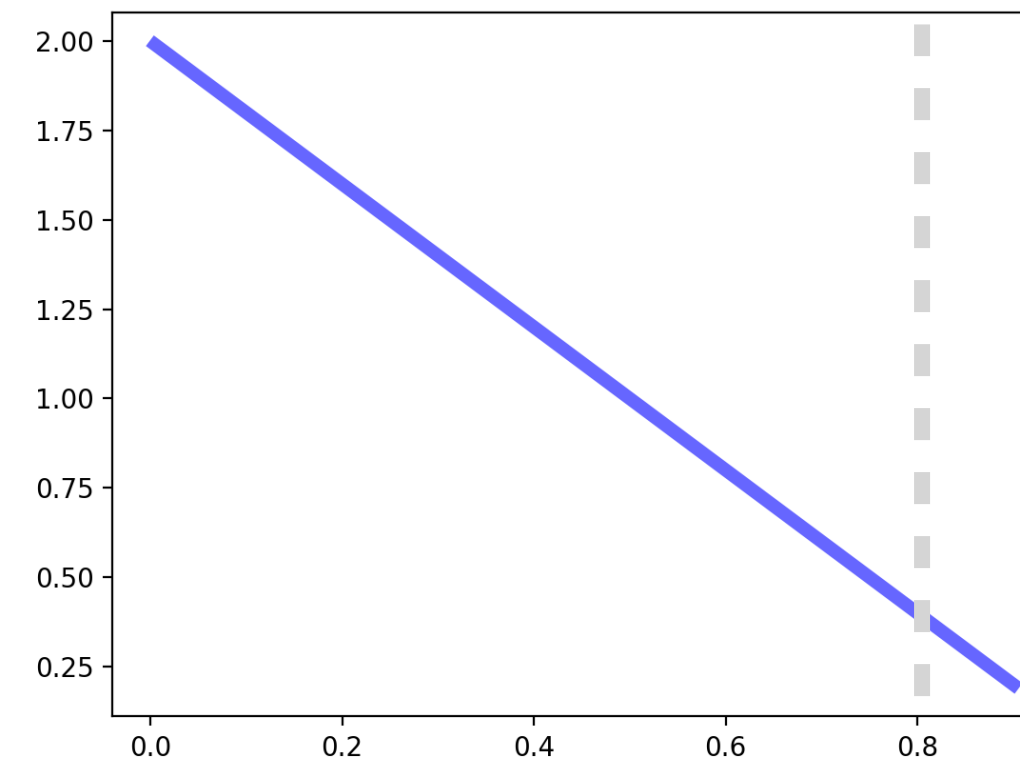
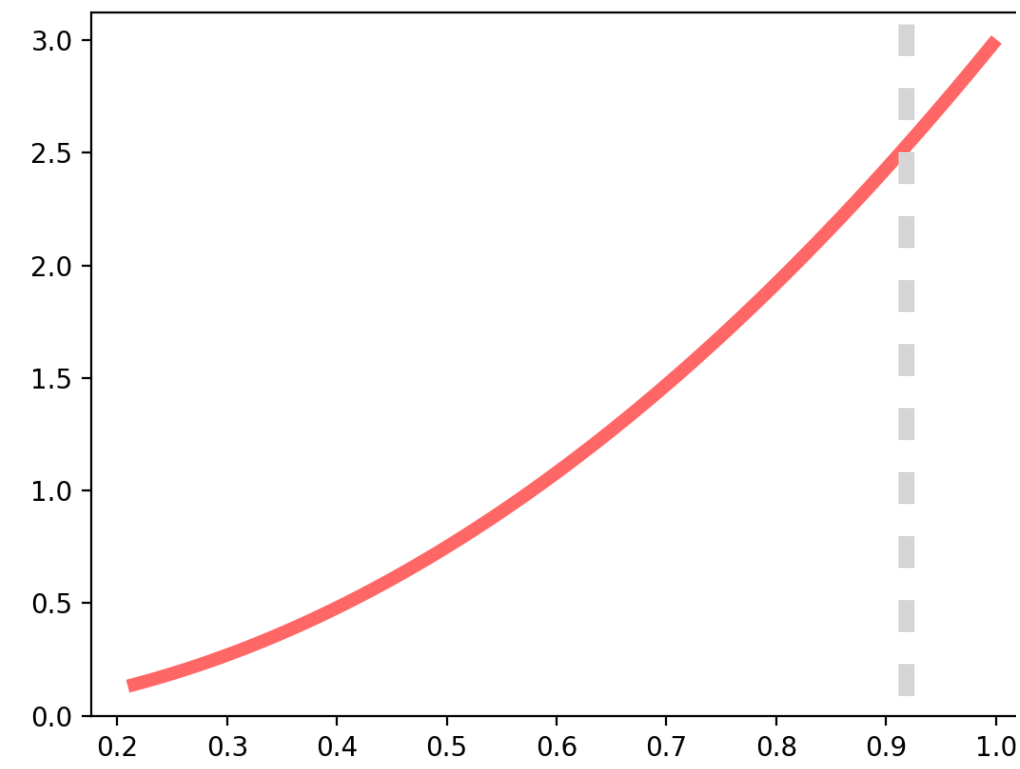
Sketching the Mechanism: Intuition



Sketching the Mechanism: Intuition

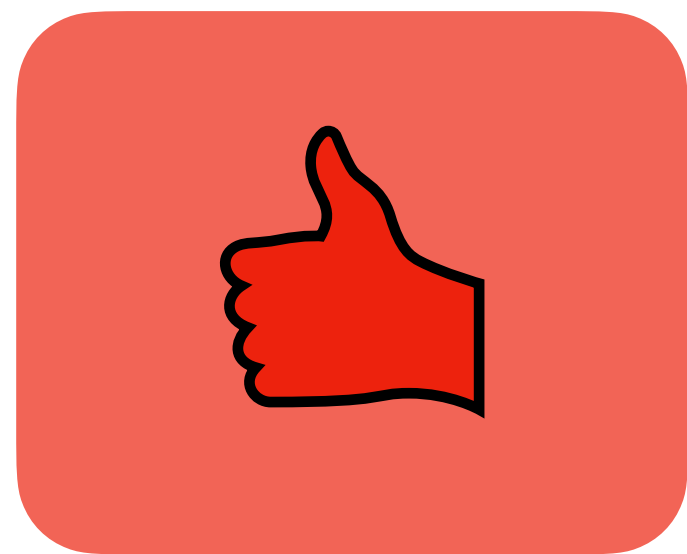
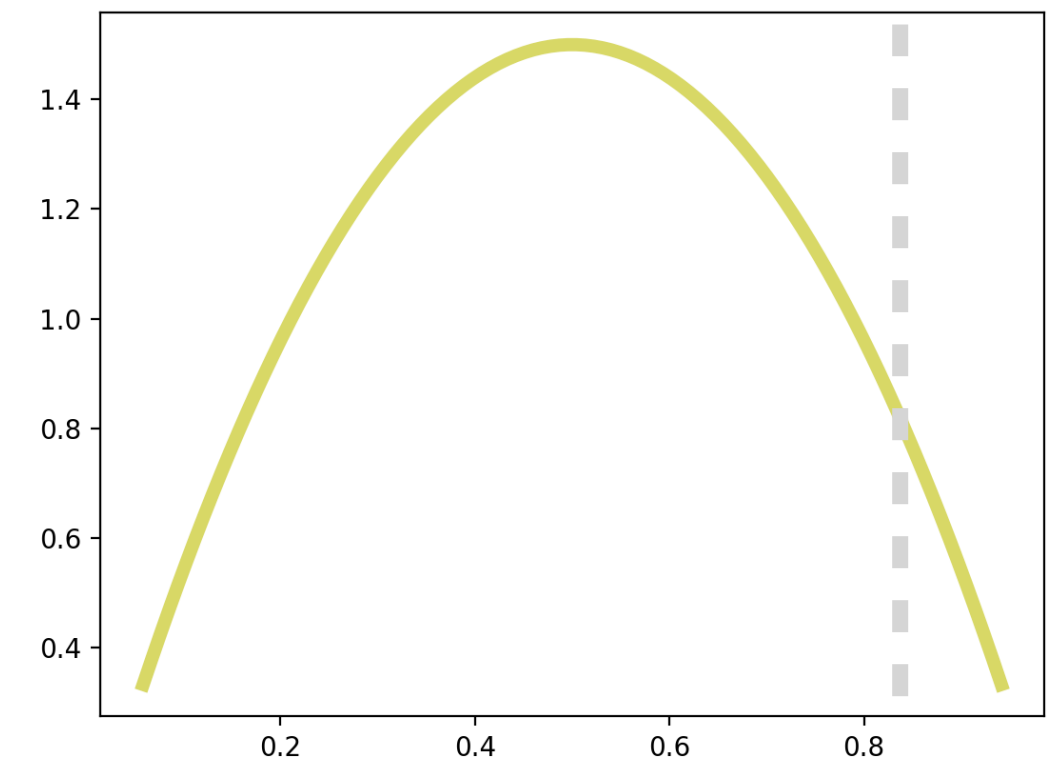
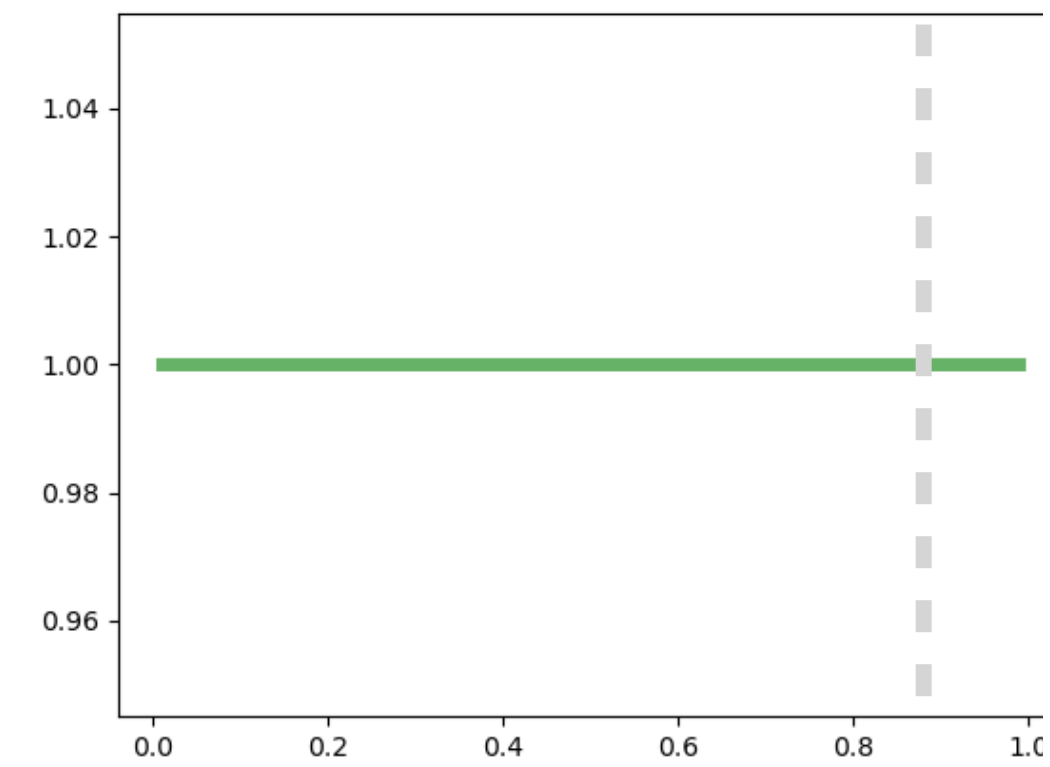
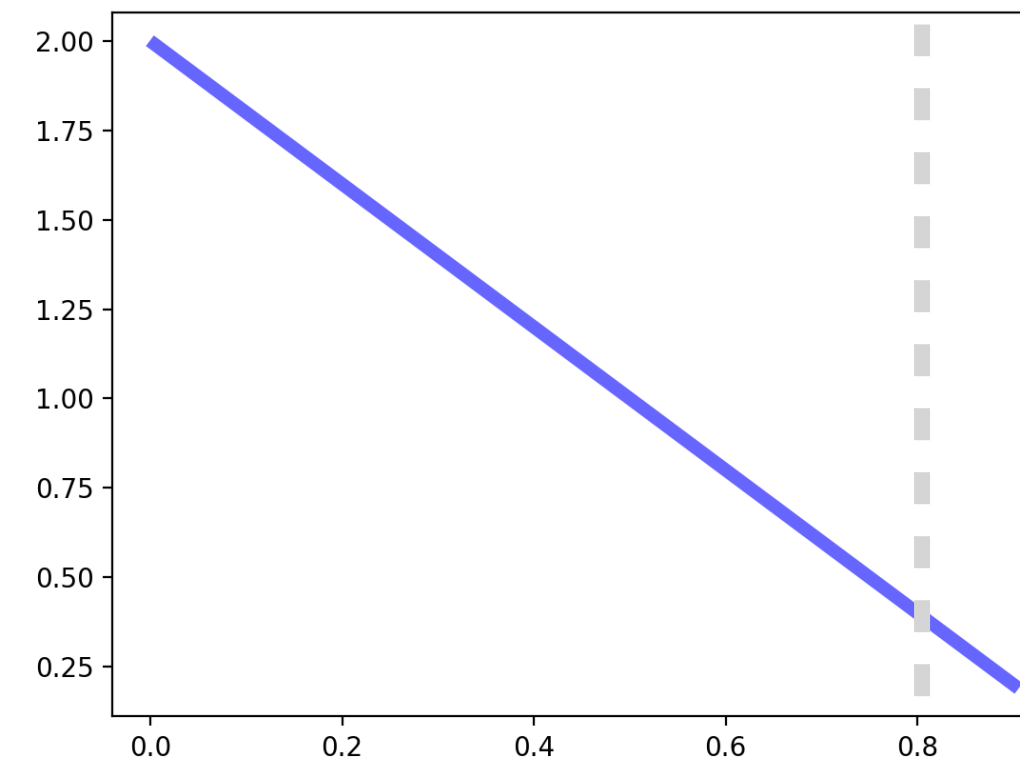
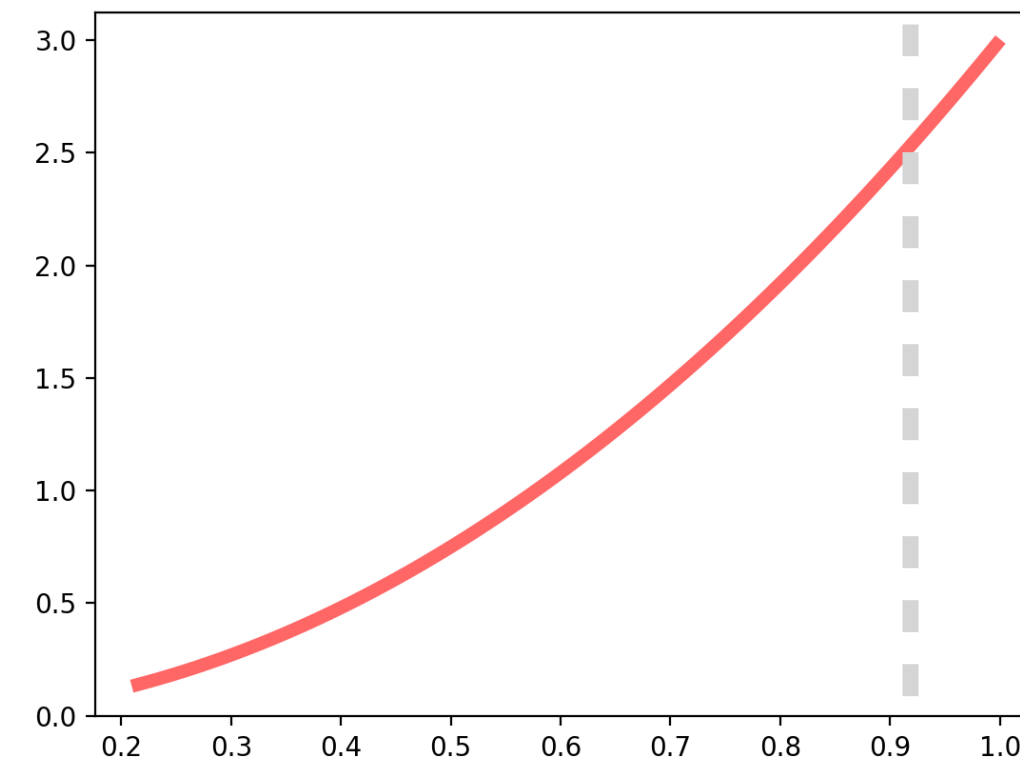


Sketching the Mechanism: Intuition

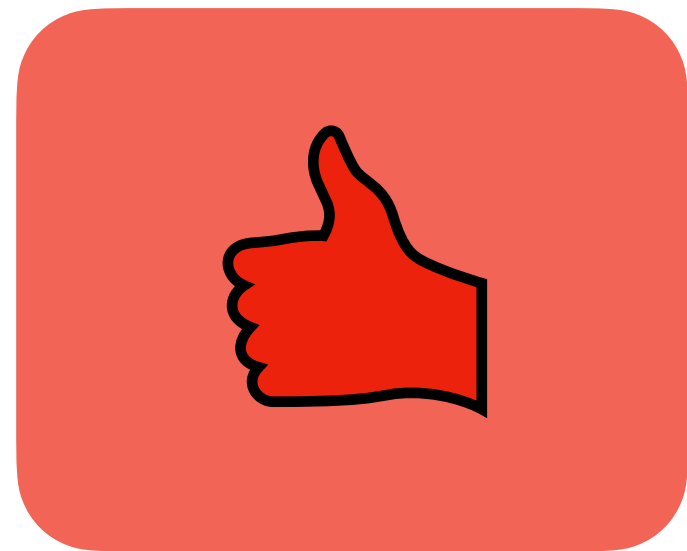
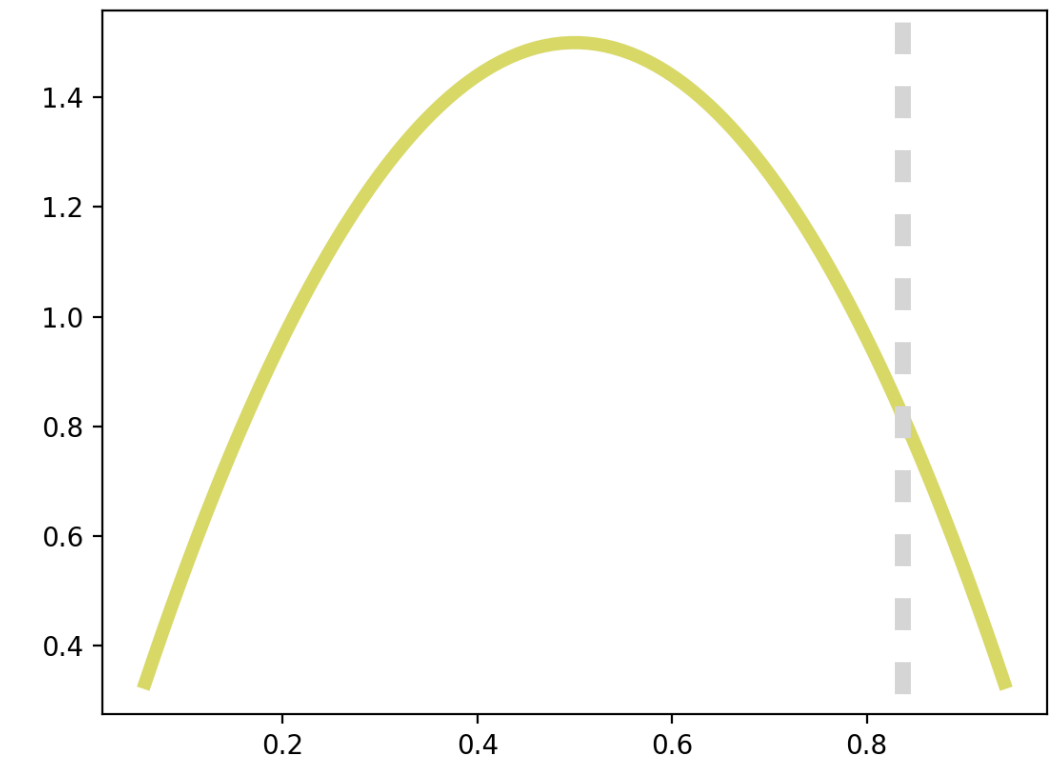
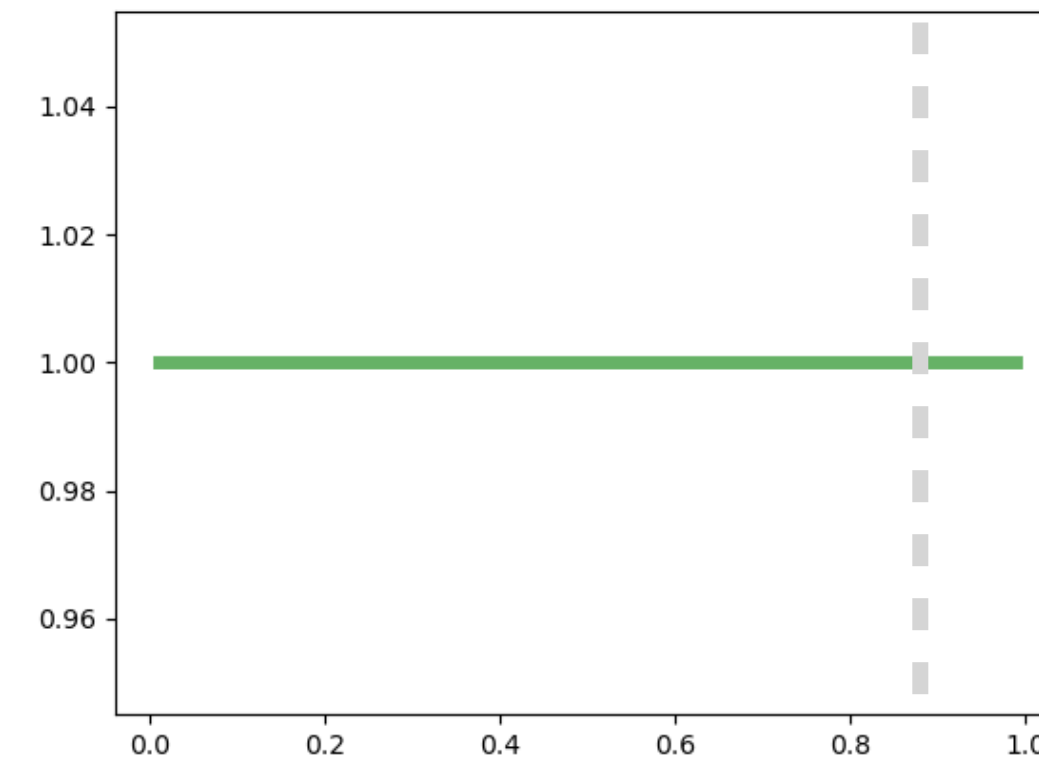
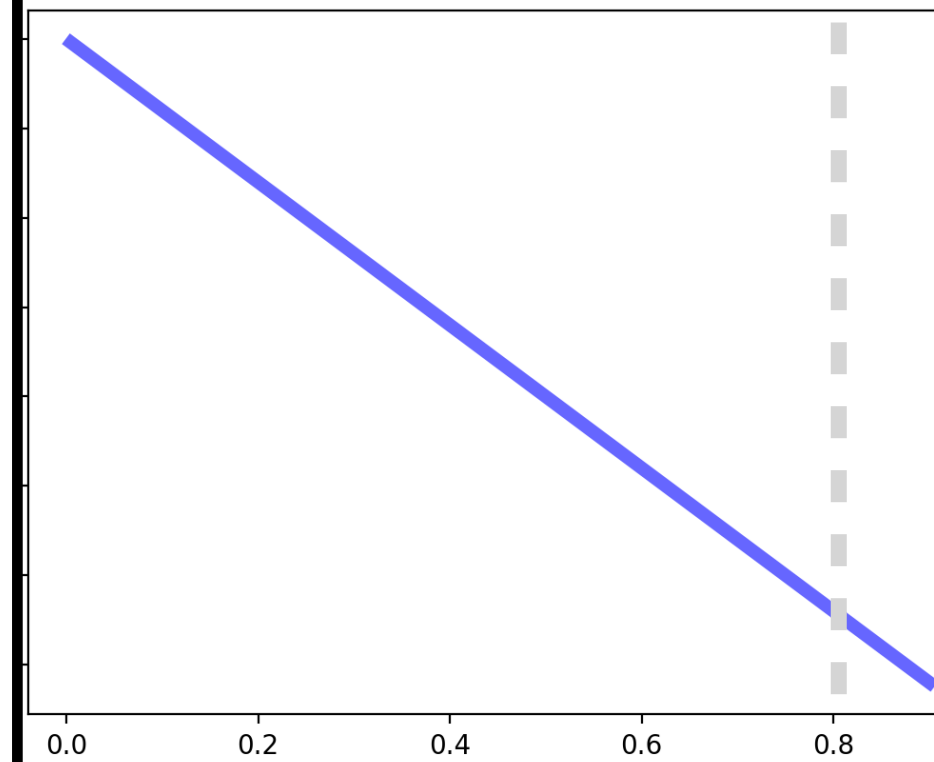
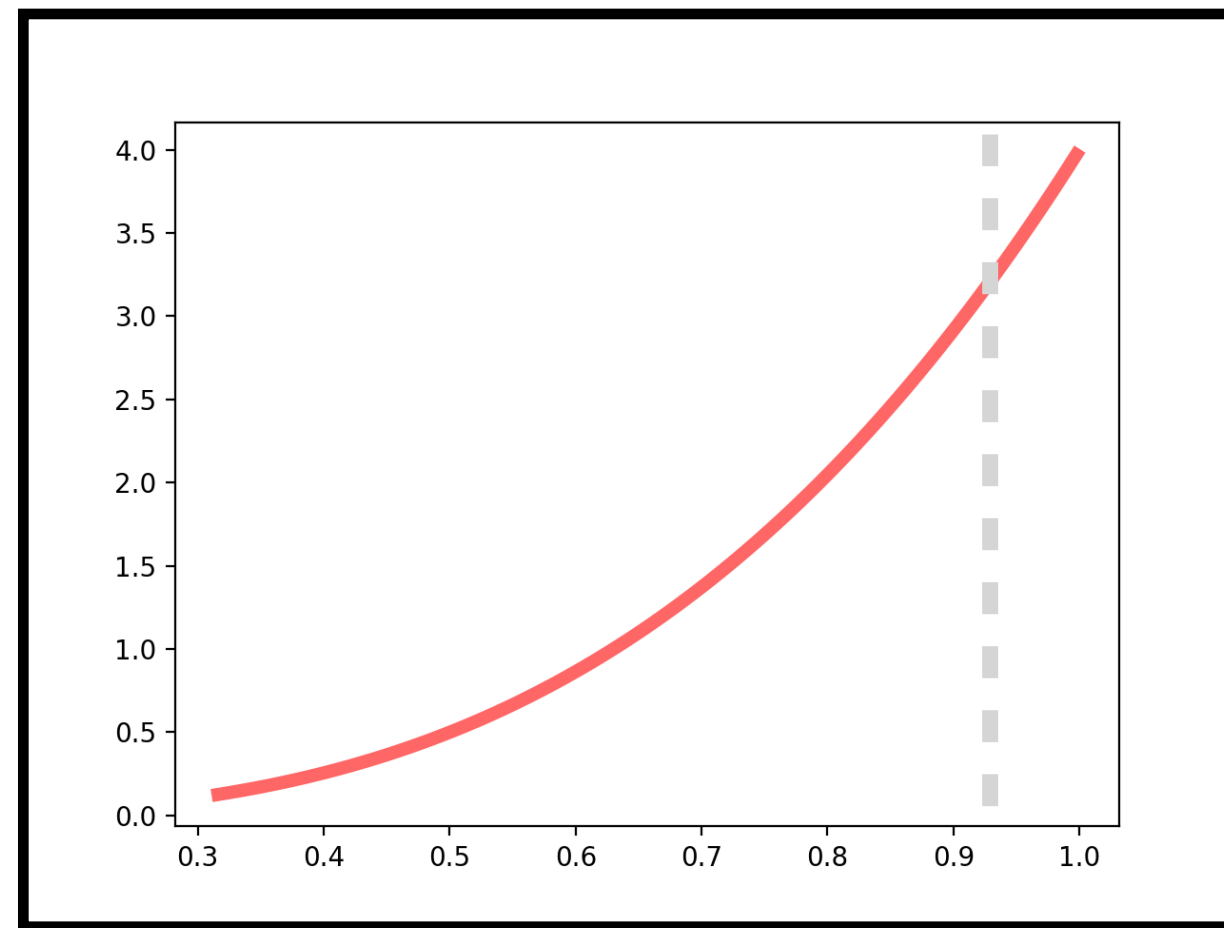


3

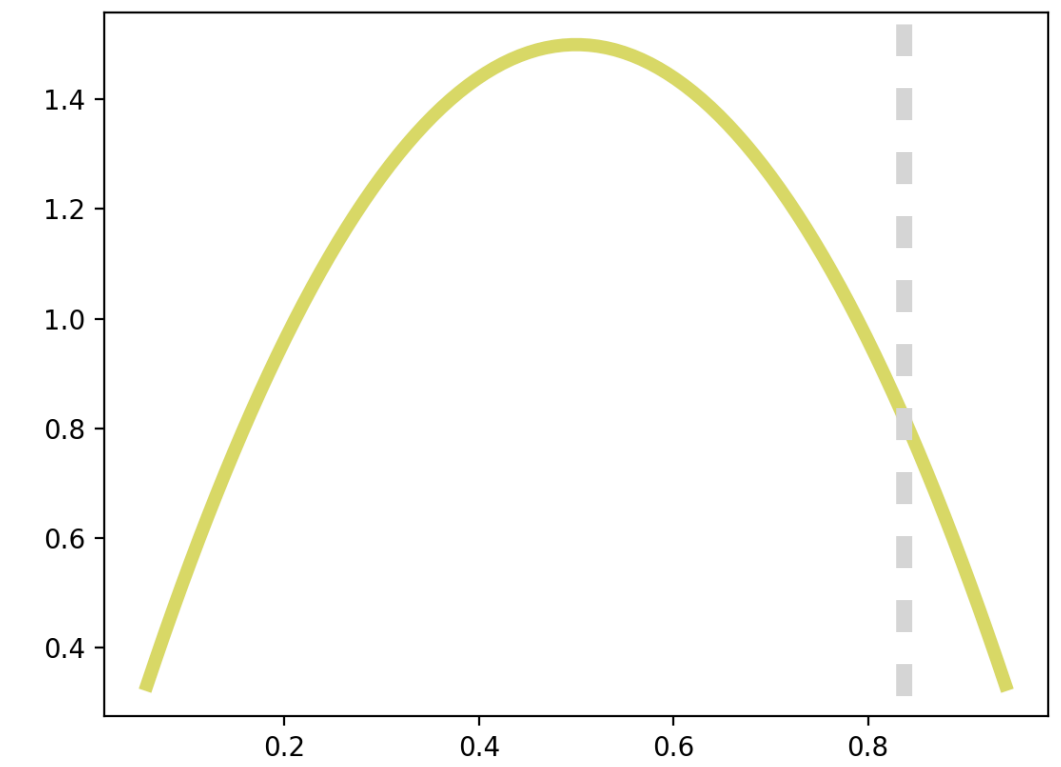
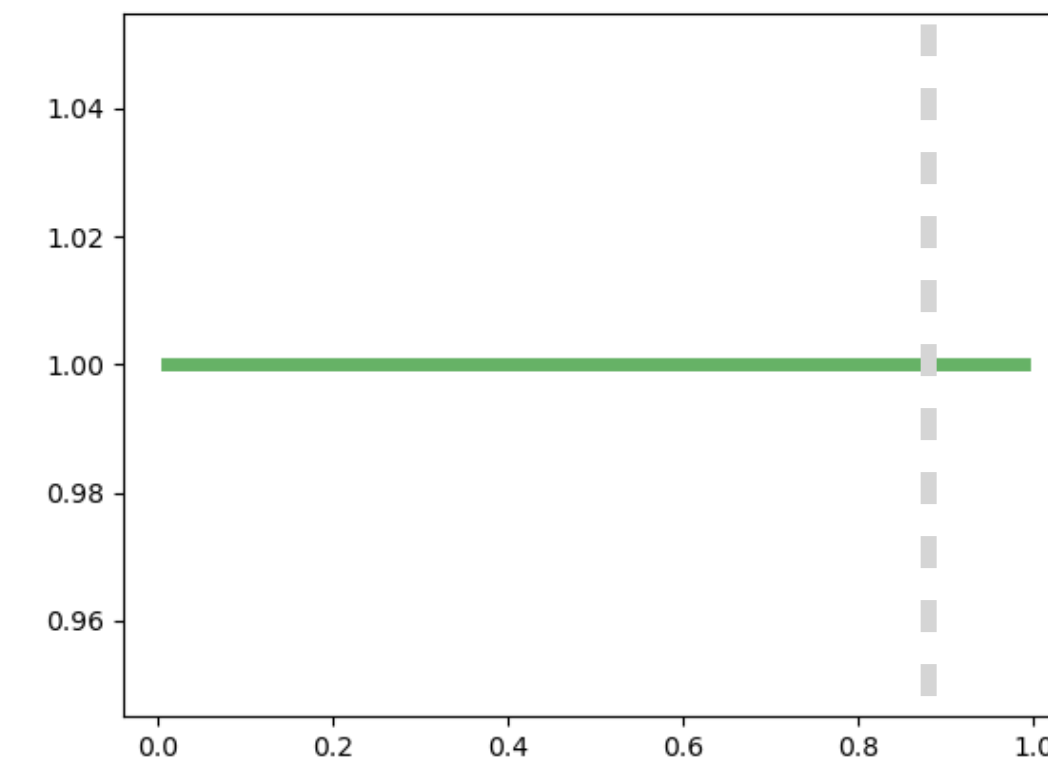
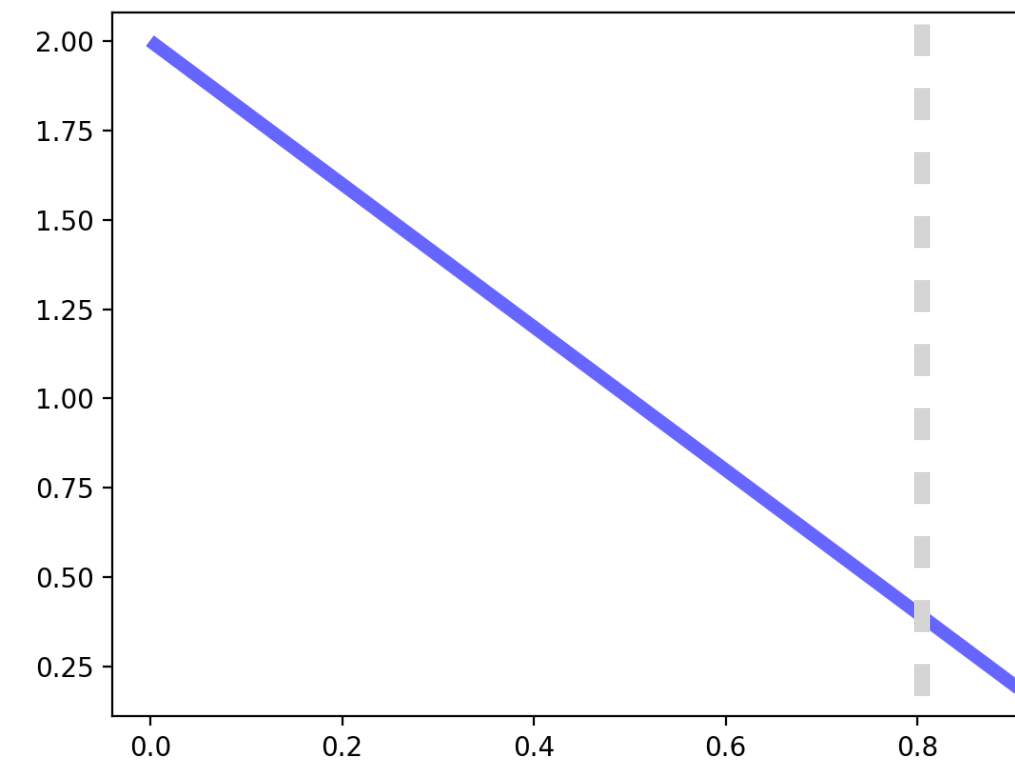
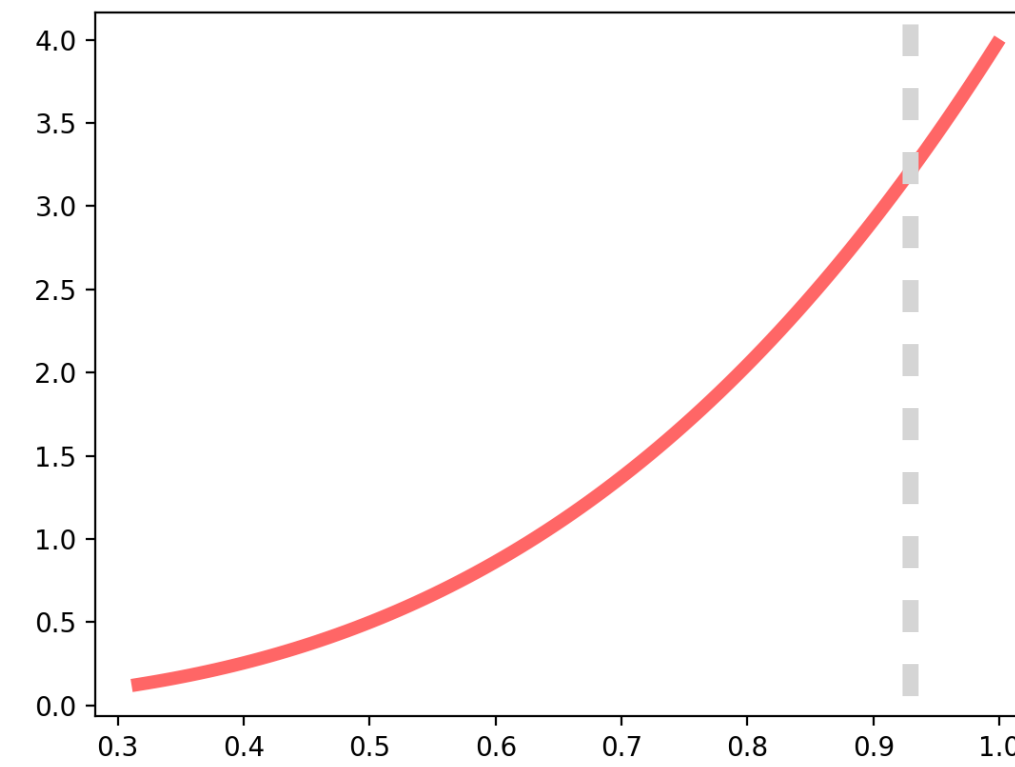
Sketching the Mechanism: Intuition



Sketching the Mechanism: Intuition

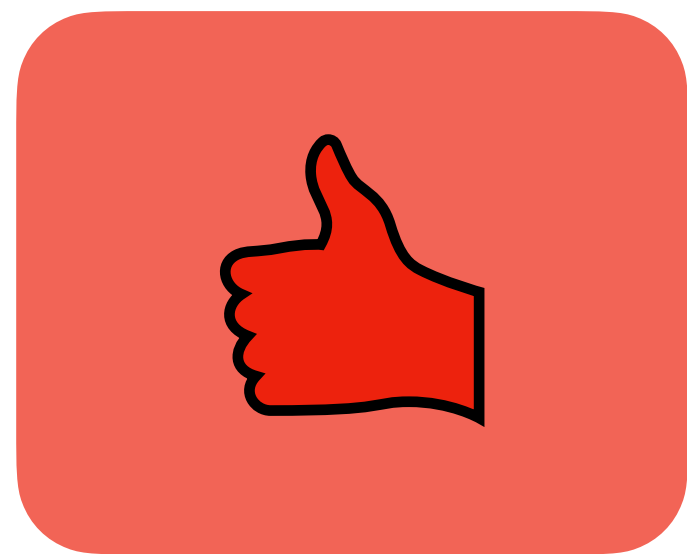
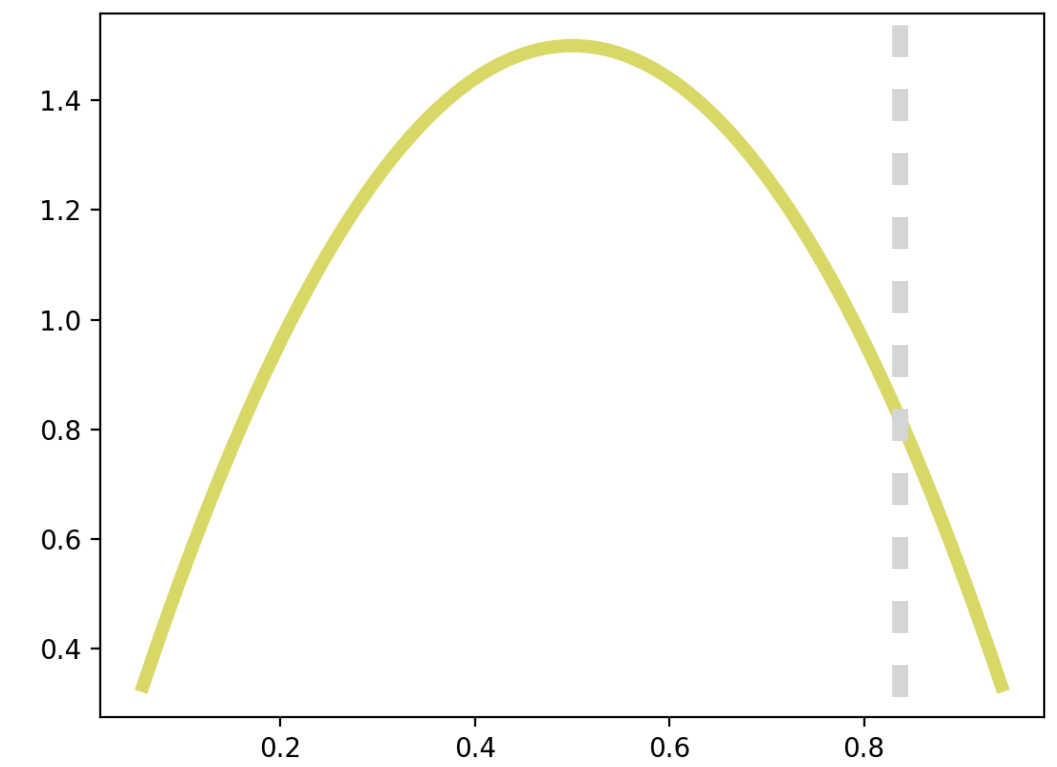
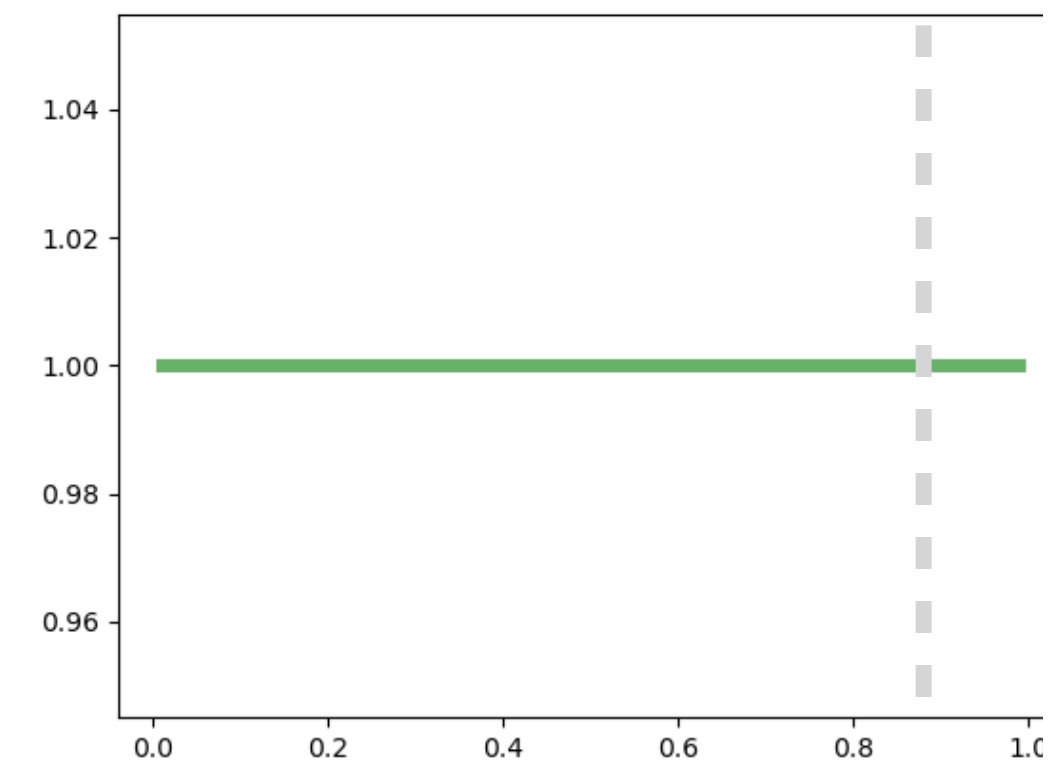
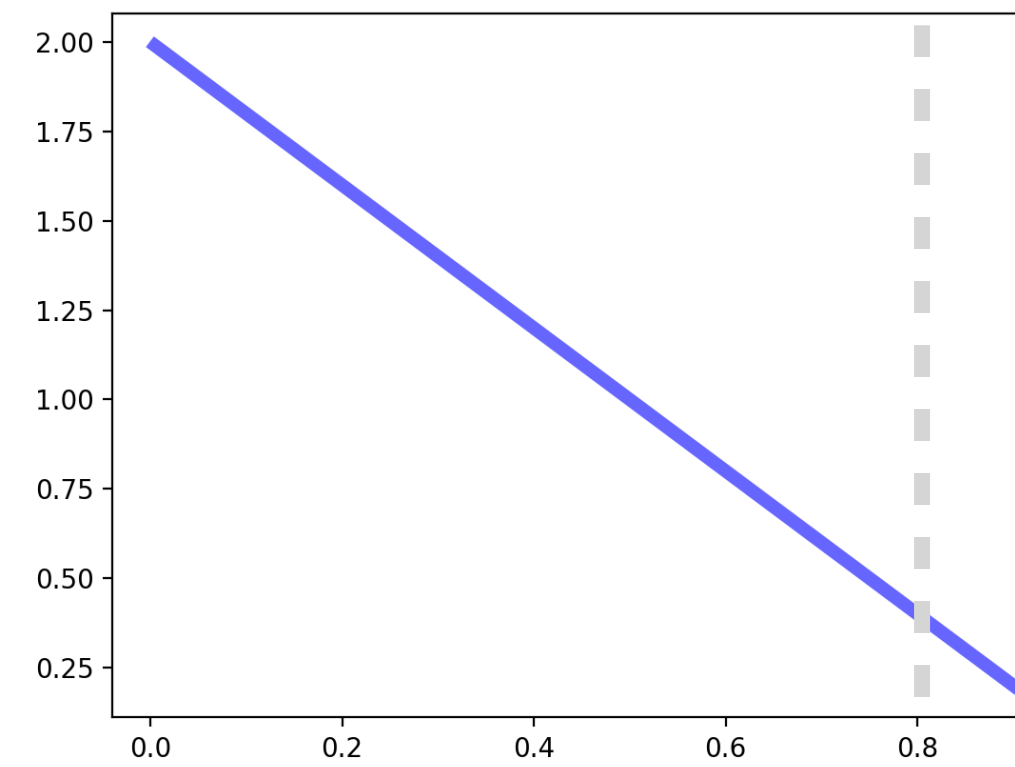
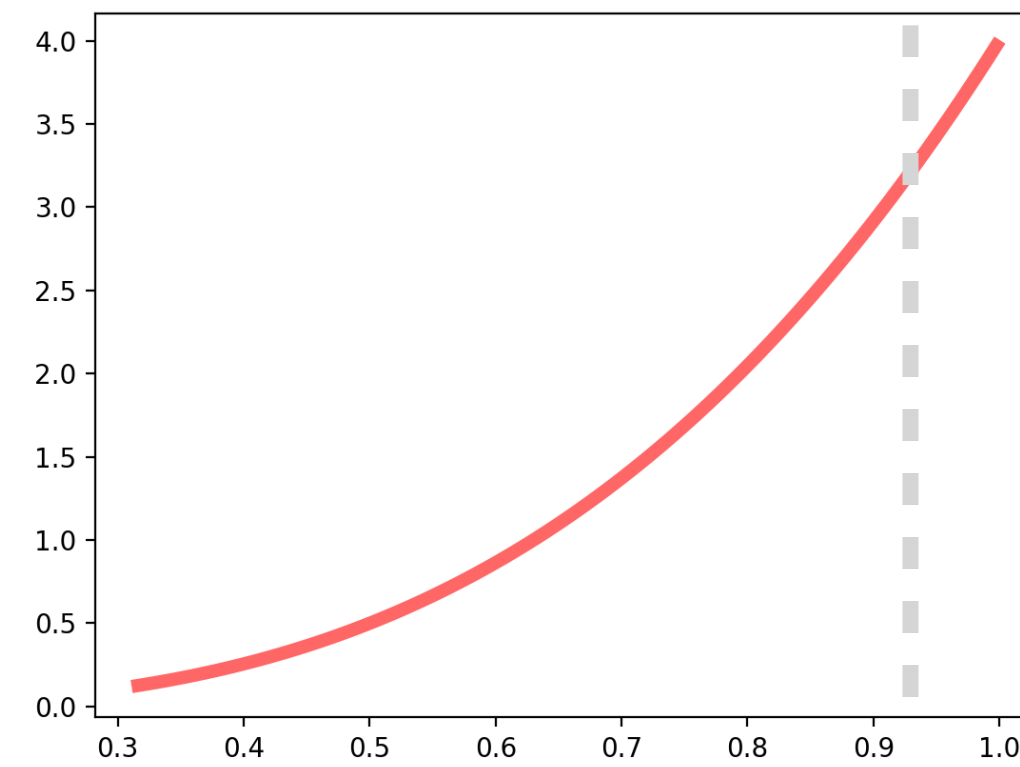


Sketching the Mechanism: Intuition

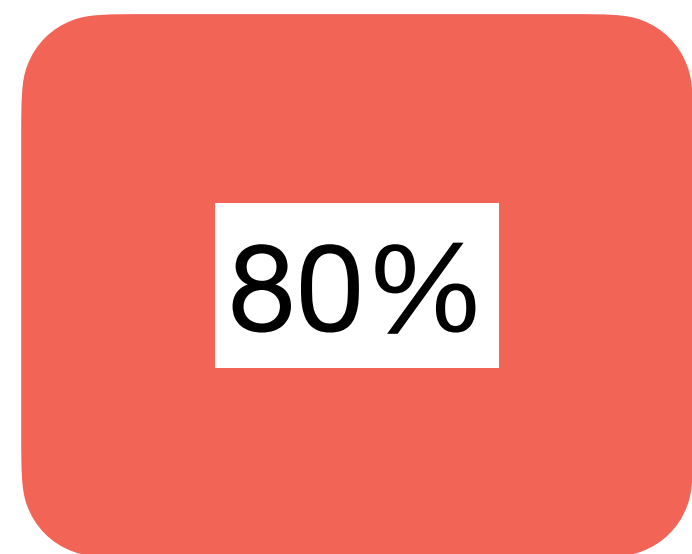
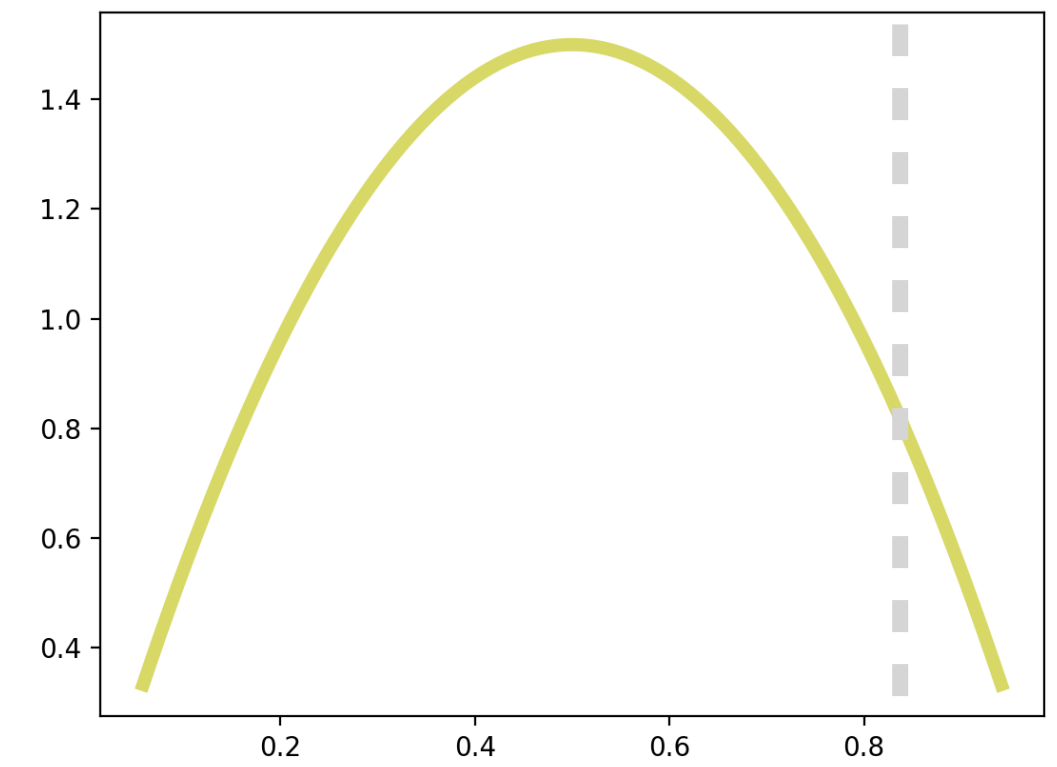
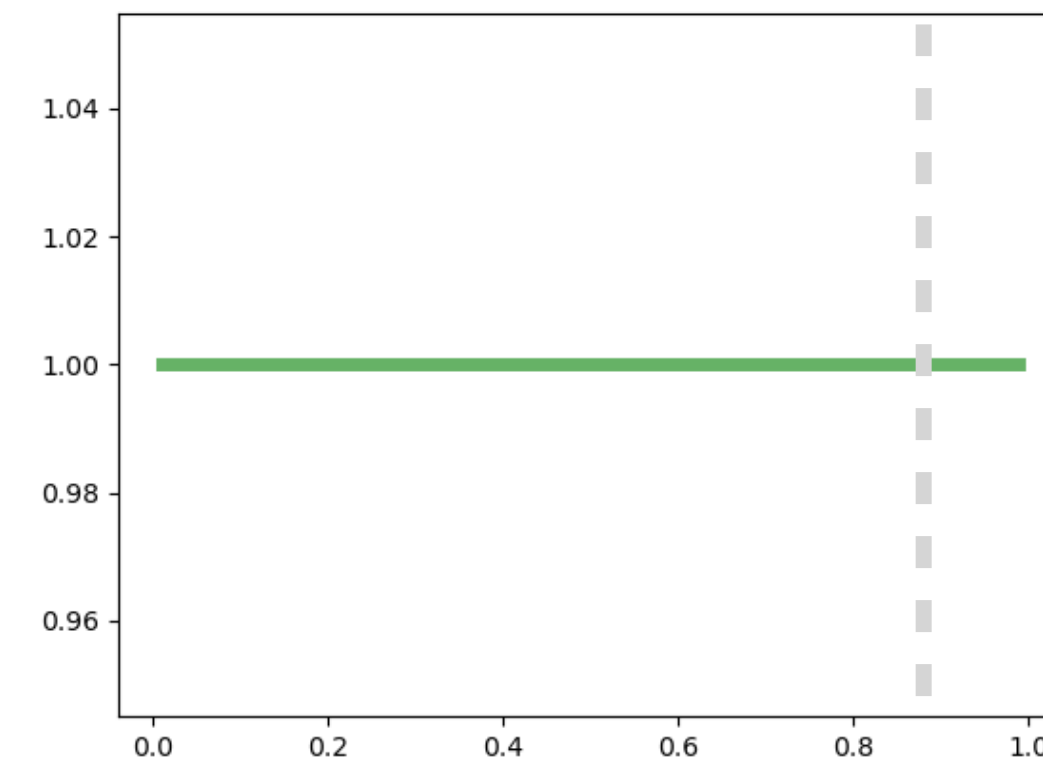
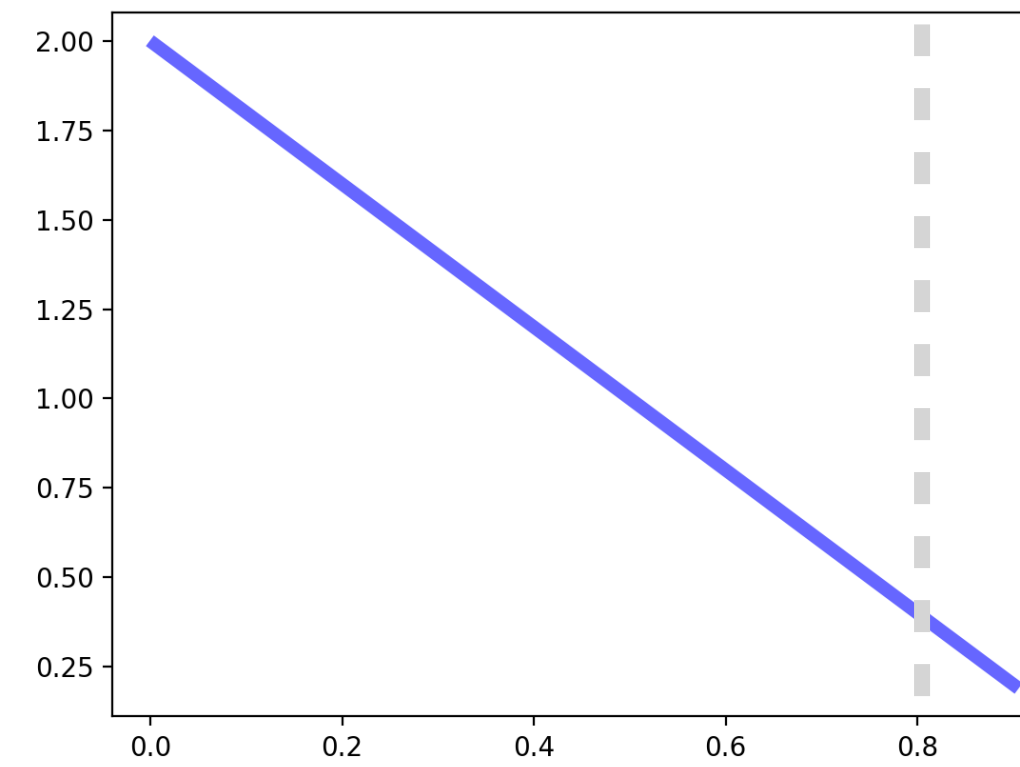
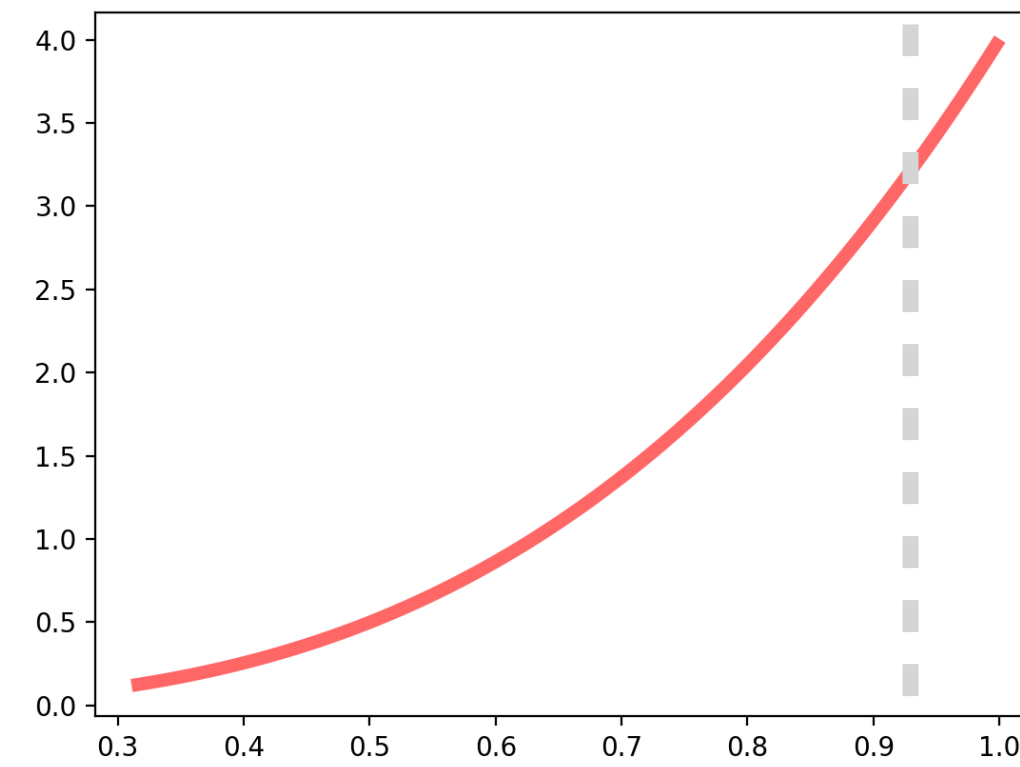


4

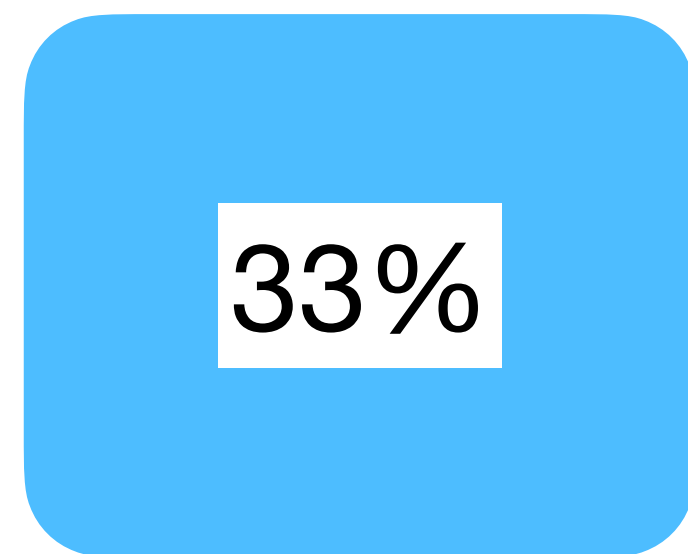
Sketching the Mechanism: Intuition



Sketching the Mechanism: Intuition



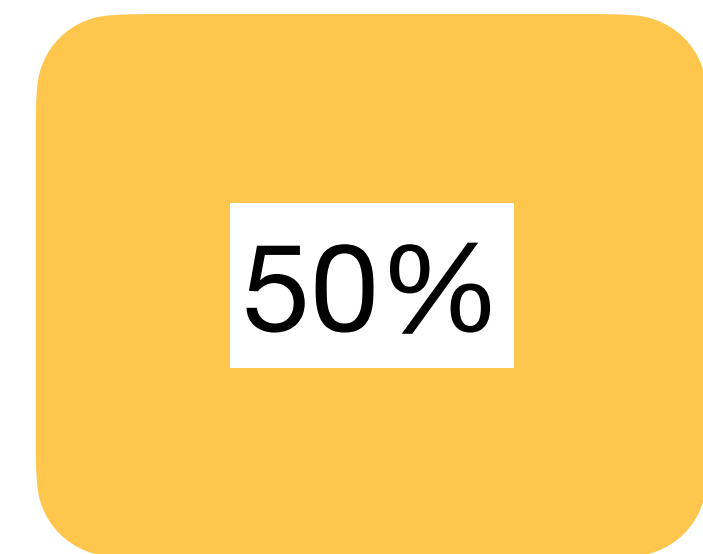
4



1

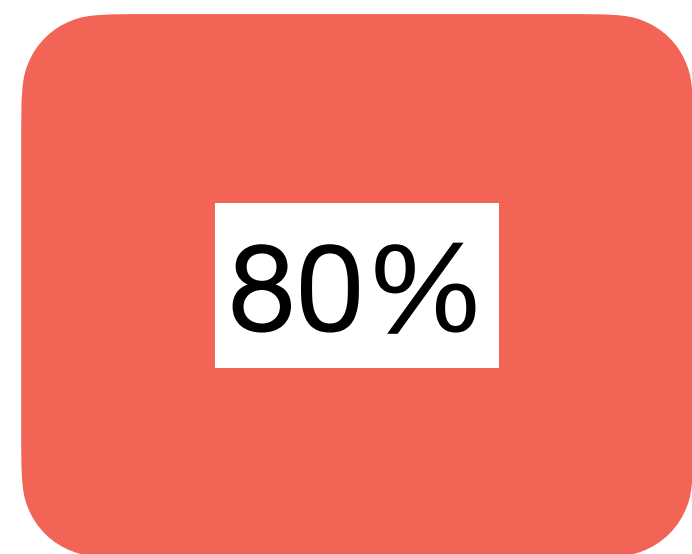
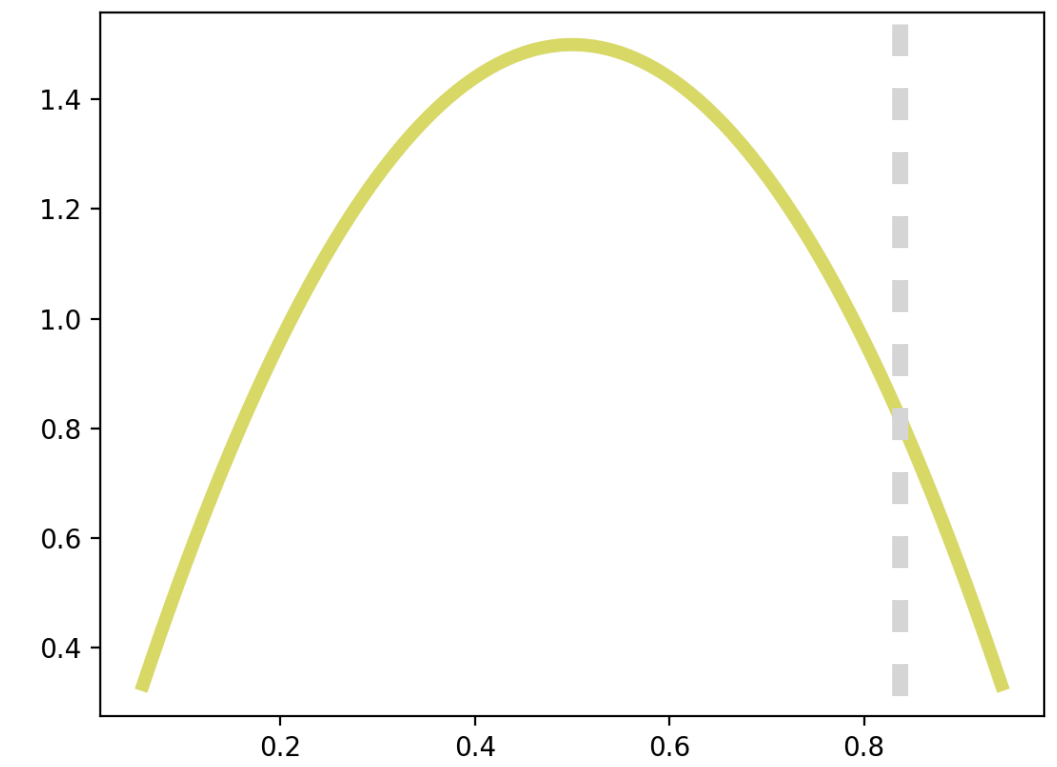
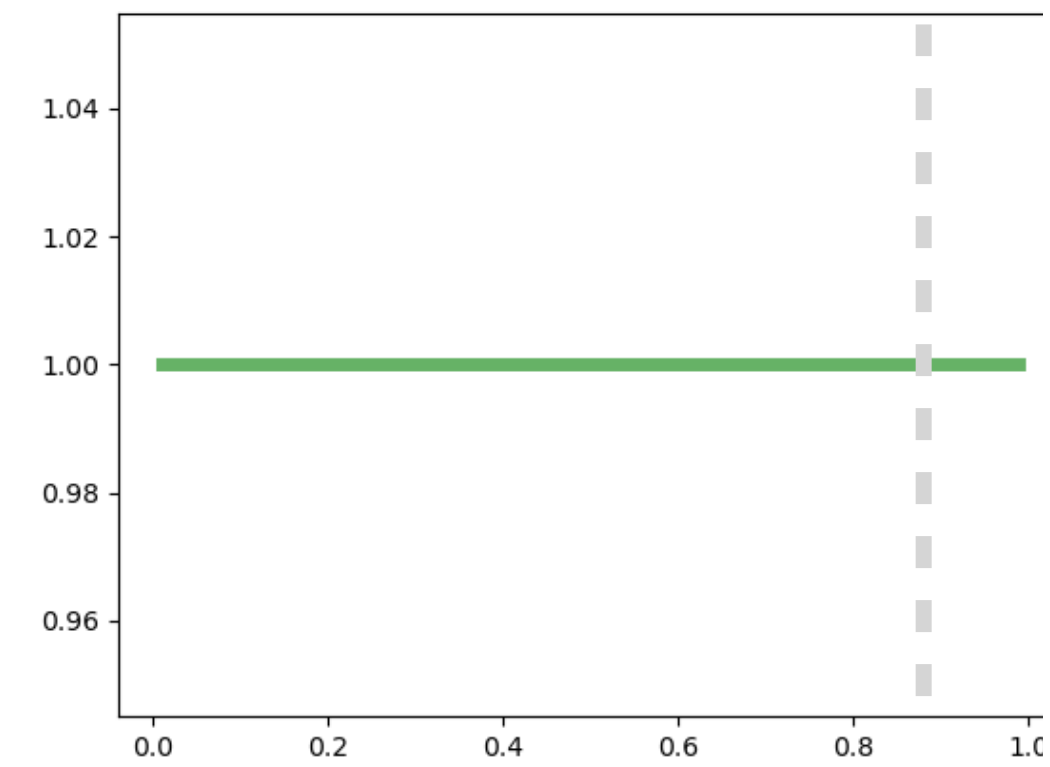
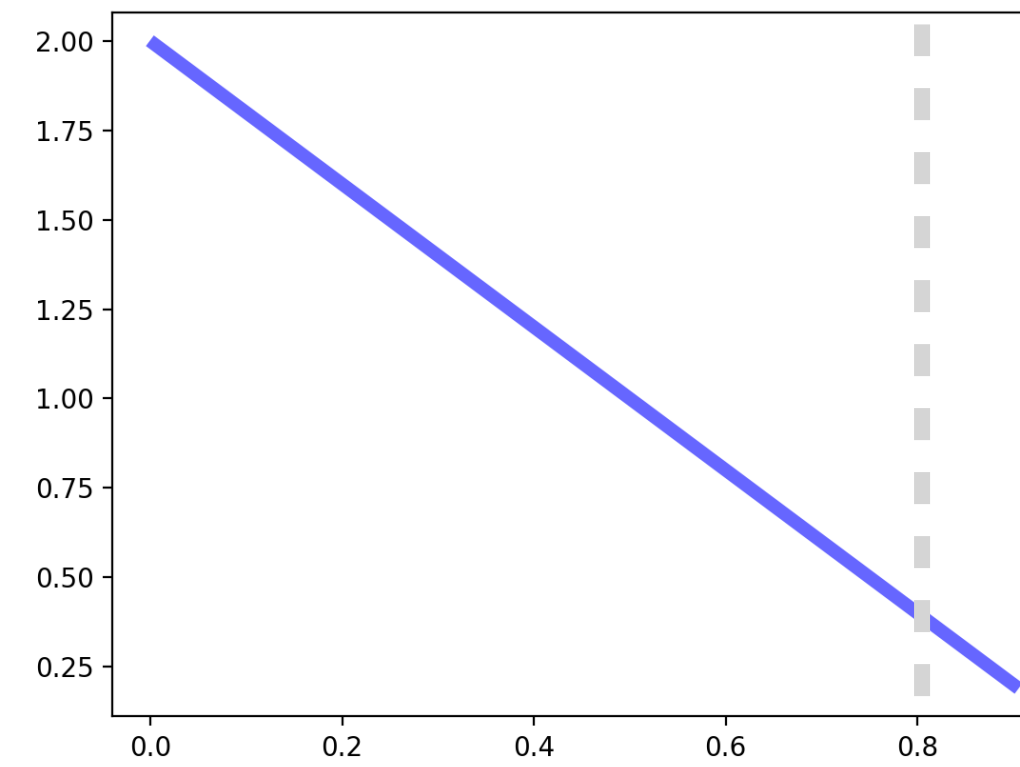
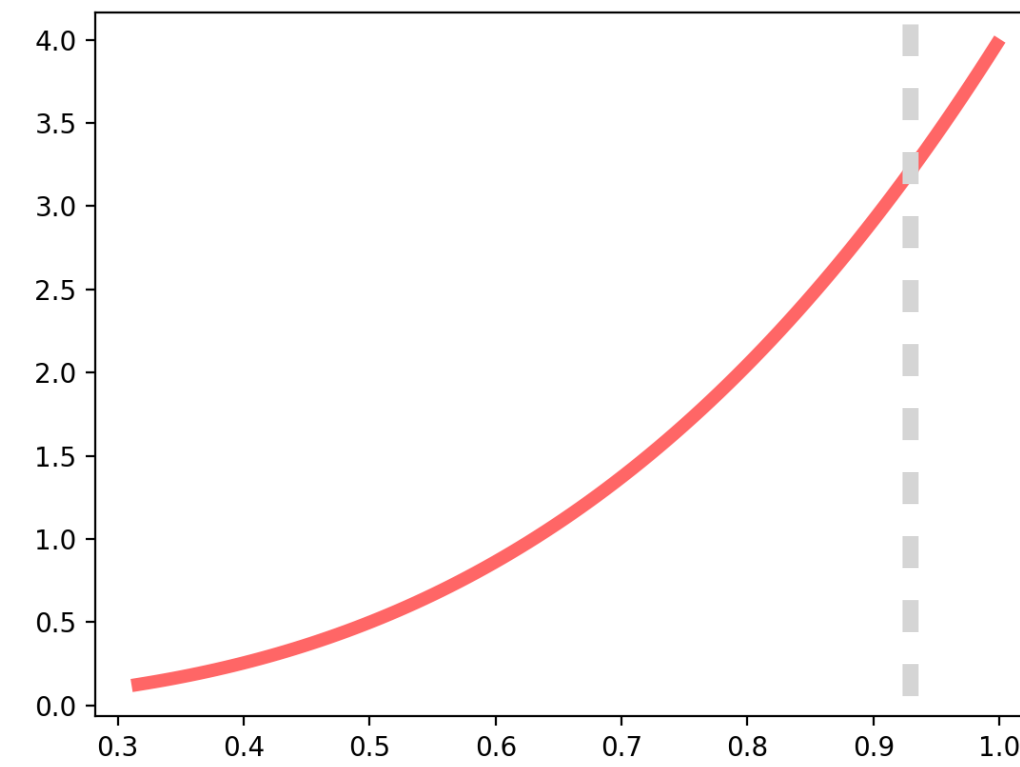


1

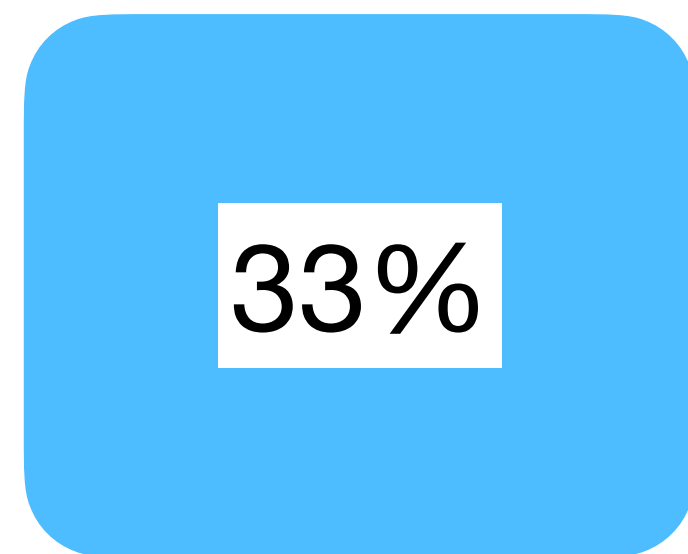


2

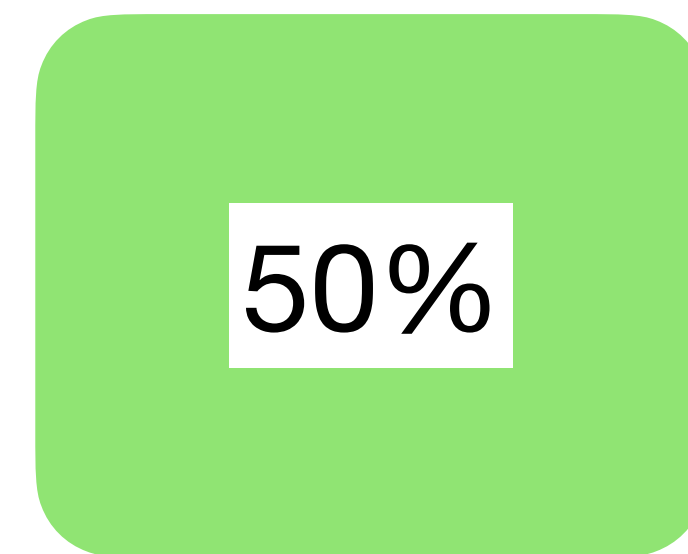
Sketching the Mechanism: Intuition



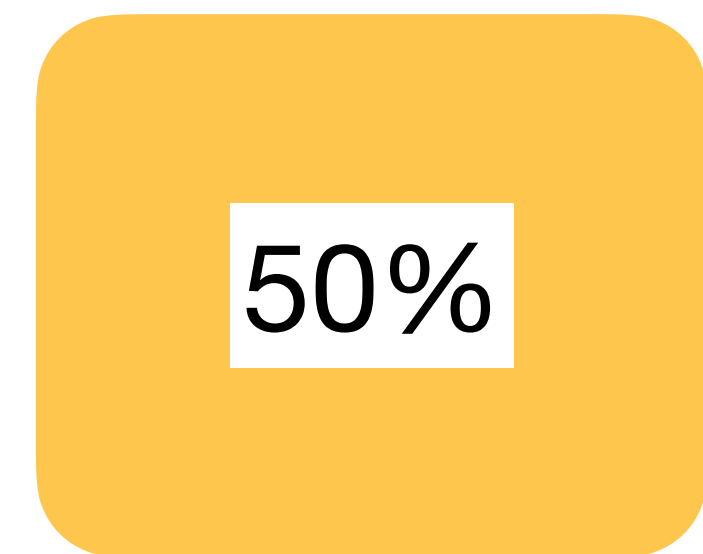
4



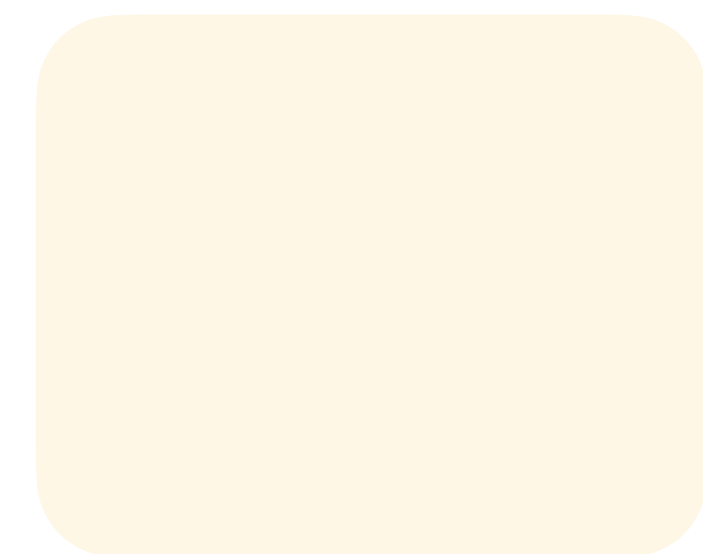
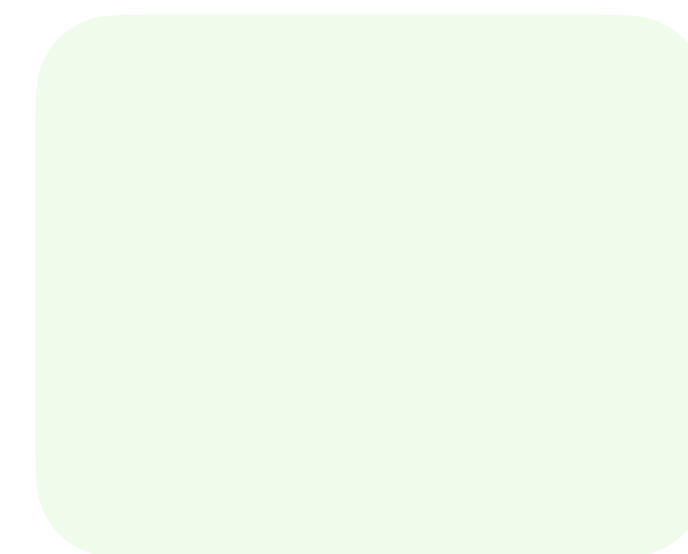
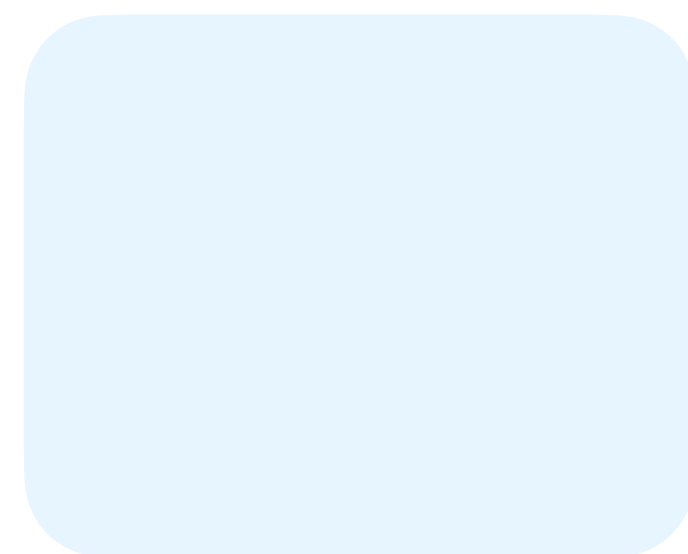
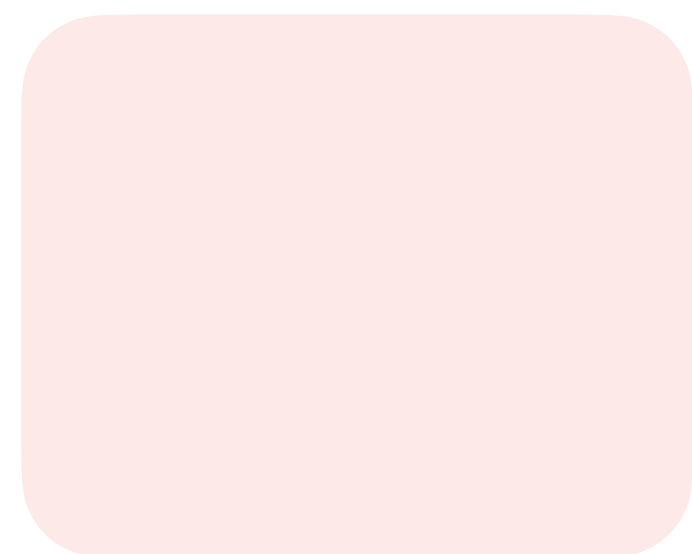
1



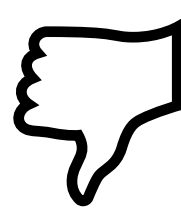
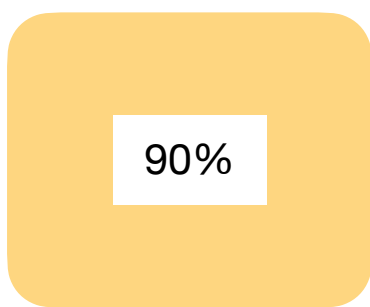
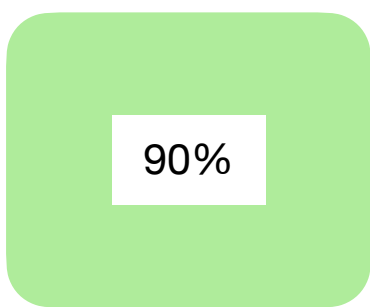
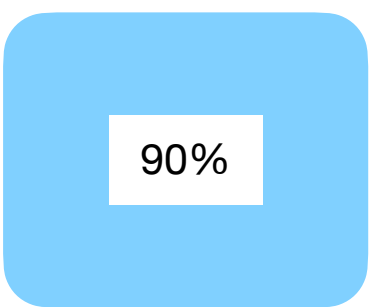
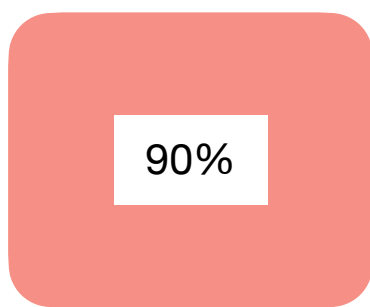
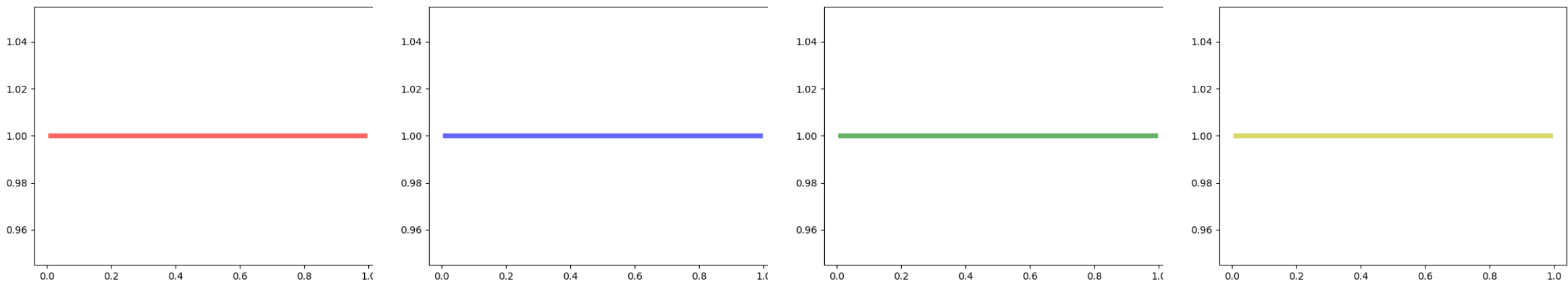
1



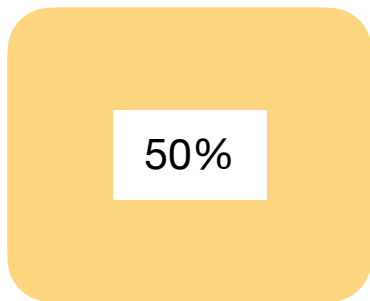
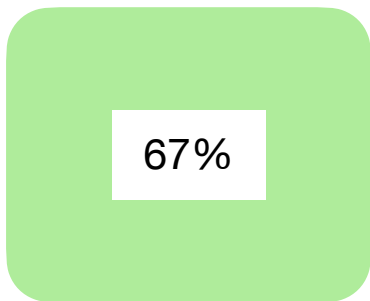
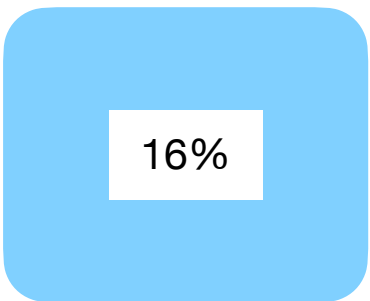
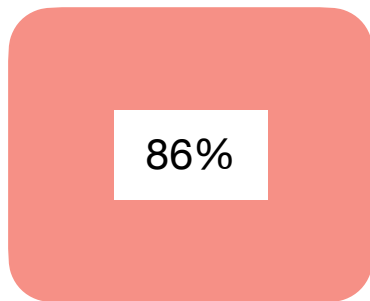
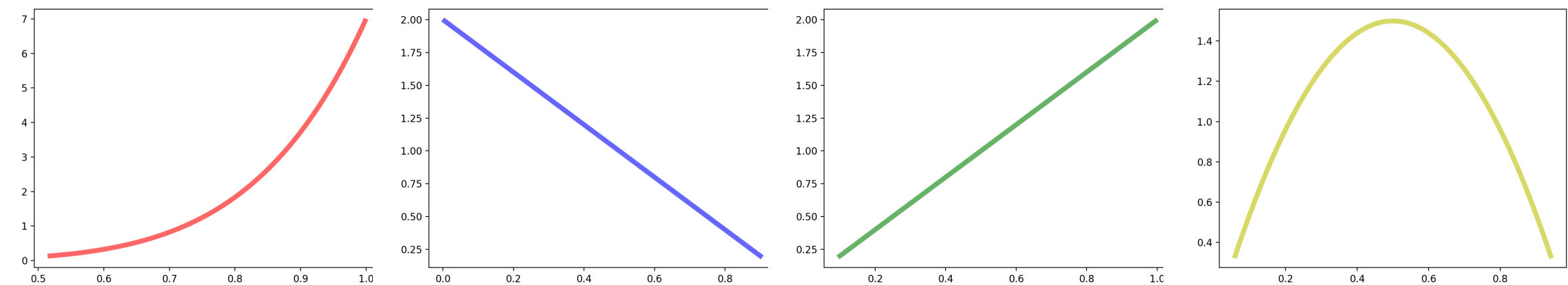
2



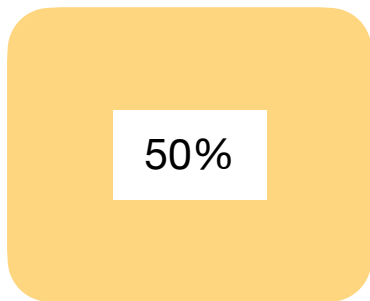
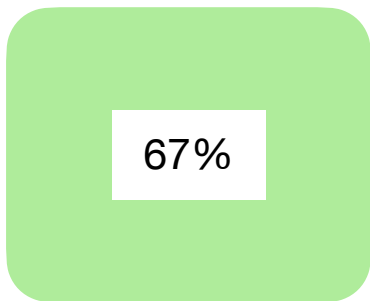
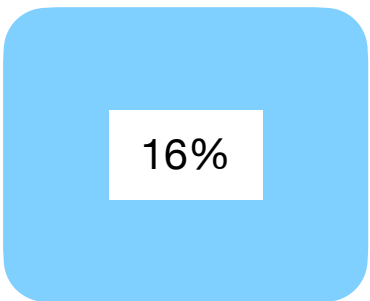
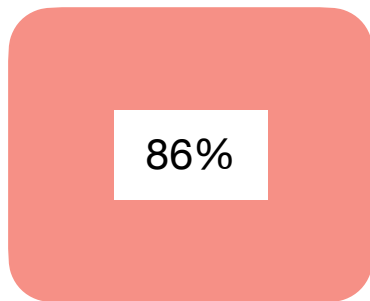
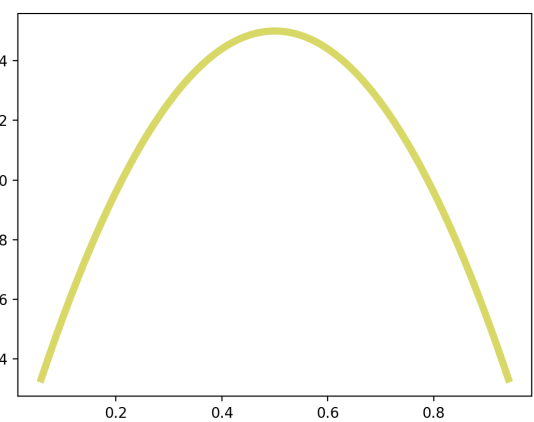
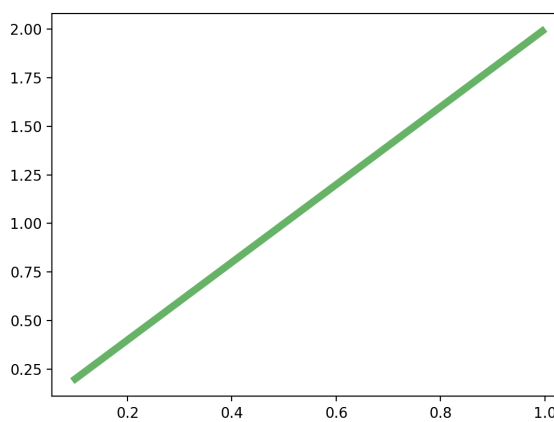
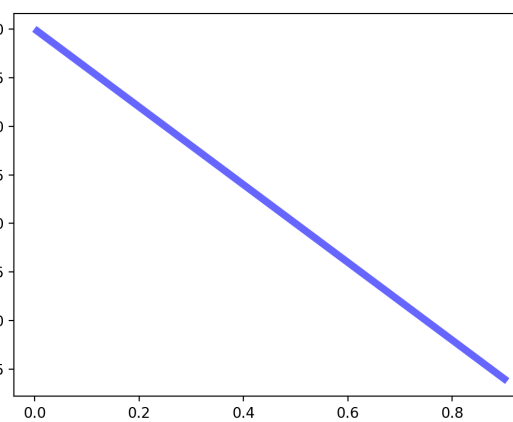
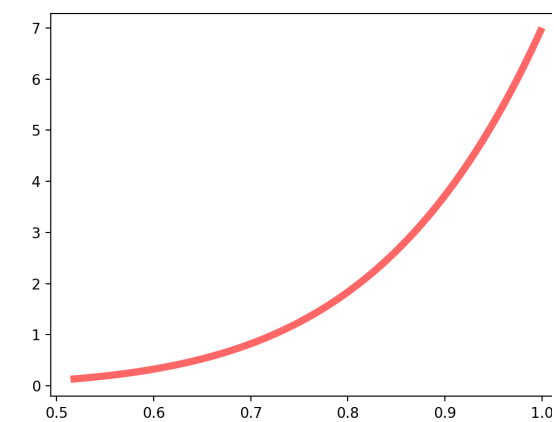
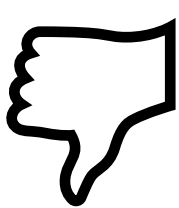
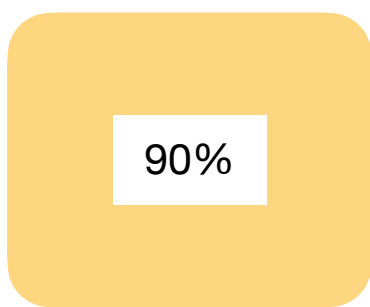
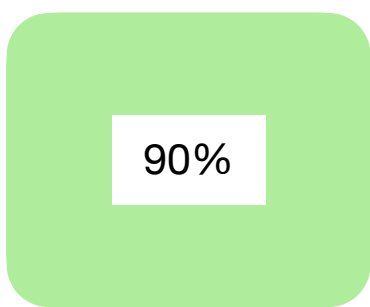
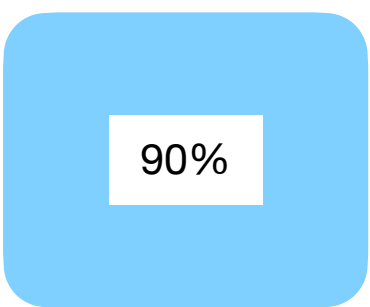
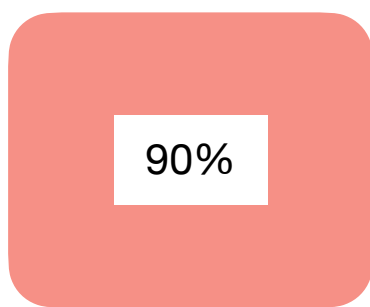
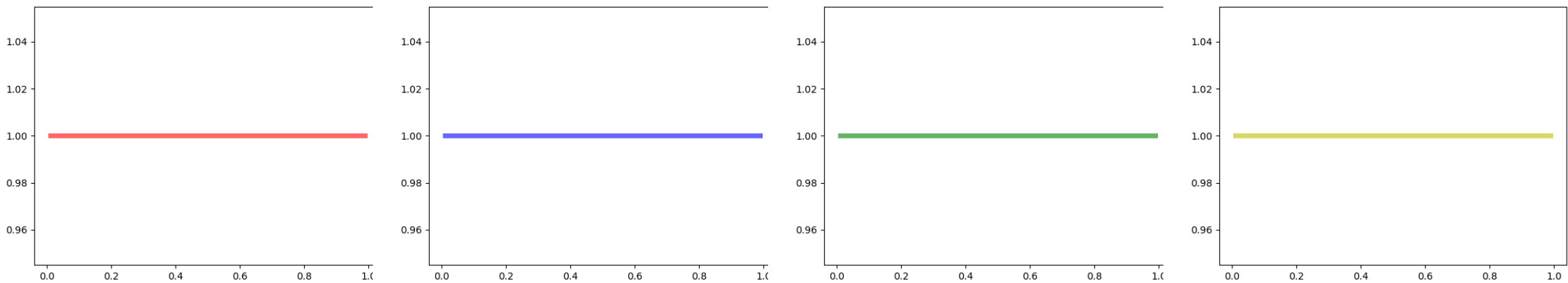
Sketching the Mechanism: Formalism



Multi-Armed Bandit Problem

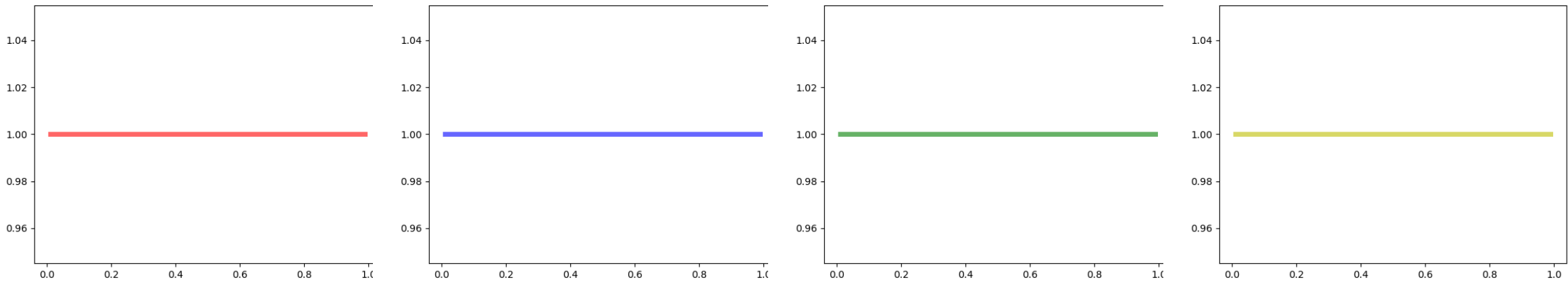


Sketching the Mechanism: Formalism



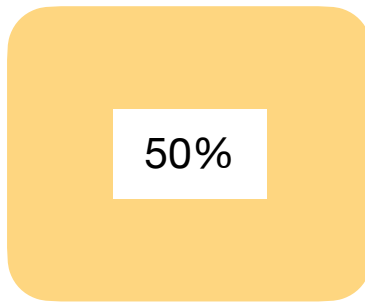
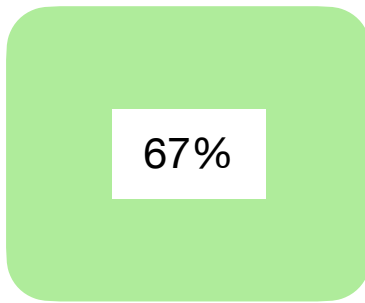
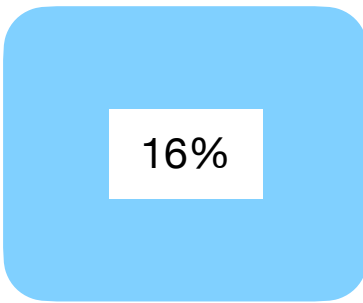
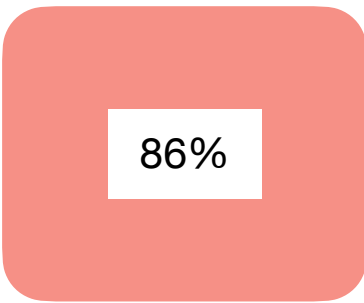
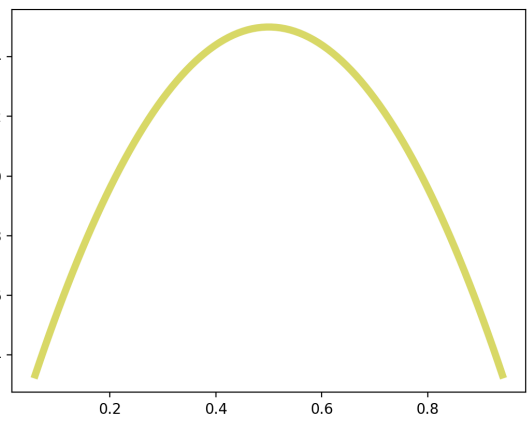
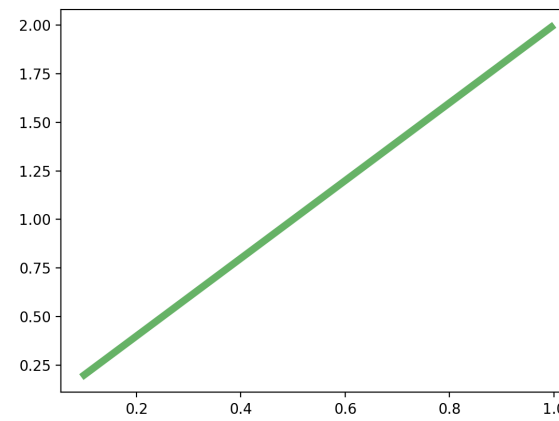
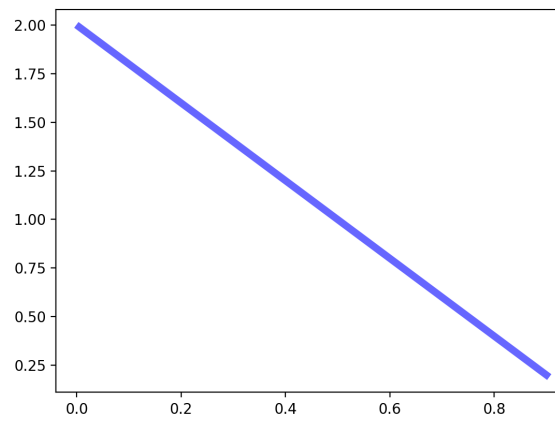
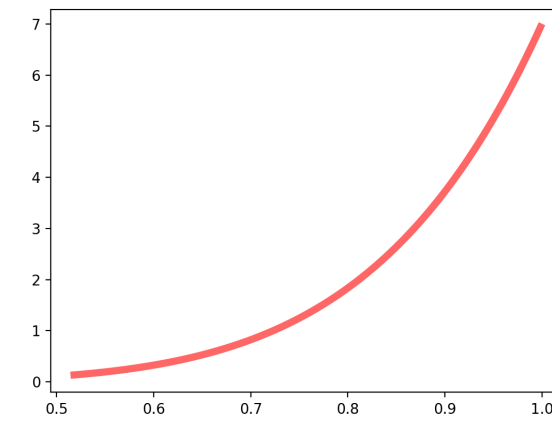
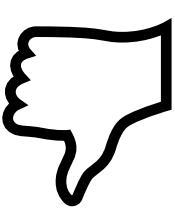
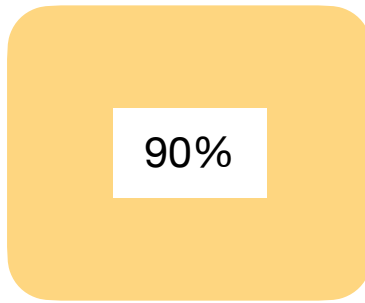
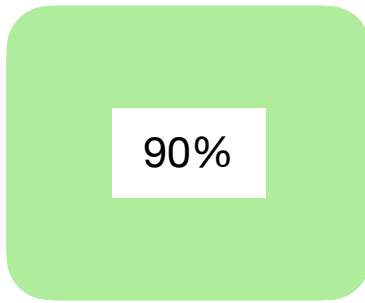
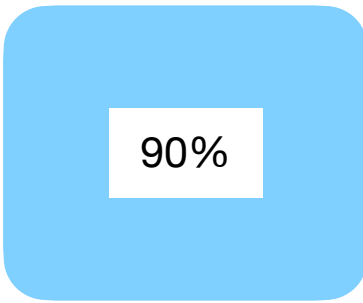
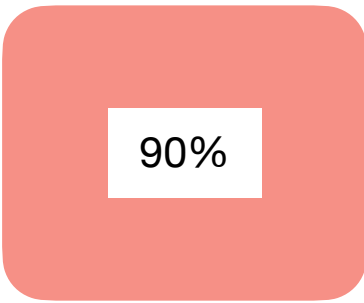
θ_k

Sketching the Mechanism: Formalism

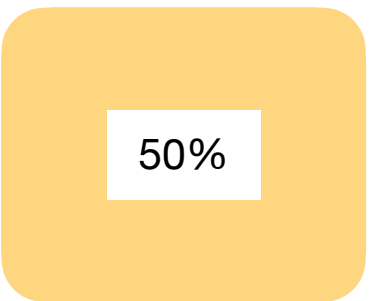
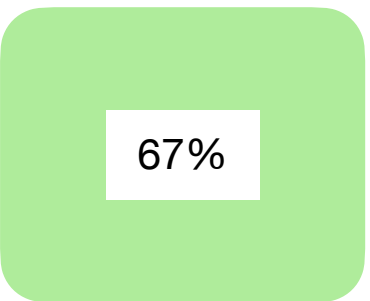
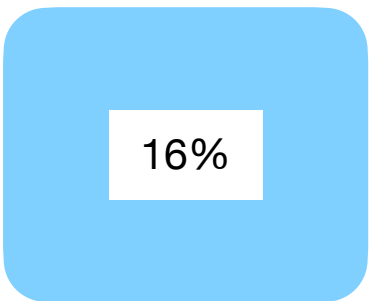
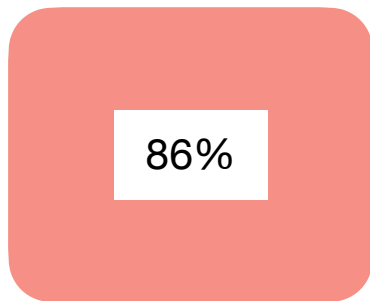
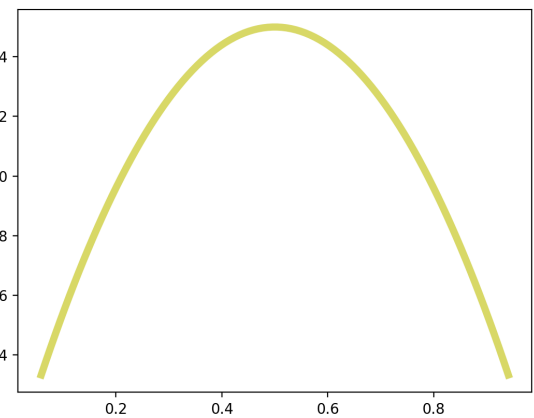
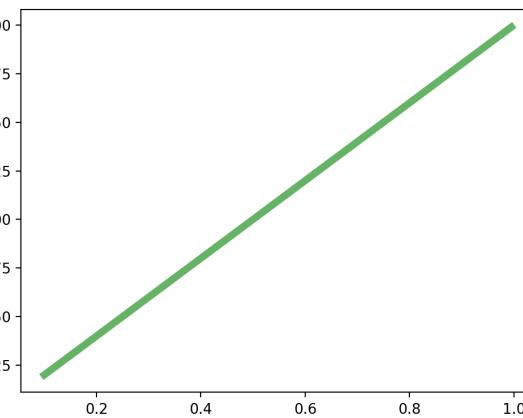
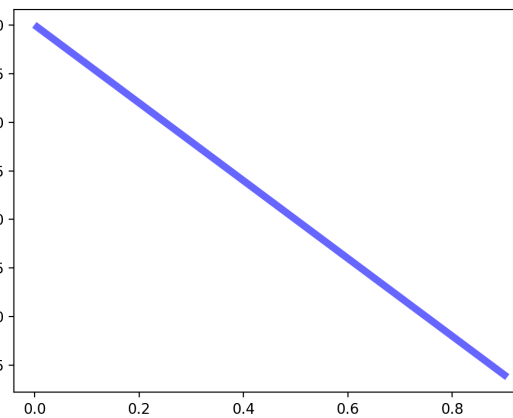
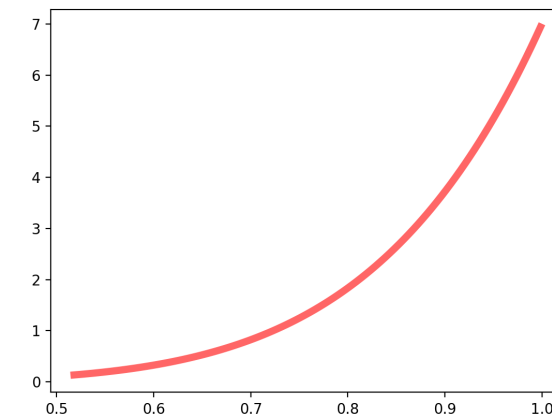
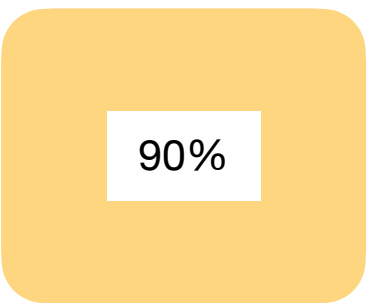
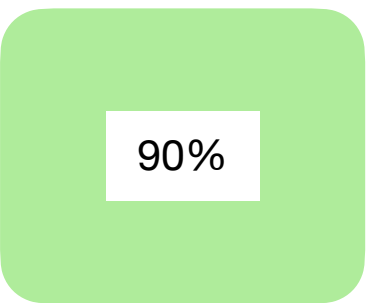
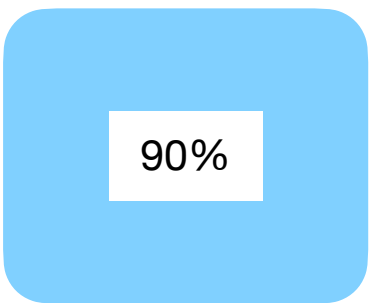
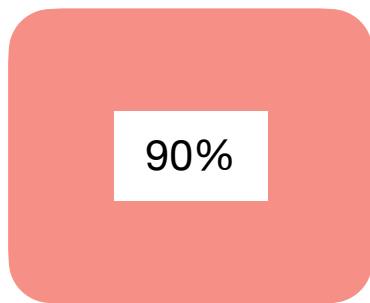
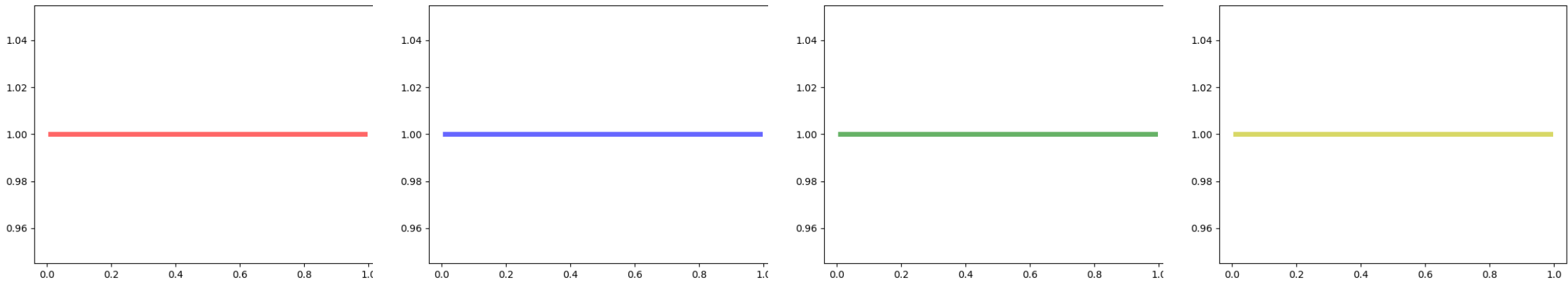


$$\theta_k$$

$$r_{t(k)} \sim \text{Bern}(\theta_k)$$



Sketching the Mechanism: **Formalism**

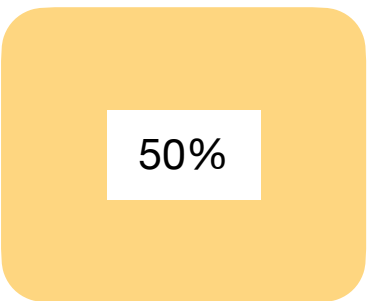
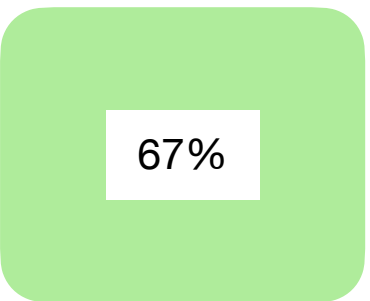
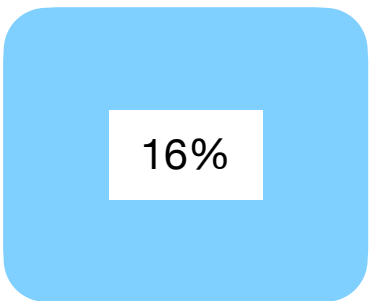
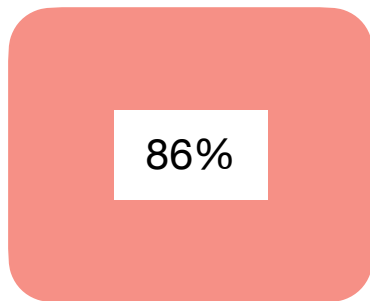
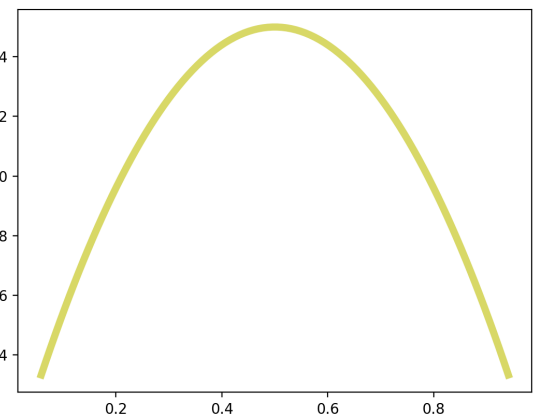
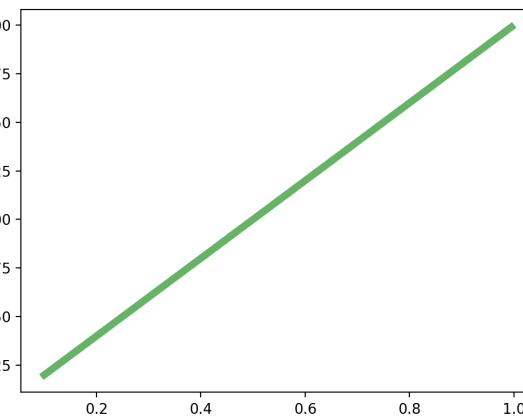
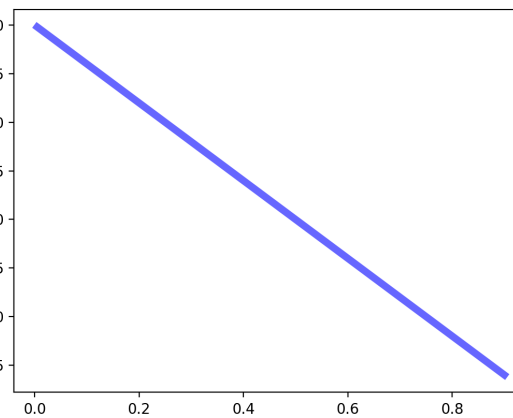
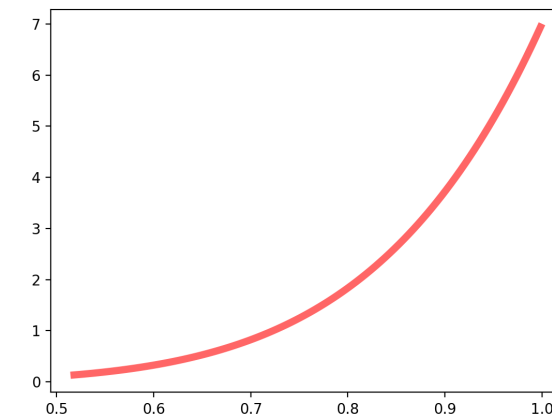
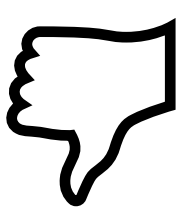
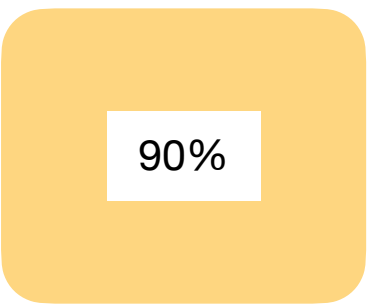
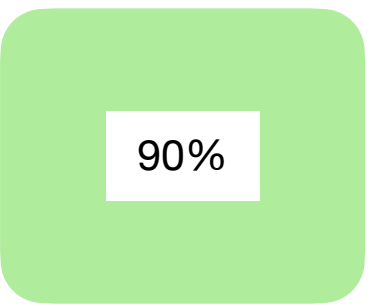
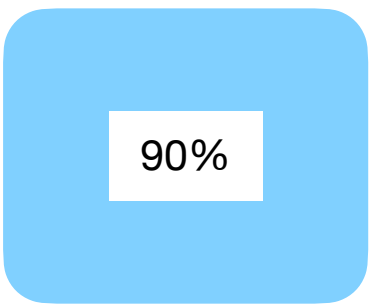
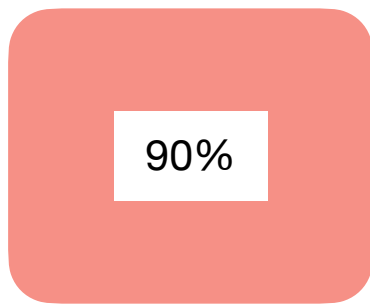
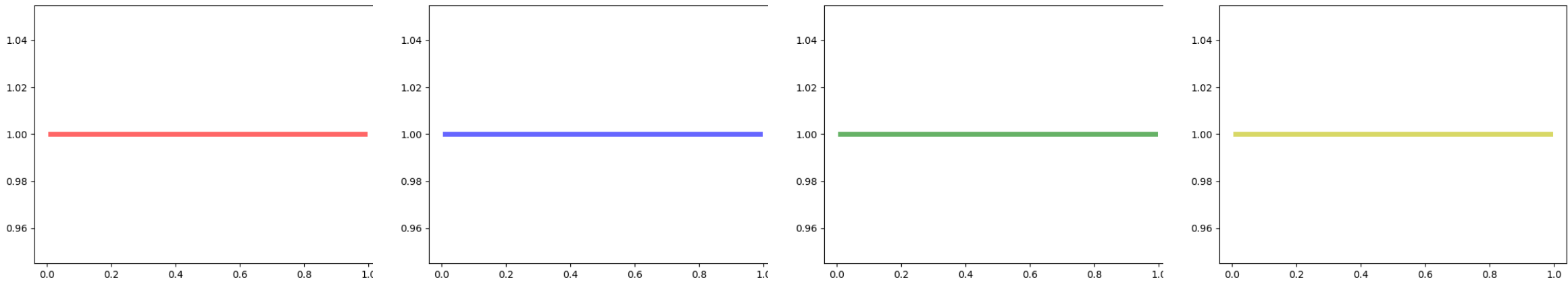


$$\theta_k$$

$$r_{t(k)} \sim \textit{Bern}(\theta_k)$$

$$\sum_{t=1}^T r_{t(k)}$$

Sketching the Mechanism: Formalism



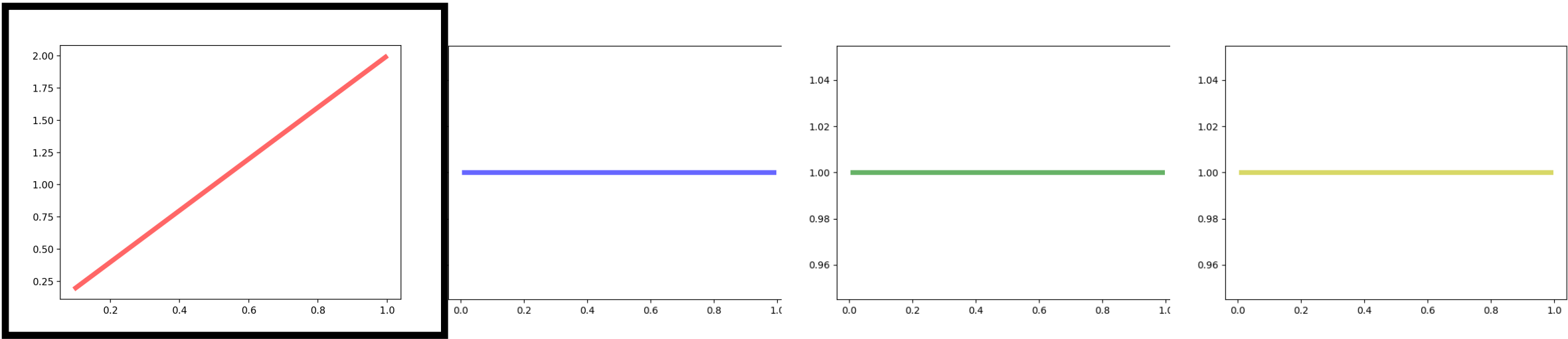
$$\theta_k$$

$$r_{t(k)} \sim \text{Bern}(\theta_k)$$

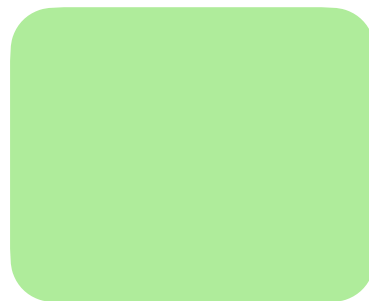
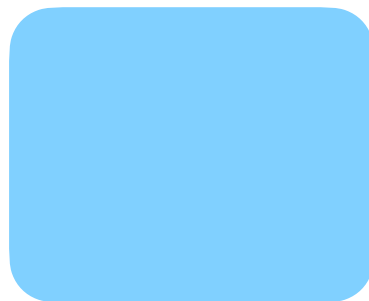
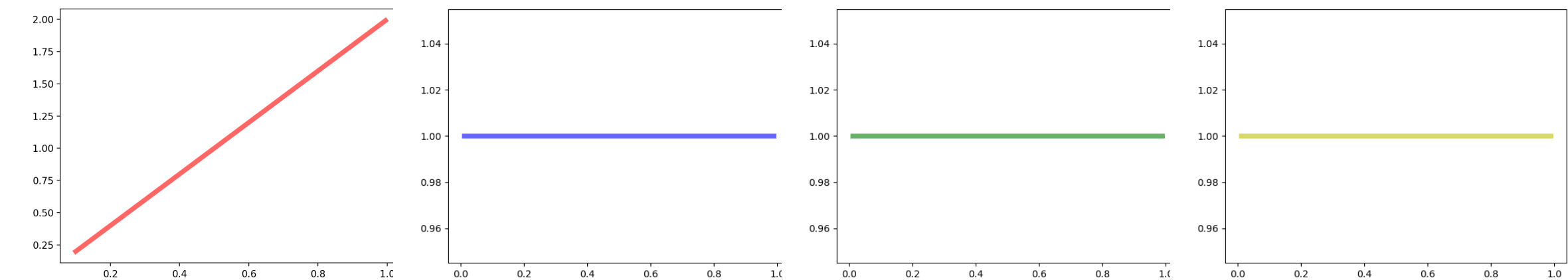
$$\sum_{t=1}^T r_{t(k)}$$

$$R = E[\sum_{t=1}^T Q_{t(k^*)} - \sum_{t=1}^T r_{t(k)}]$$

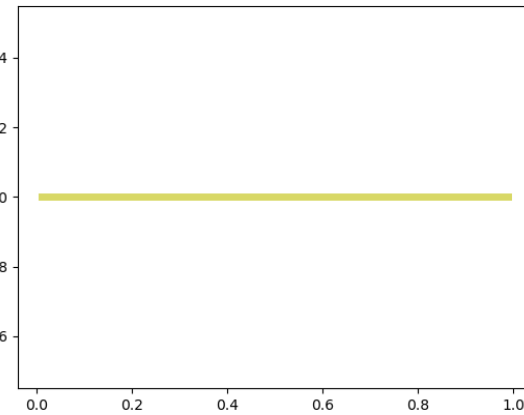
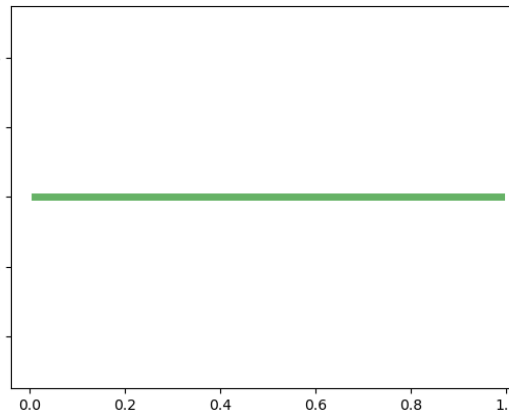
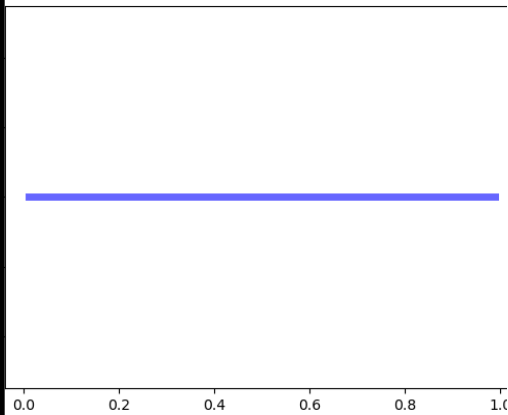
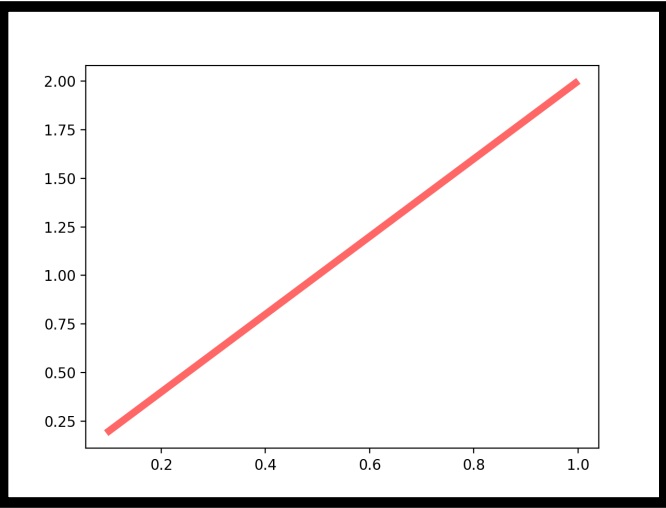
Sketching the Mechanism: Formalism



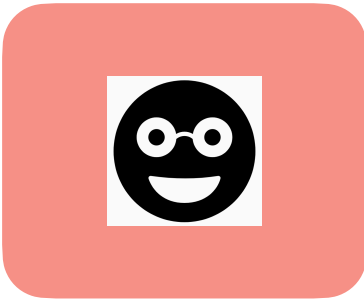
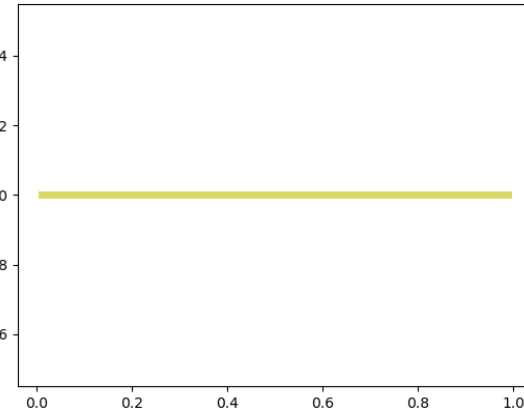
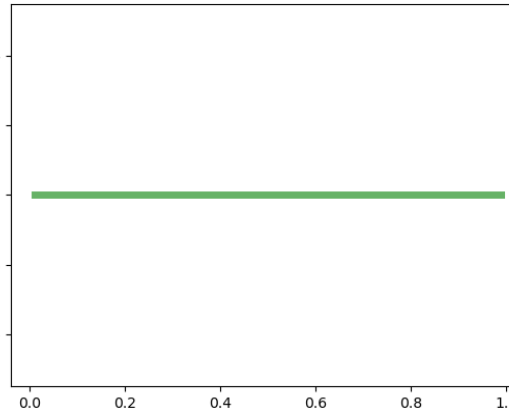
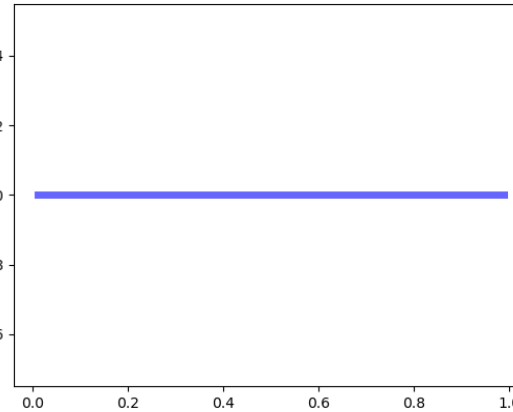
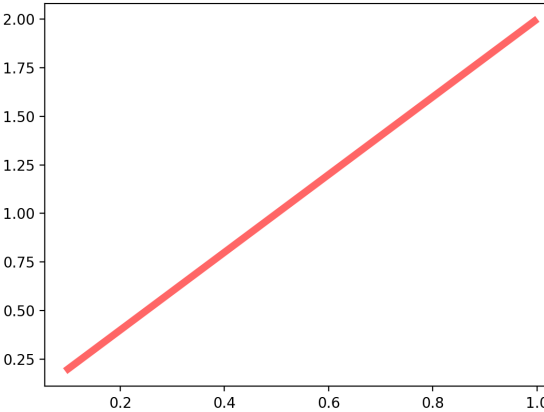
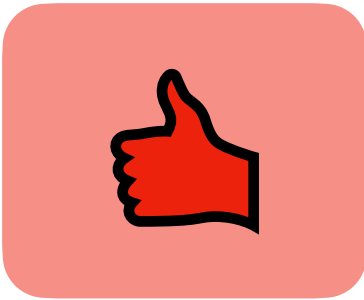
Thompson Sampling



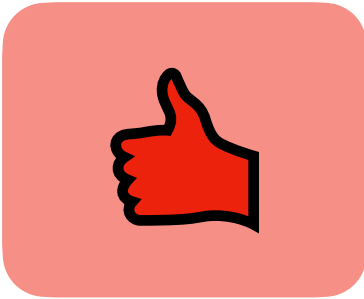
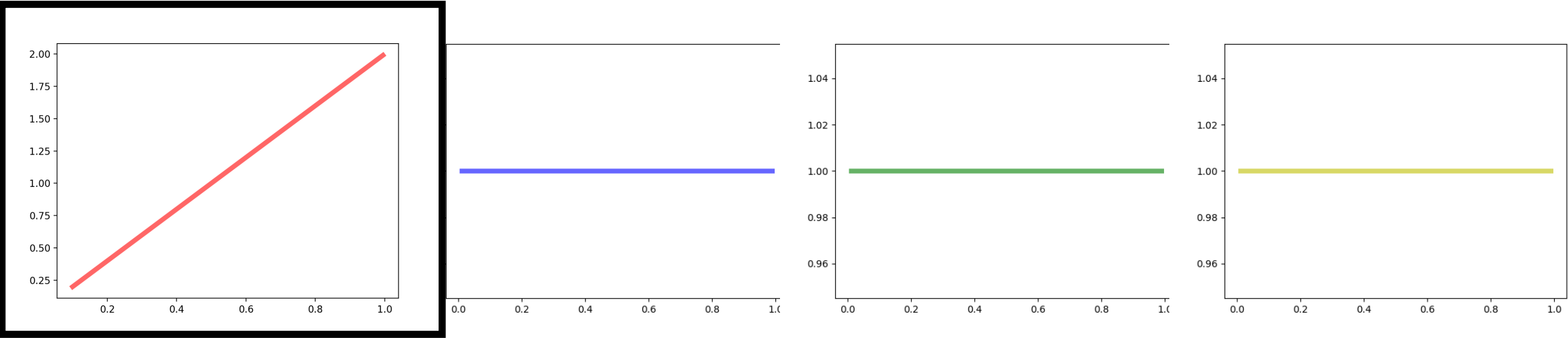
Sketching the Mechanism: Formalism



$Beta(\alpha, \beta)$

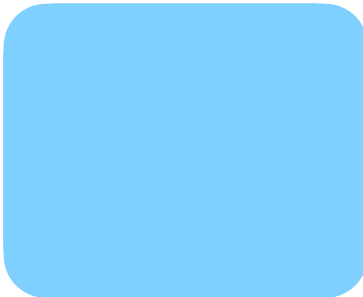
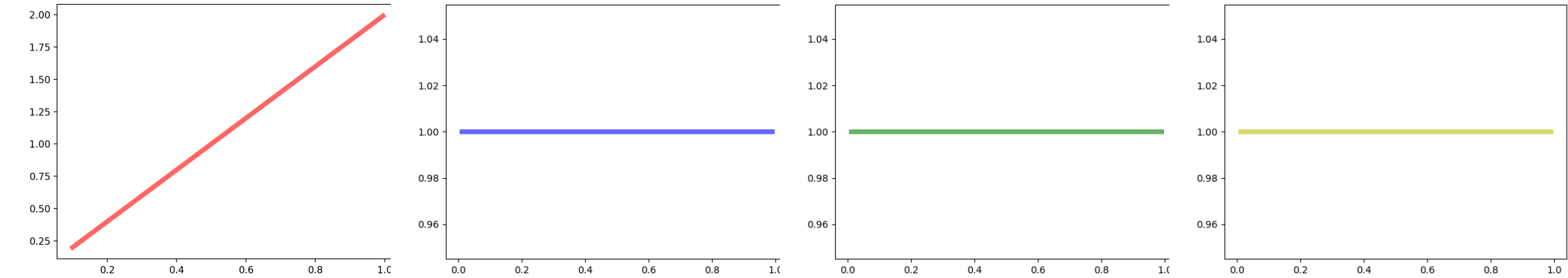


Sketching the Mechanism: Formalism

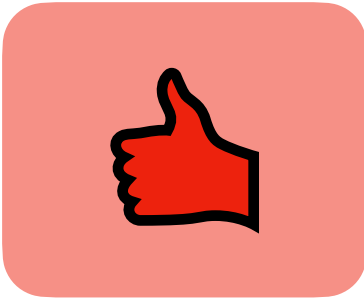
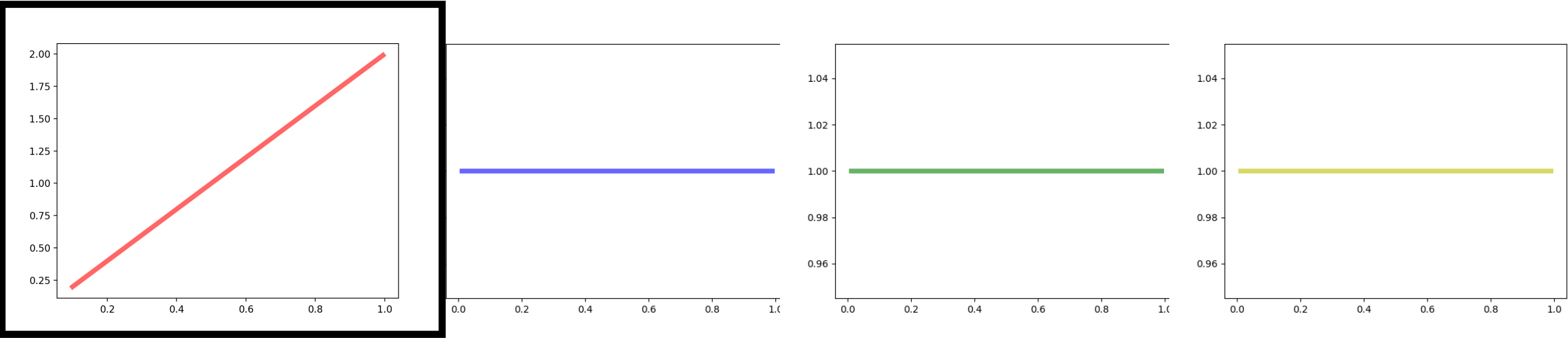


$Beta(\alpha, \beta)$

$\hat{\theta}_k \sim Beta(\alpha_k, \beta_k)$



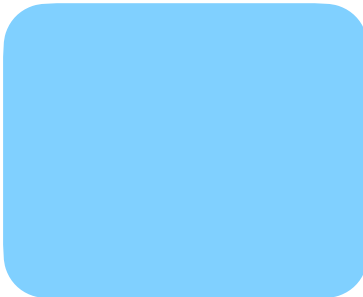
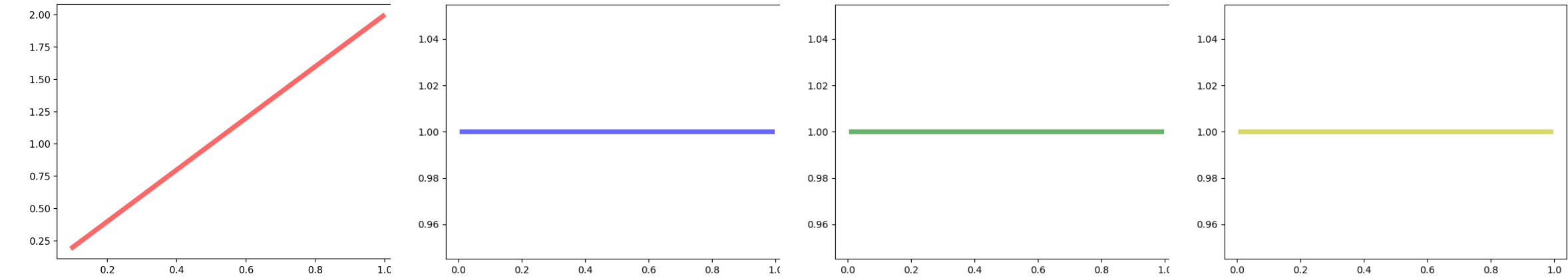
Sketching the Mechanism: **Formalism**



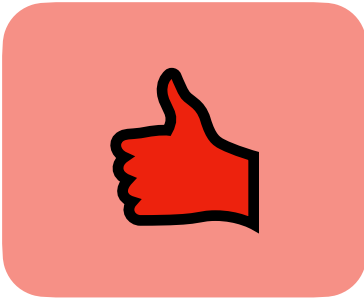
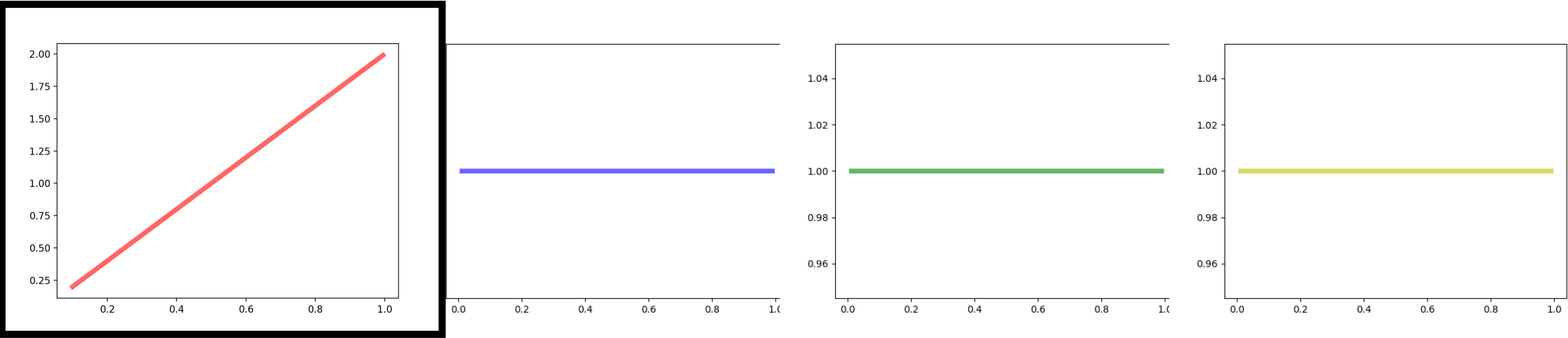
$$Beta(\alpha, \beta)$$

$$\hat{\theta}_k \sim Beta(\alpha_k, \beta_k)$$

$$x_{t(k)} \leftarrow argmax_k \hat{\theta}_k$$



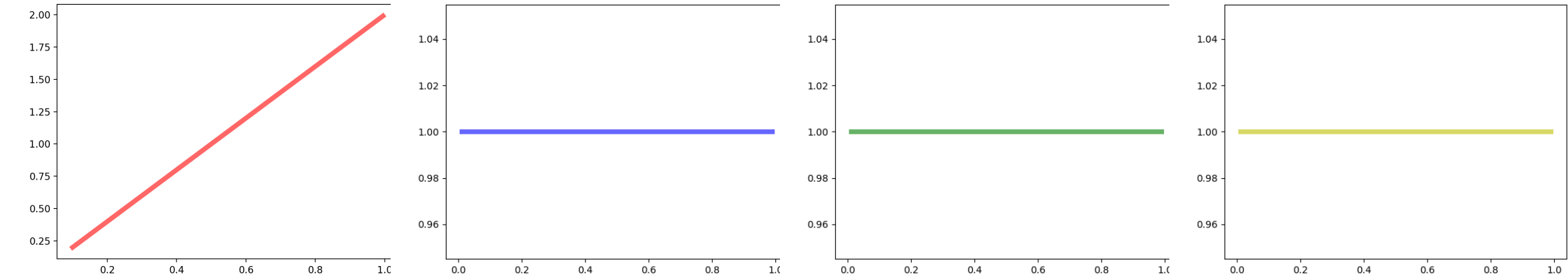
Sketching the Mechanism: **Formalism**



$Beta(\alpha, \beta)$

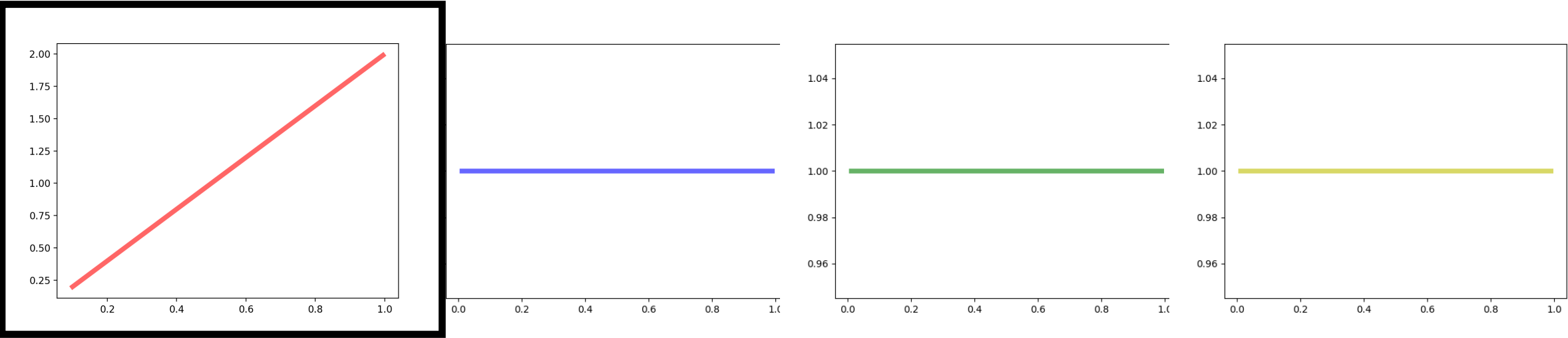
$\hat{\theta}_k \sim Beta(\alpha_k, \beta_k)$

$x_{t(k)} \leftarrow argmax_k \hat{\theta}_k$



Play with $x_{t(k)}$, observe $r_{t(k)}$

Sketching the Mechanism: **Formalism**



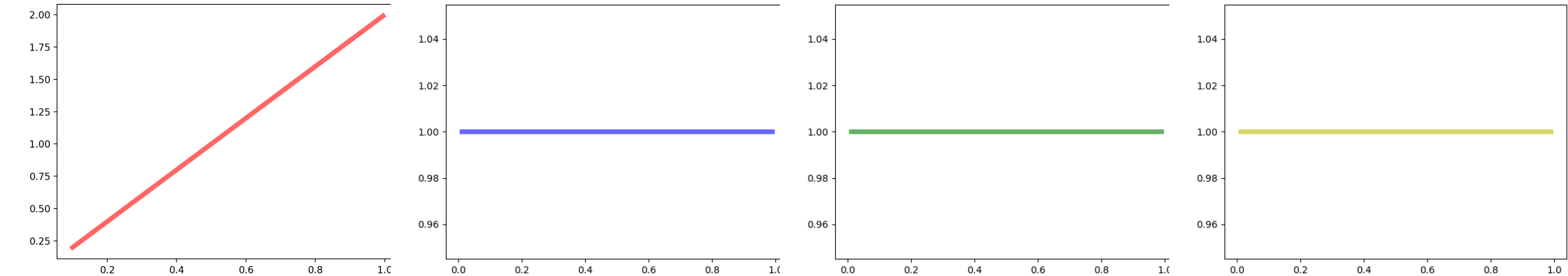
$$Beta(\alpha, \beta)$$

$$\hat{\theta}_k \sim Beta(\alpha_k, \beta_k)$$

$$x_{t(k)} \leftarrow argmax_k \hat{\theta}_k$$

Play with $x_{t(k)}$, observe $r_{t(k)}$

$$(\alpha_{xt}, \beta_{xt}) \leftarrow (\alpha_{xt} + r_t, \beta_{xt} + 1 - r_t)$$

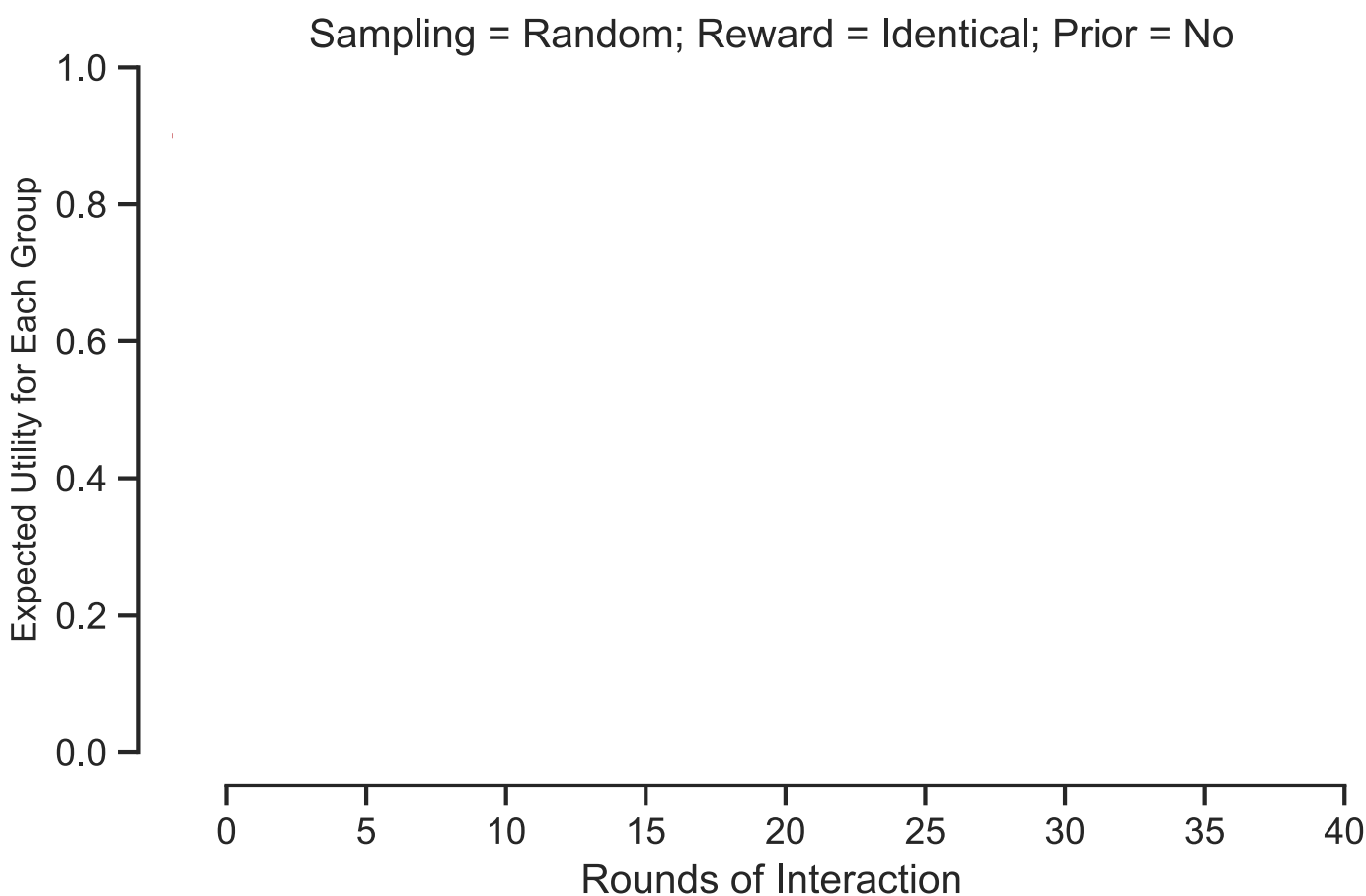
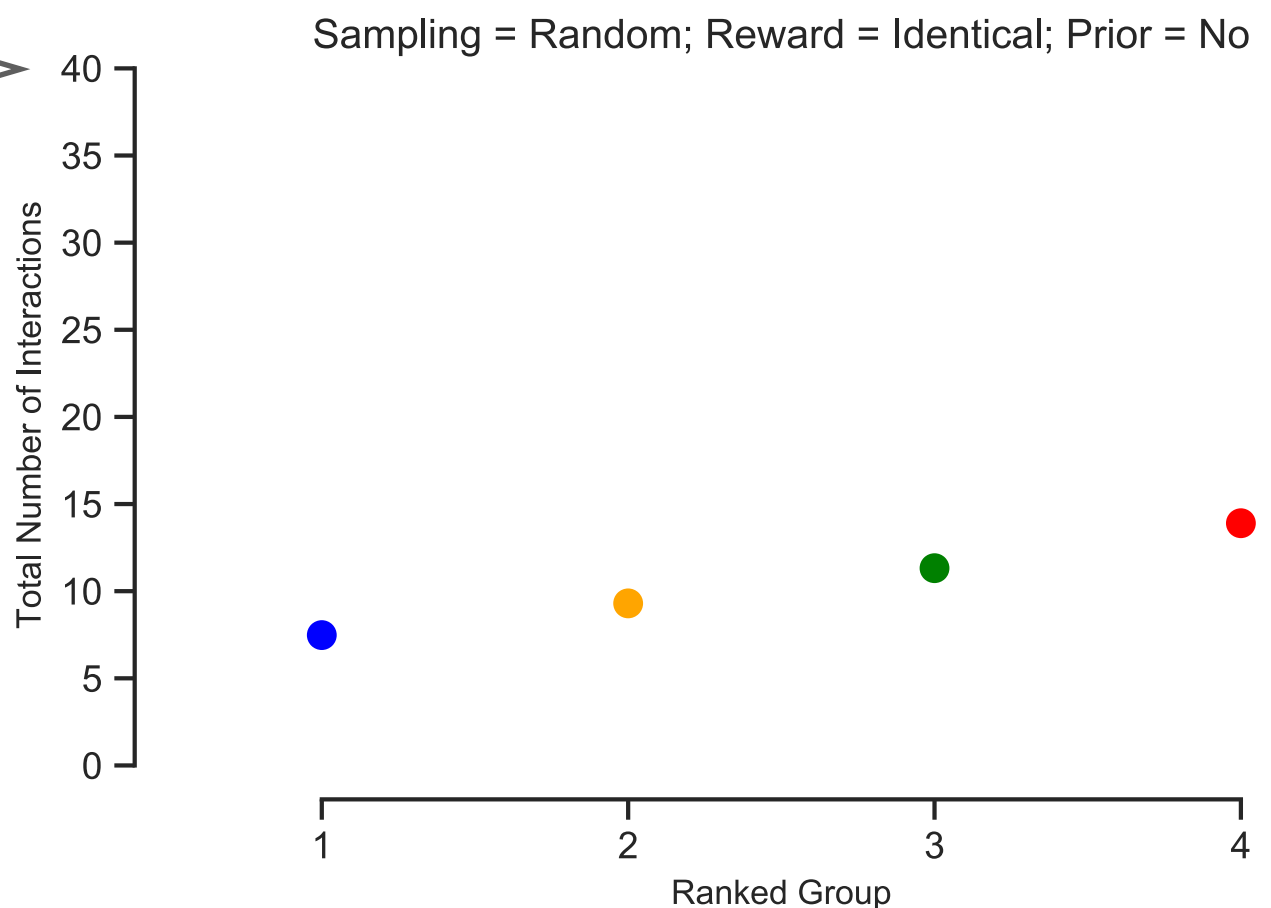
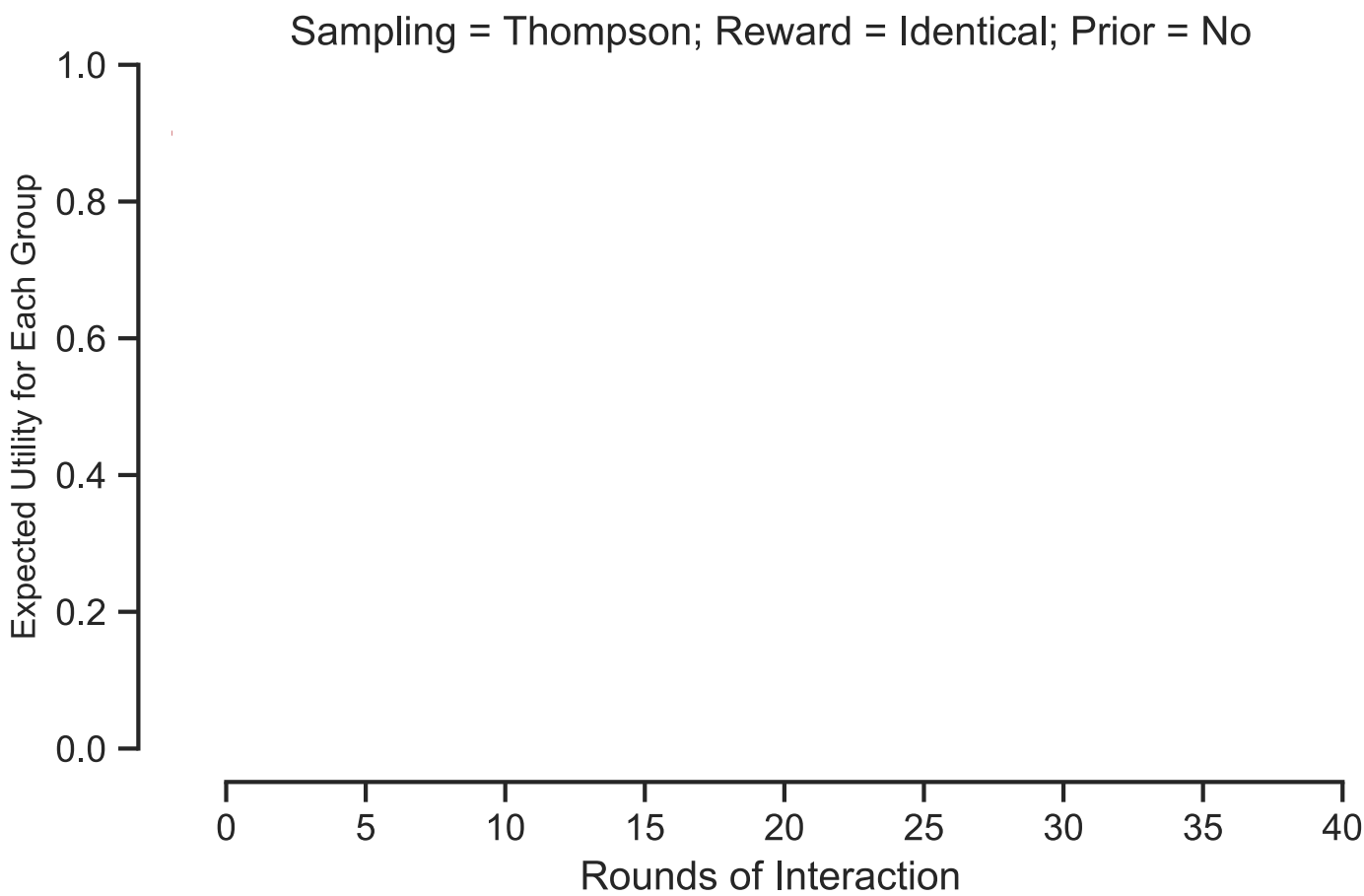
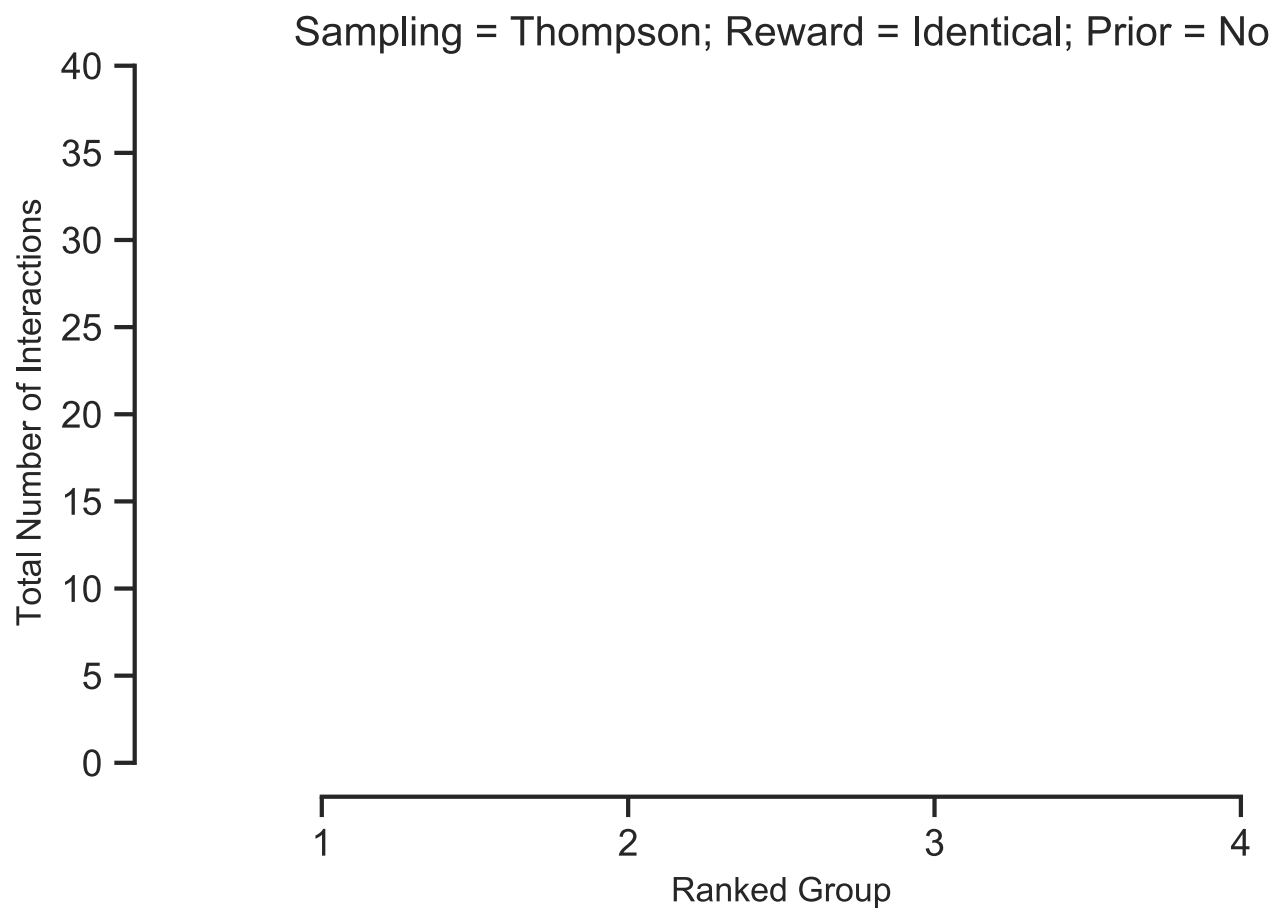


Sketching the Mechanism: **Hypothesis**

Control group: random sampling

- Random-sampling agent, prior beta (1,1), 4 arms, identical rewards ($\theta=0.9$), 40 rounds, 100 seeds.

How many times did the agent interact with each group?

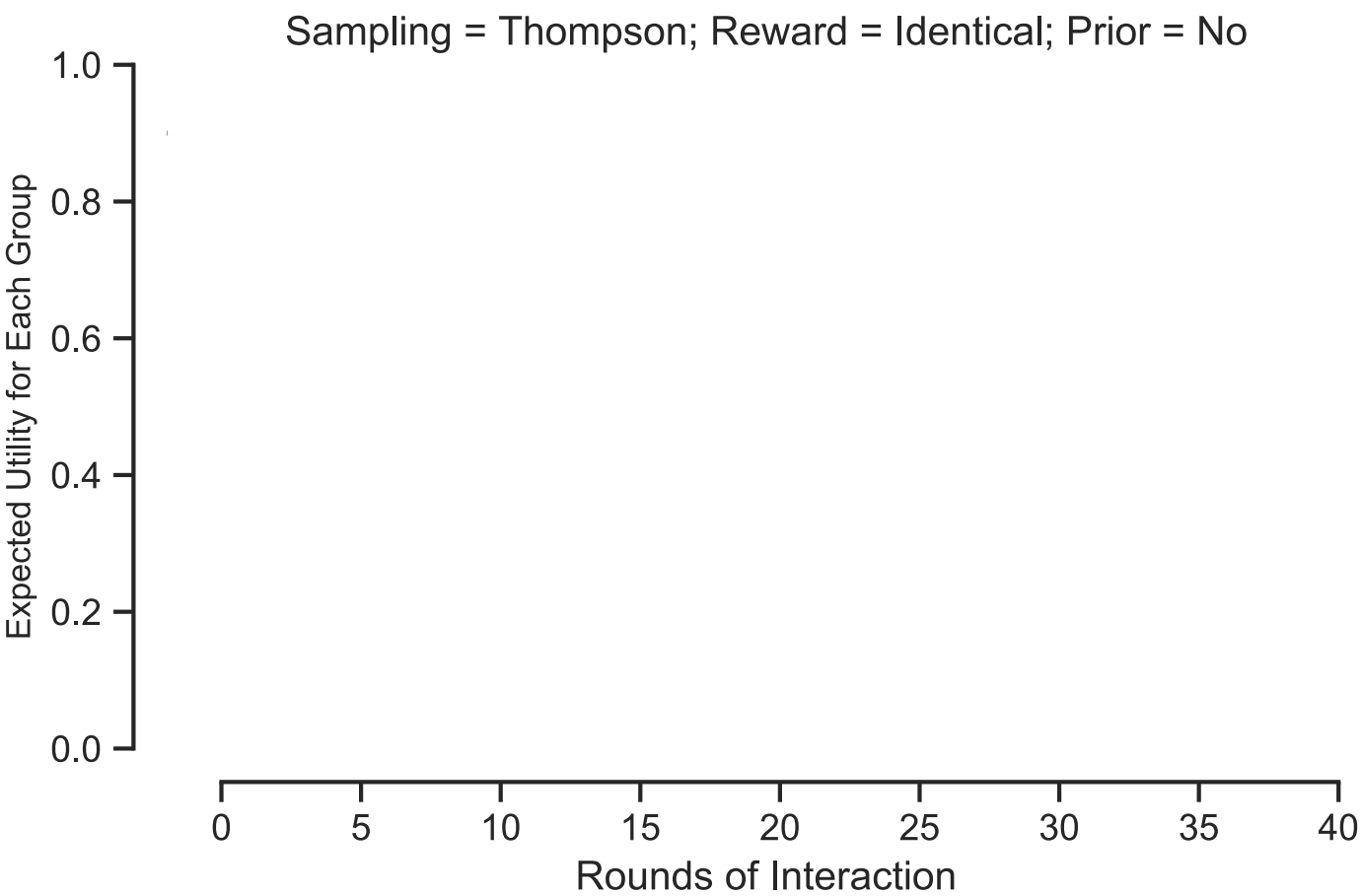
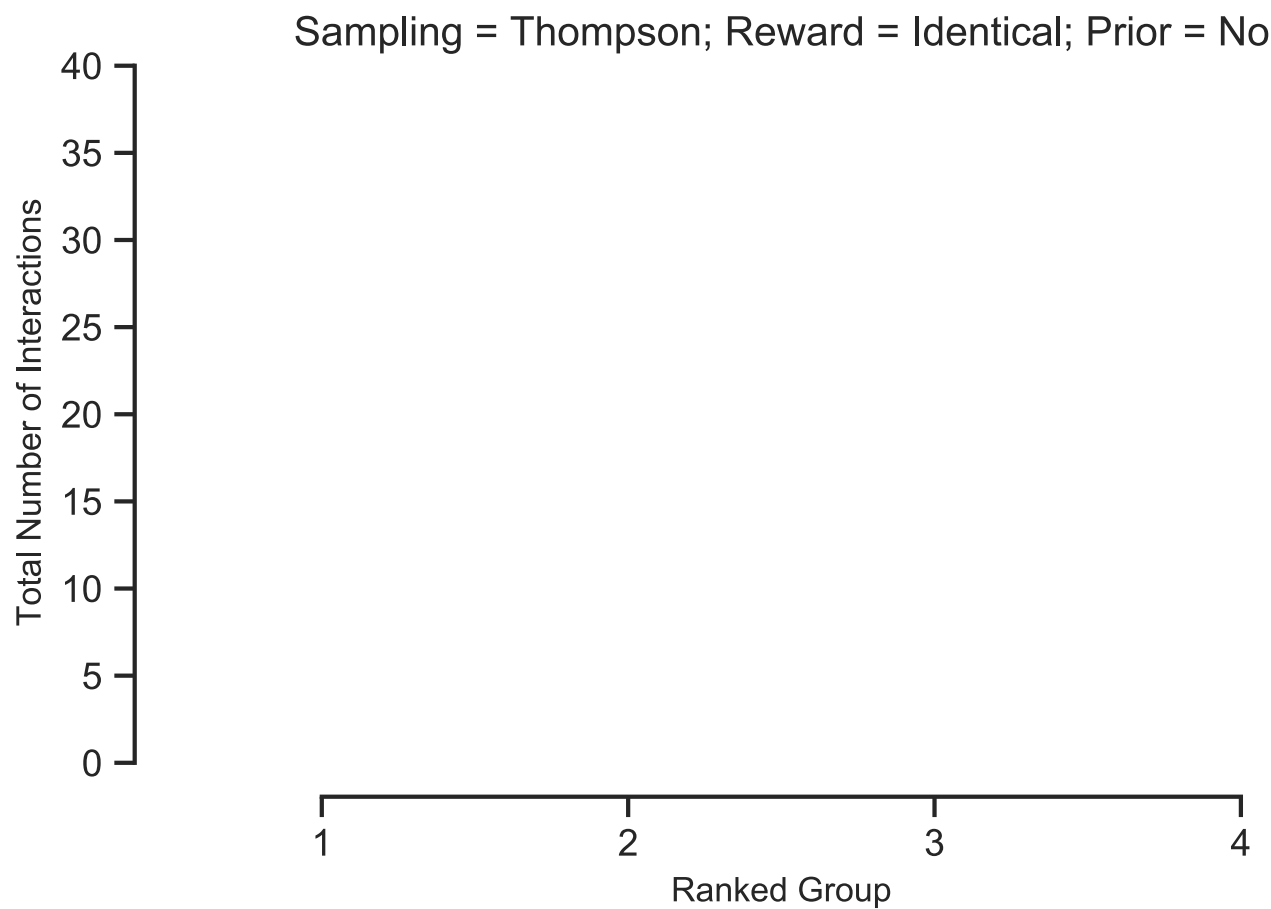


Sketching the Mechanism: Hypothesis

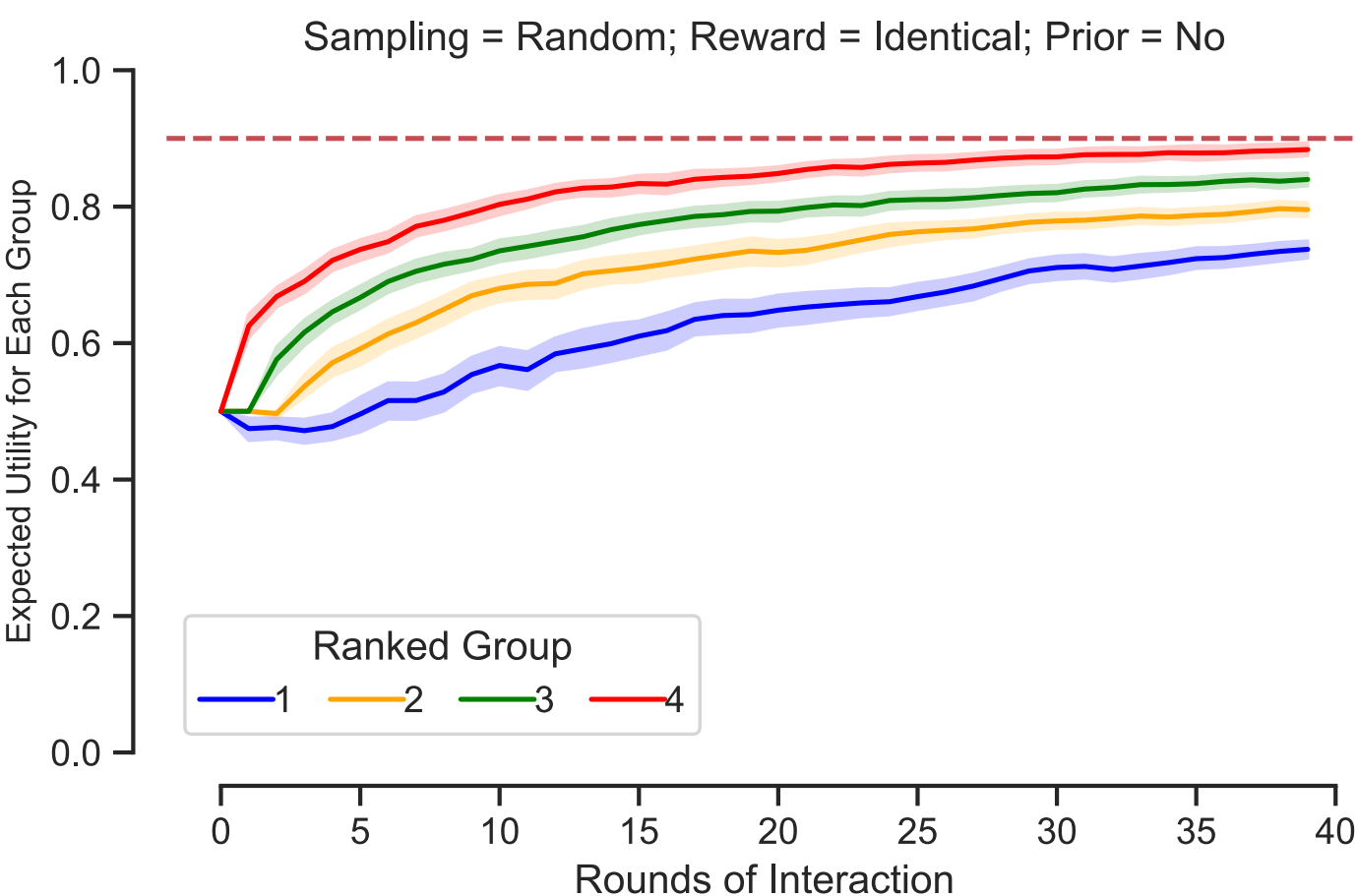
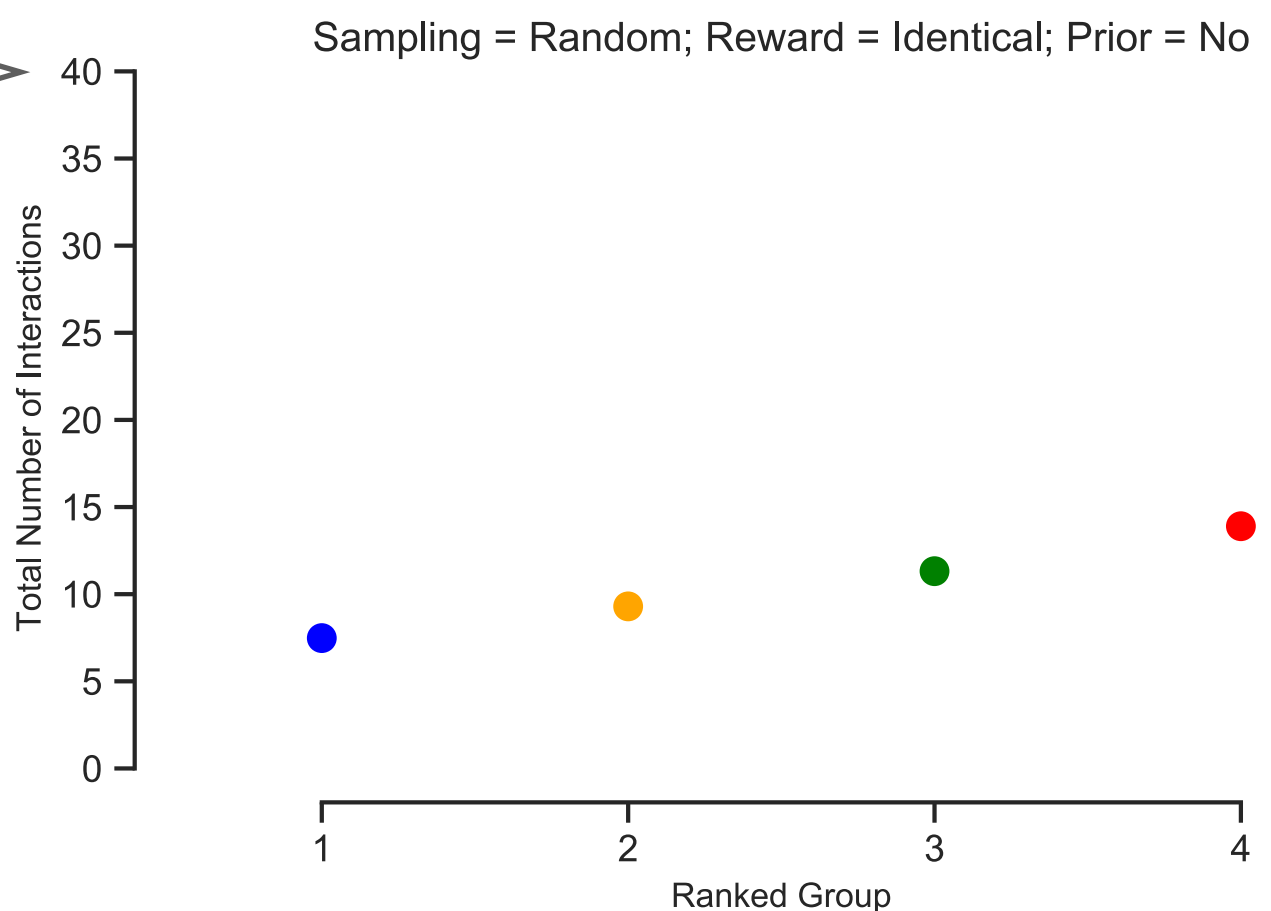
Control group: random sampling

- Random-sampling agent, prior beta (1,1), 4 arms, identical rewards ($\theta=0.9$), 40 rounds, 100 seeds.

How many times did the agent interact with each group?



What is the estimated reward for each group?

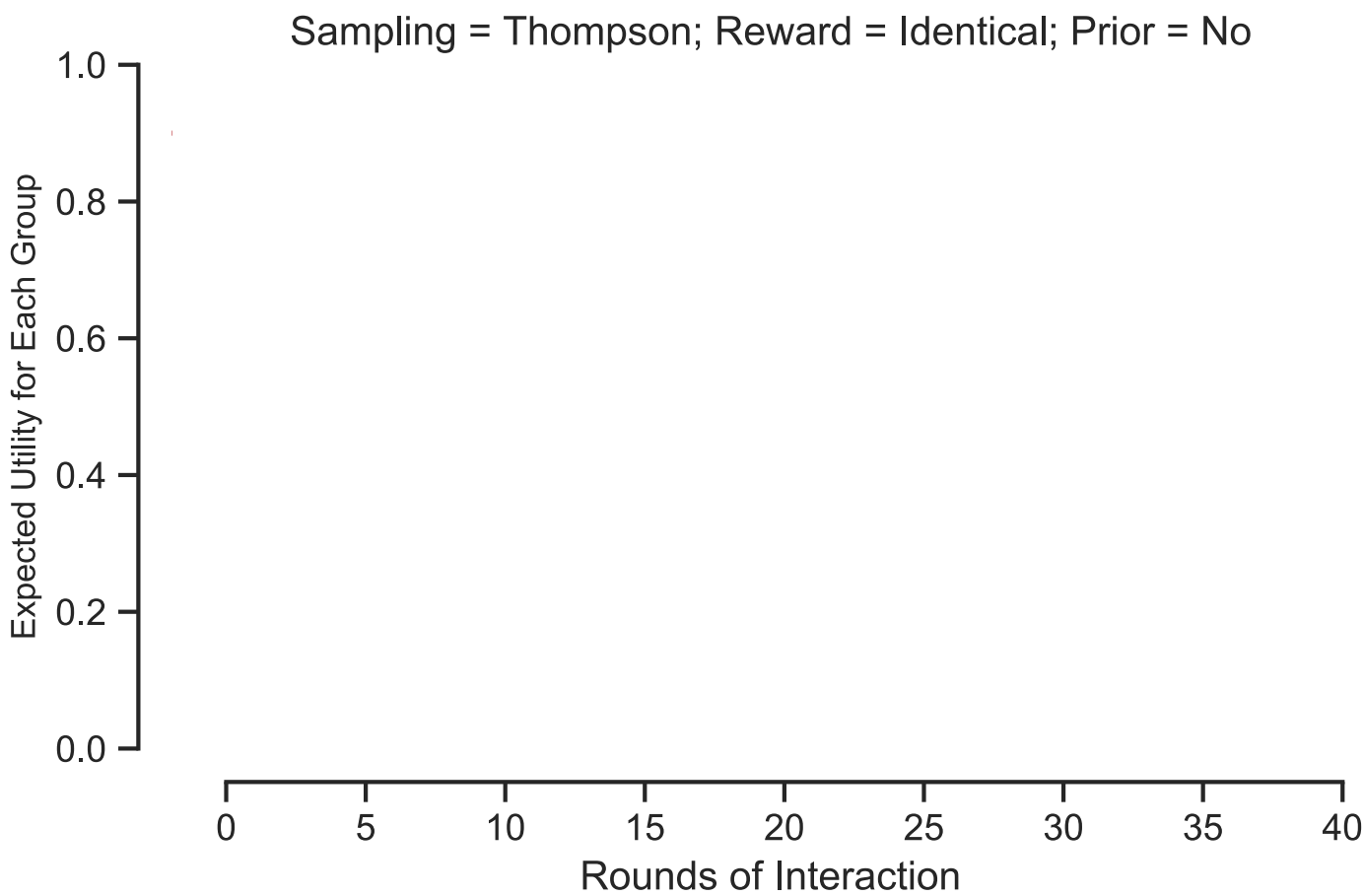
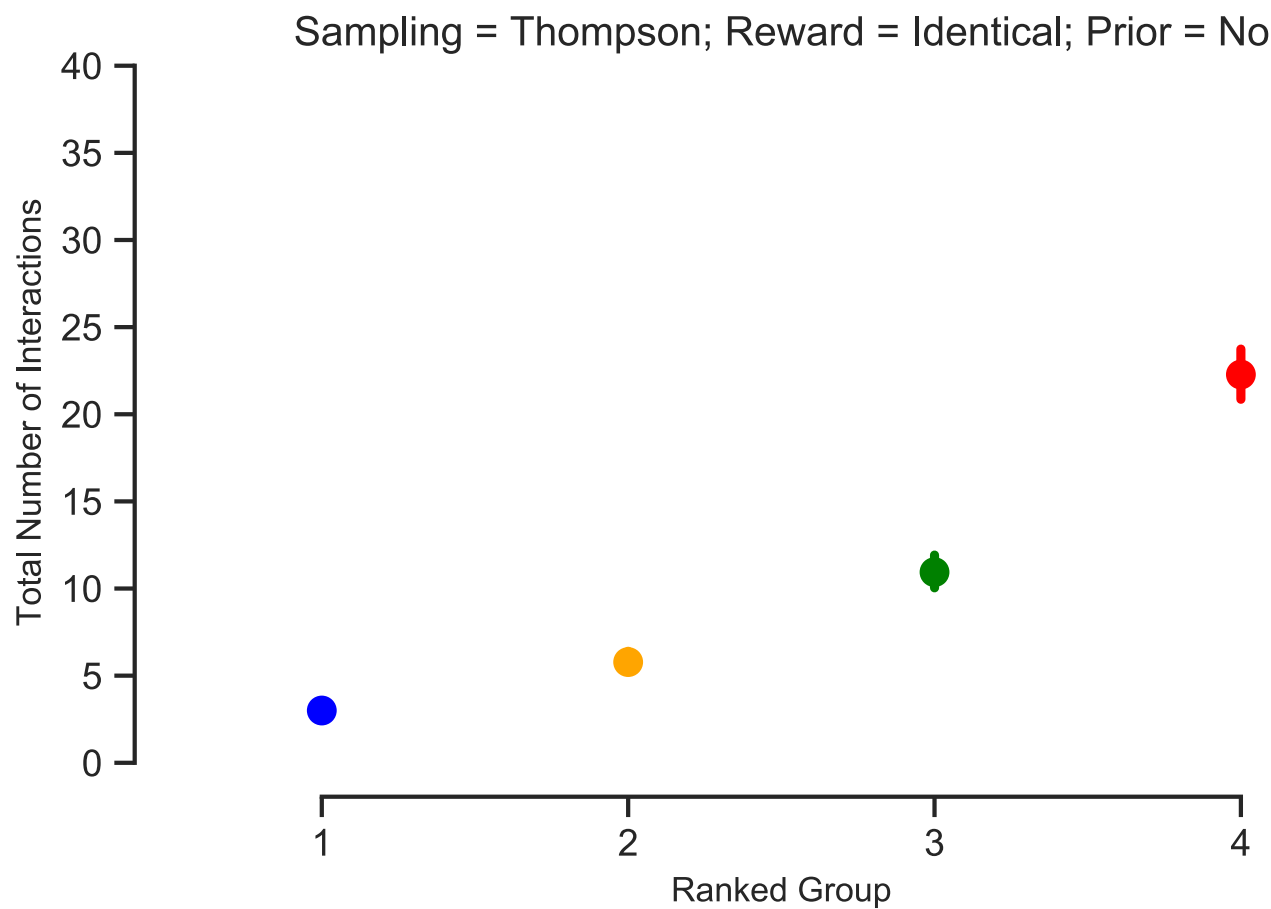


Sketching the Mechanism: Hypothesis

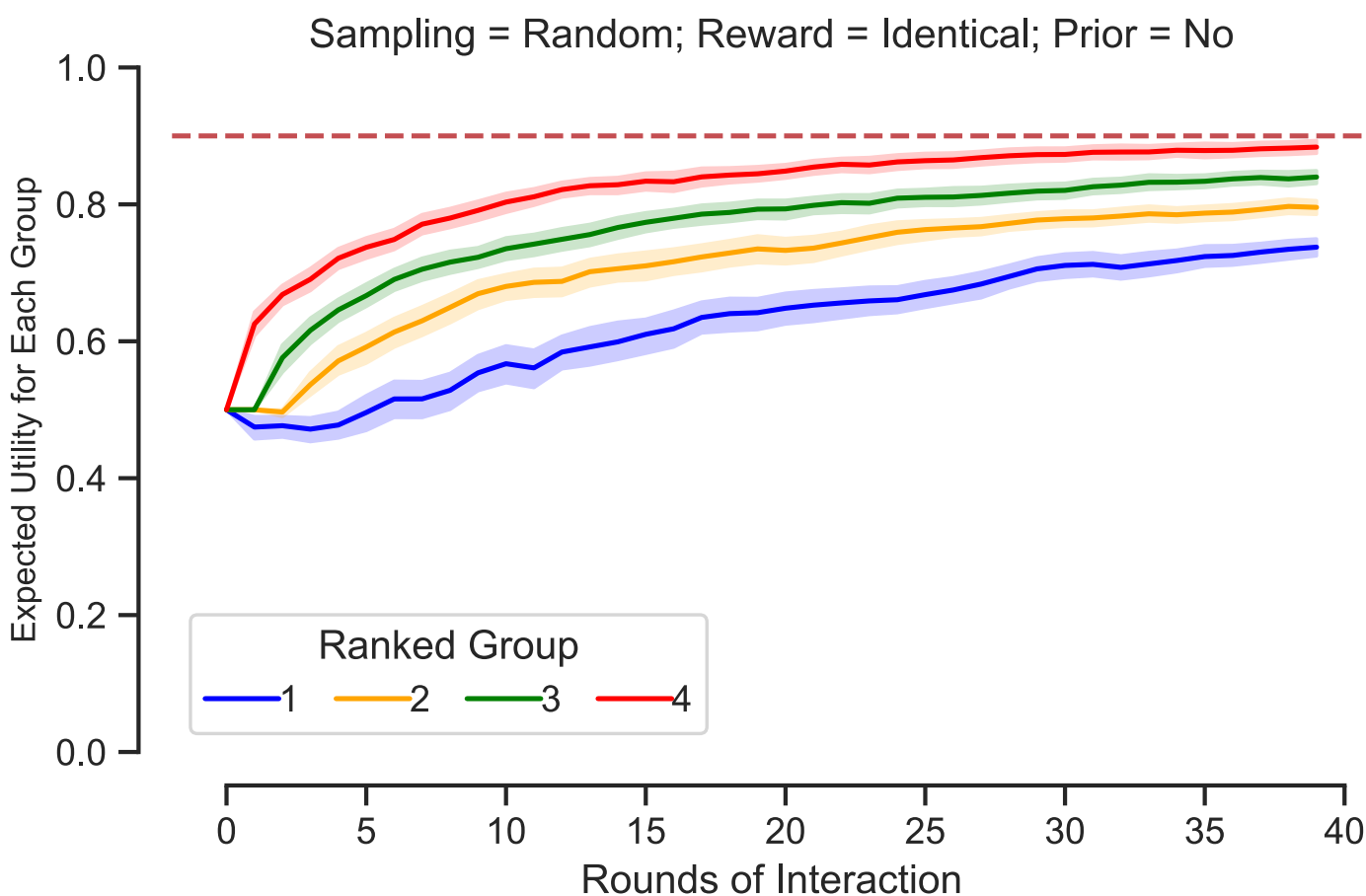
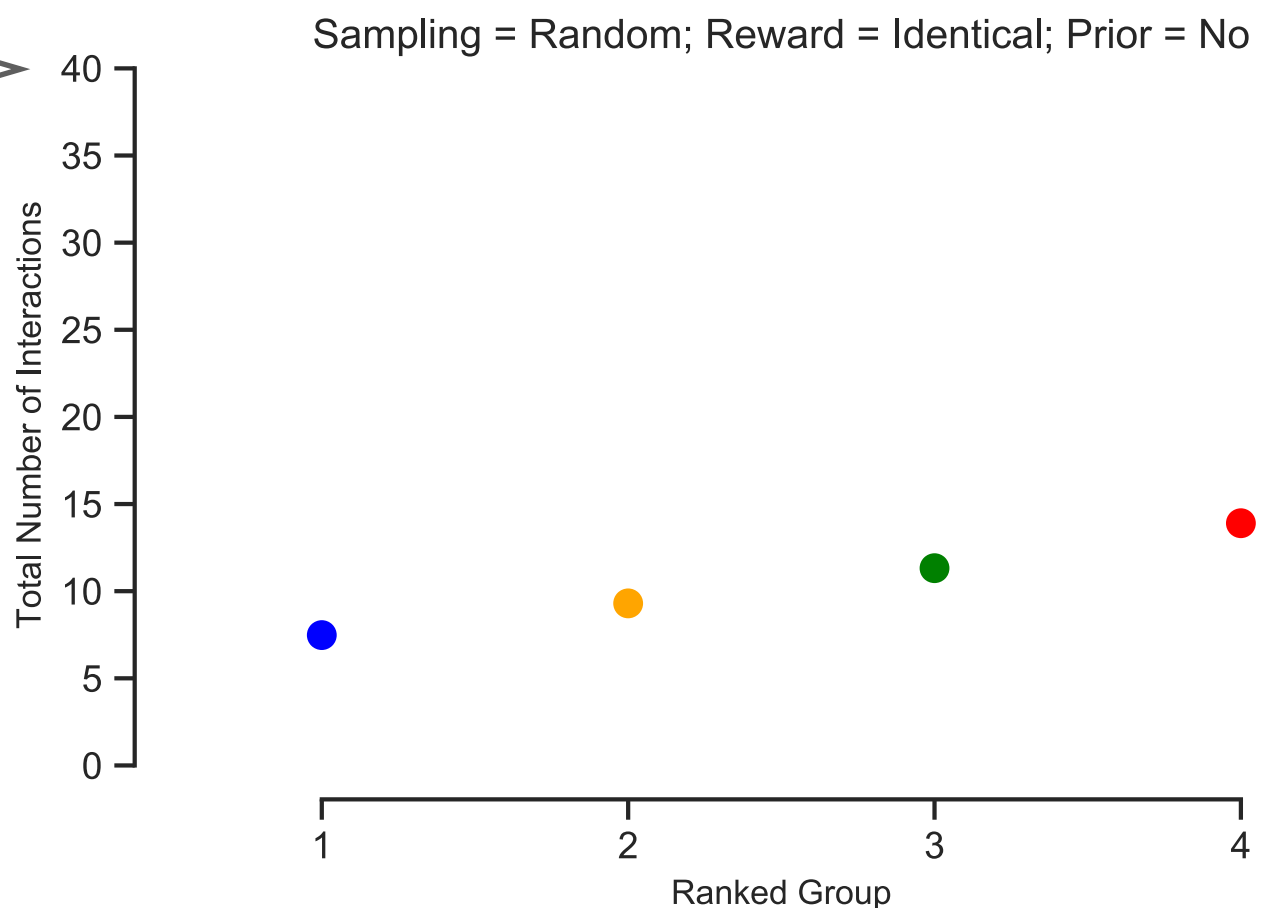
Treatment group: Thompson sampling

- Bayesian agent, prior beta (1,1), 4 arms, identical rewards ($\theta=0.9$), 40 rounds, 100 seeds.

How many times did the agent interact with each group?



What is the estimated reward for each group?

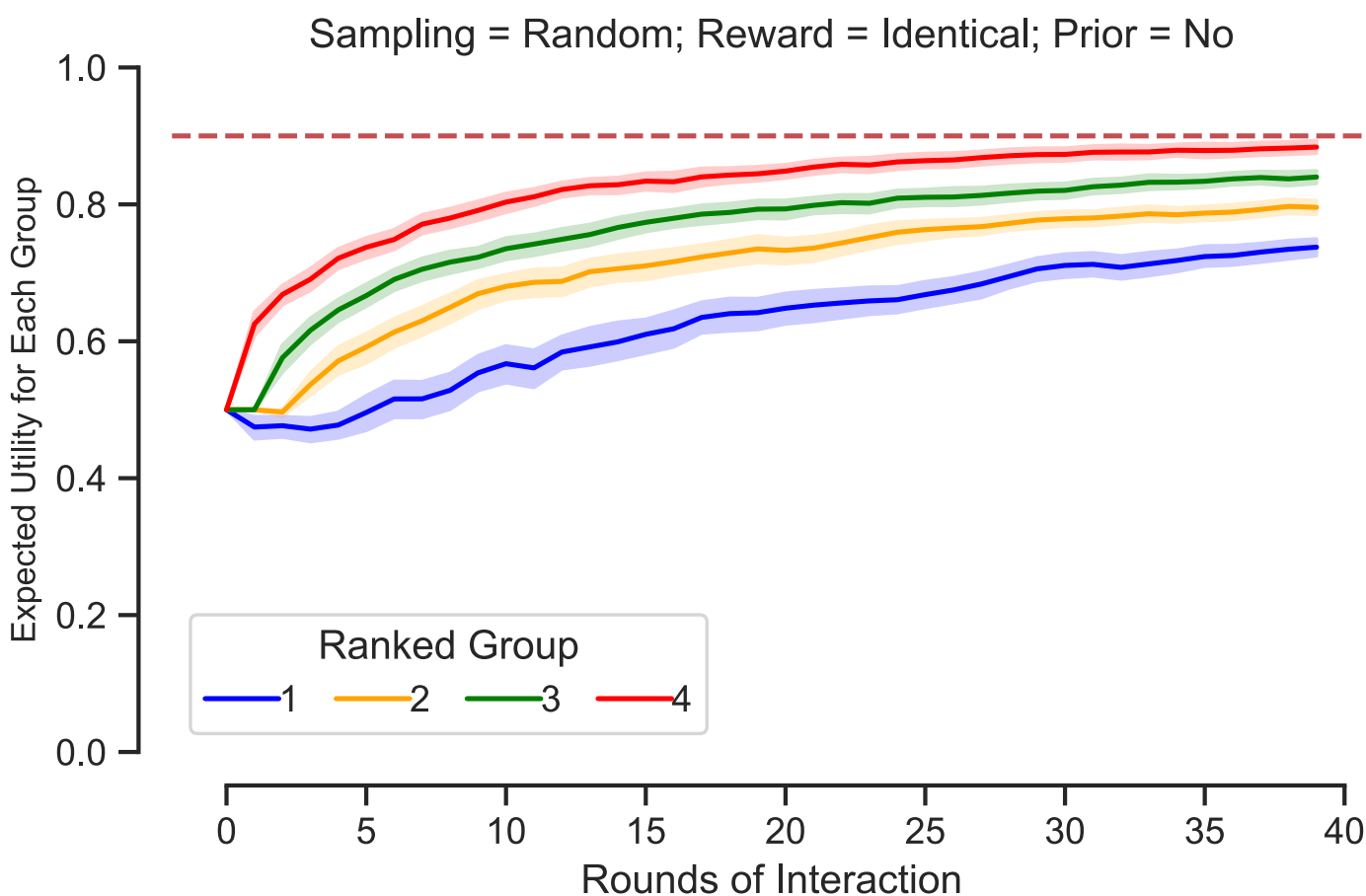
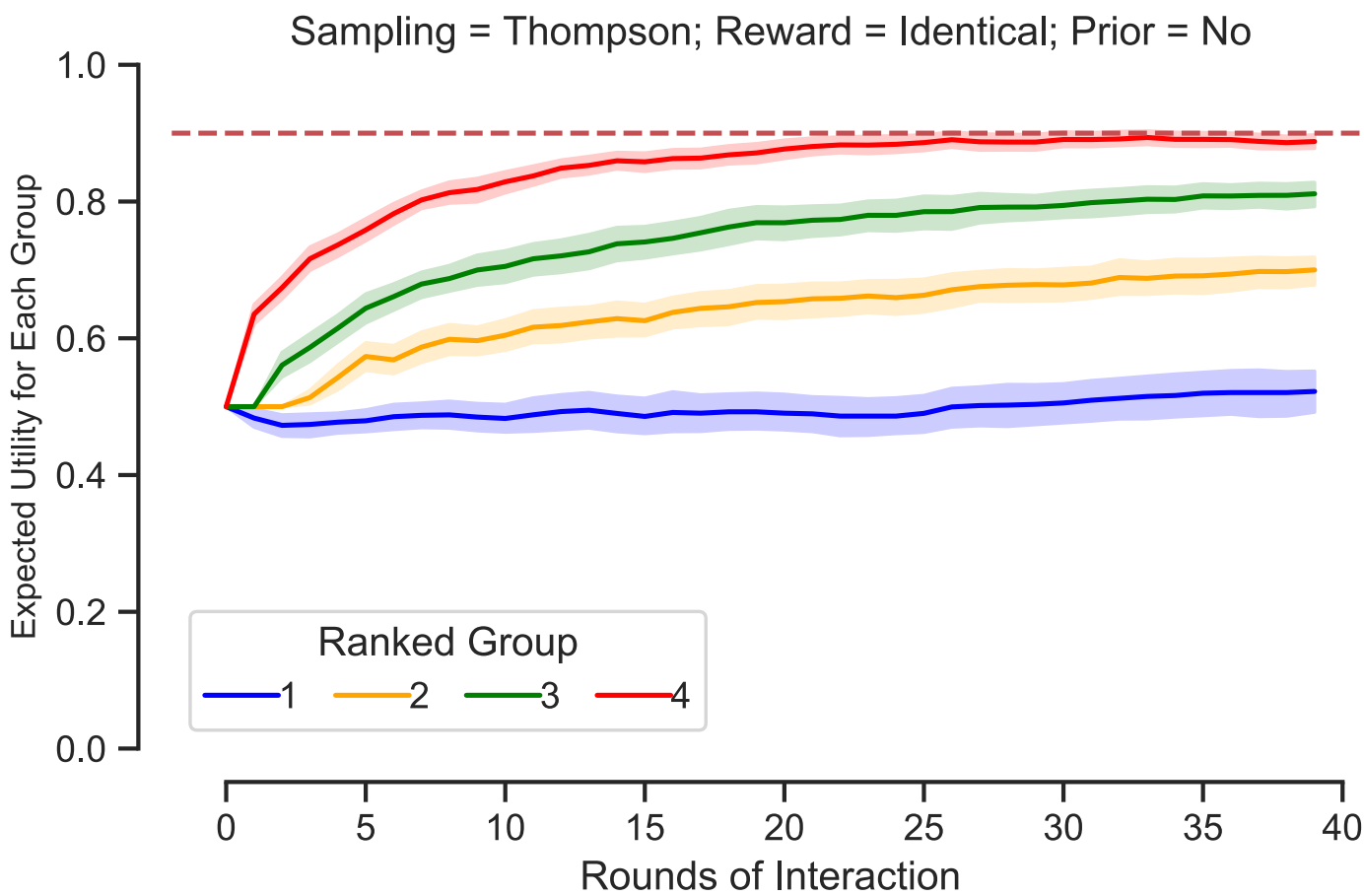
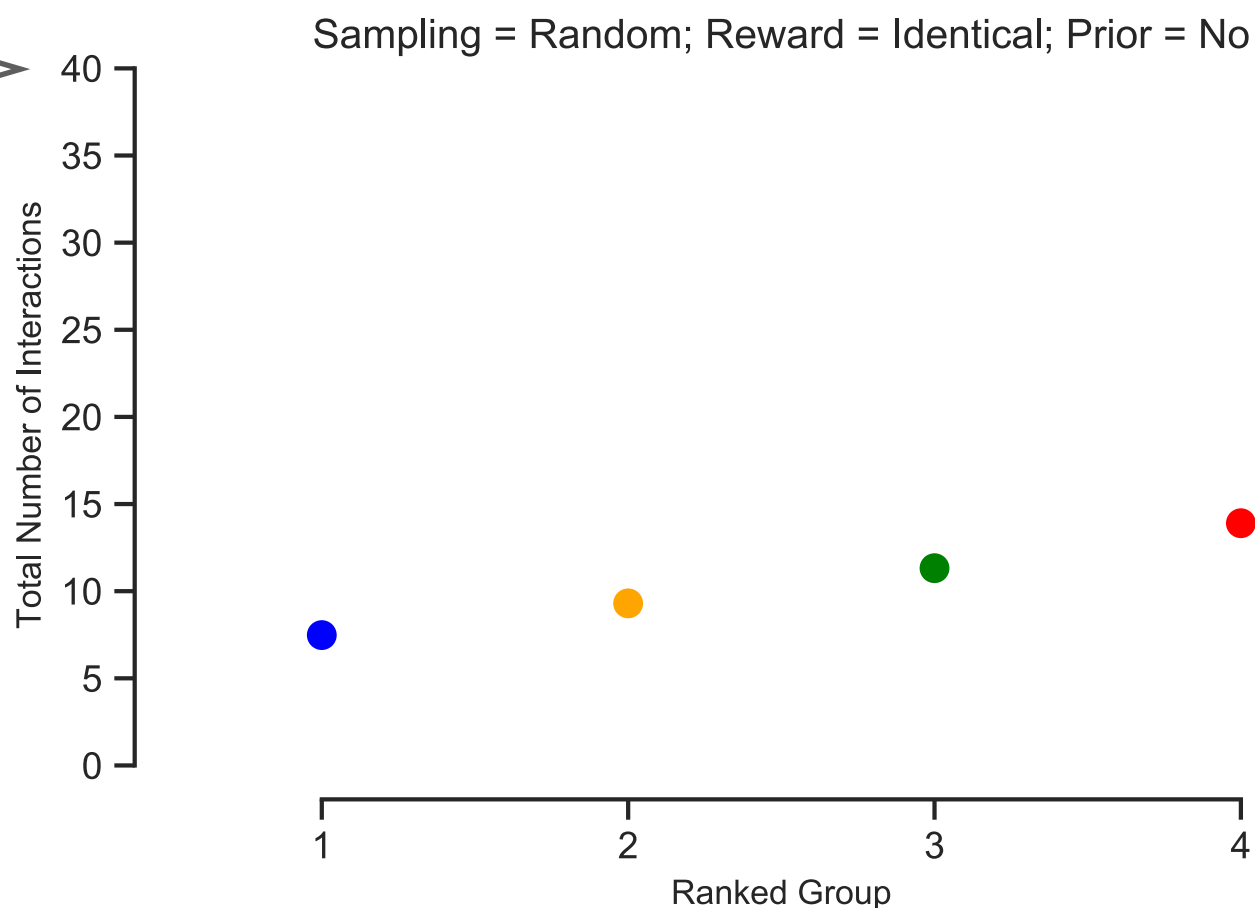
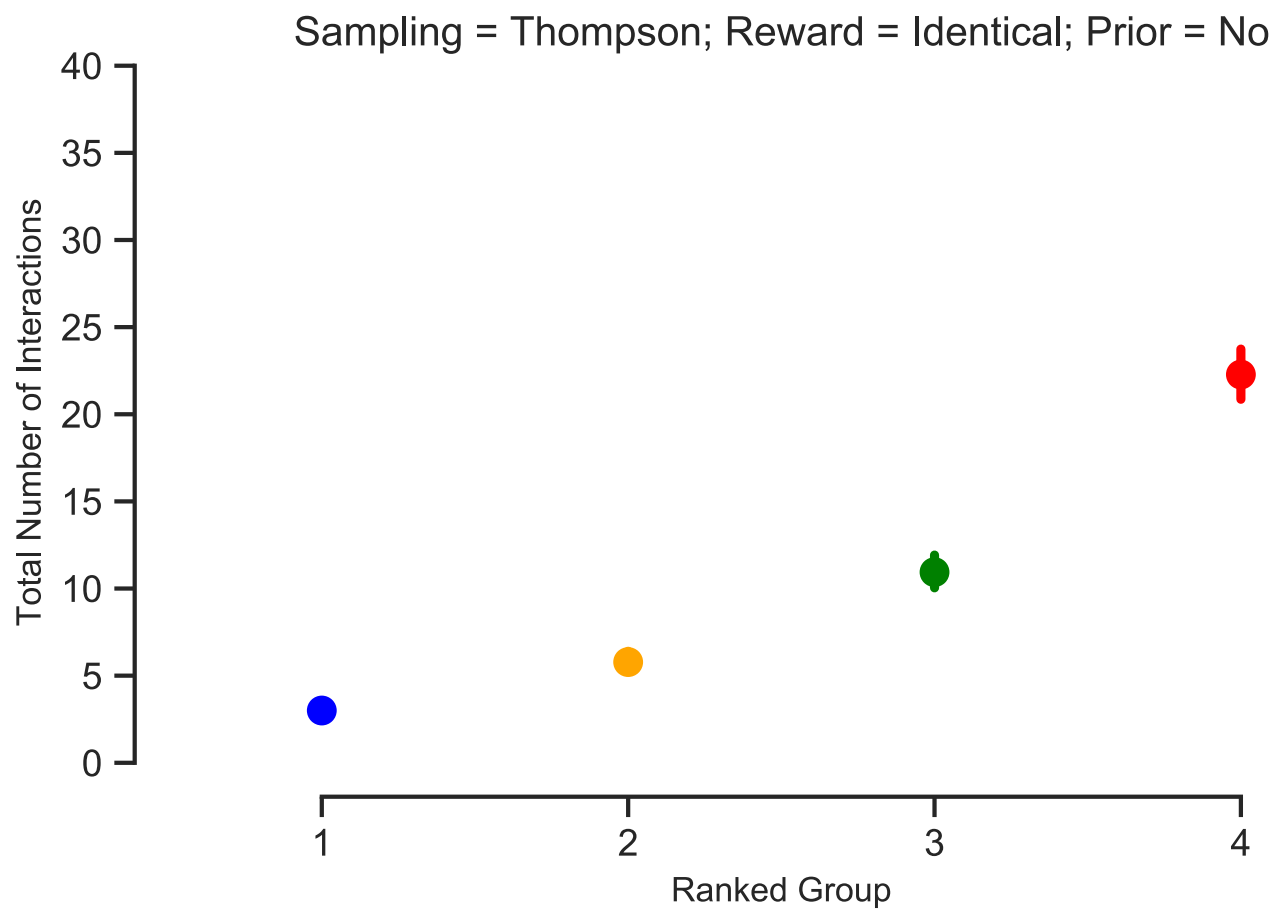


Sketching the Mechanism: Hypothesis

Treatment group: Thompson sampling

- Bayesian agent, prior beta (1,1), 4 arms, identical rewards ($\theta=0.9$), 40 rounds, 100 seeds.

How many times did the agent interact with each group?



What is the estimated reward for each group?

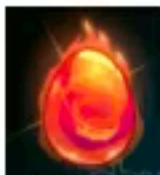
Explore Toma City:



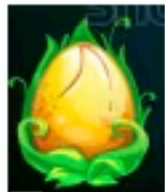
Ready to play?

Let's meet some new Tufas, Aimas, Rekus, and Wekis!

Tufa



Aima



Reku



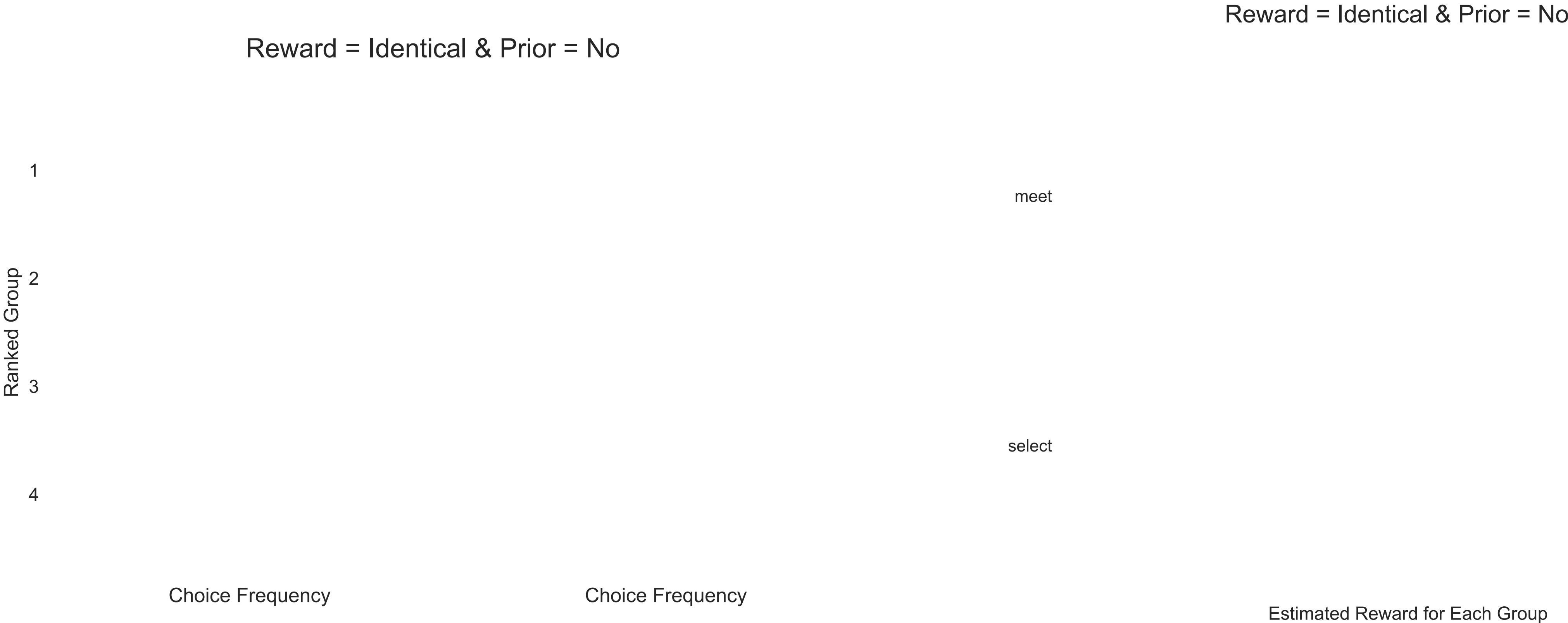
Weki



Sketching the Mechanism: Experiment

Explore Toma City:

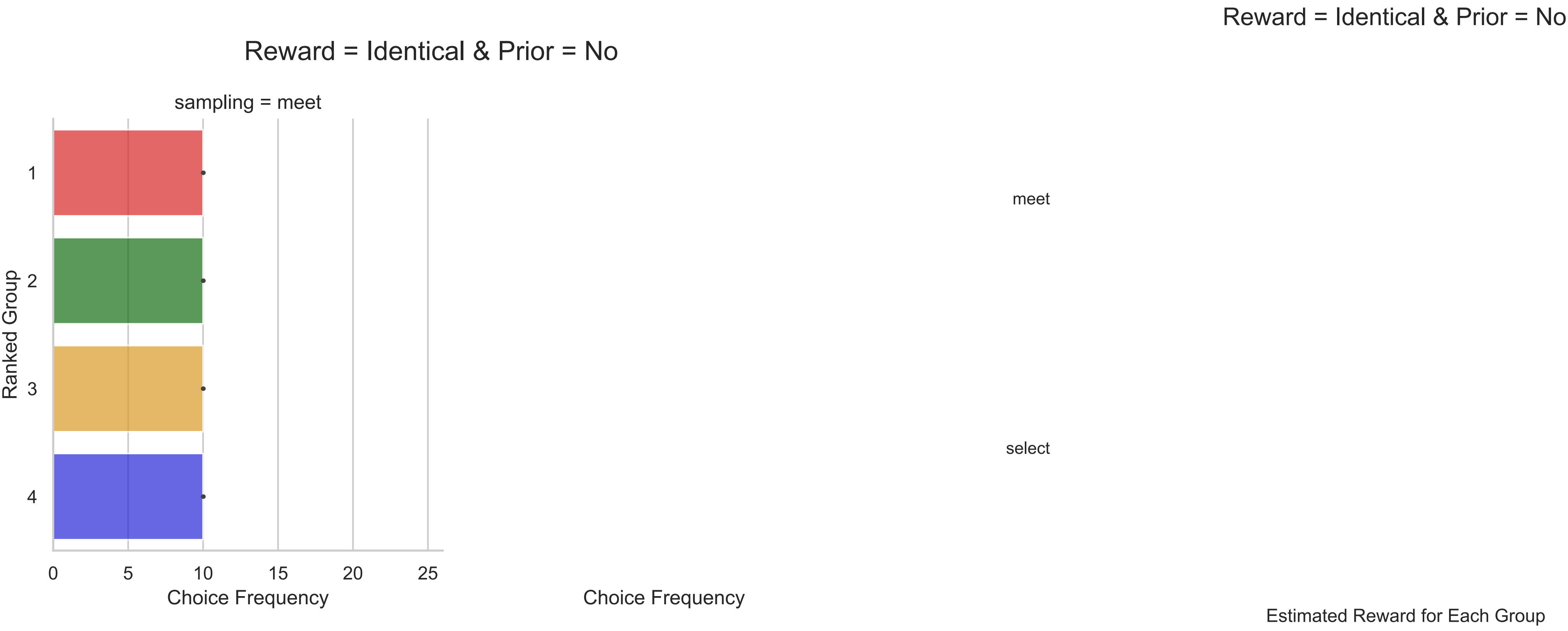
400 online participants in Study 1 (N = 2000 in Study 2)



Sketching the Mechanism: **Experiment**

Explore Toma City:

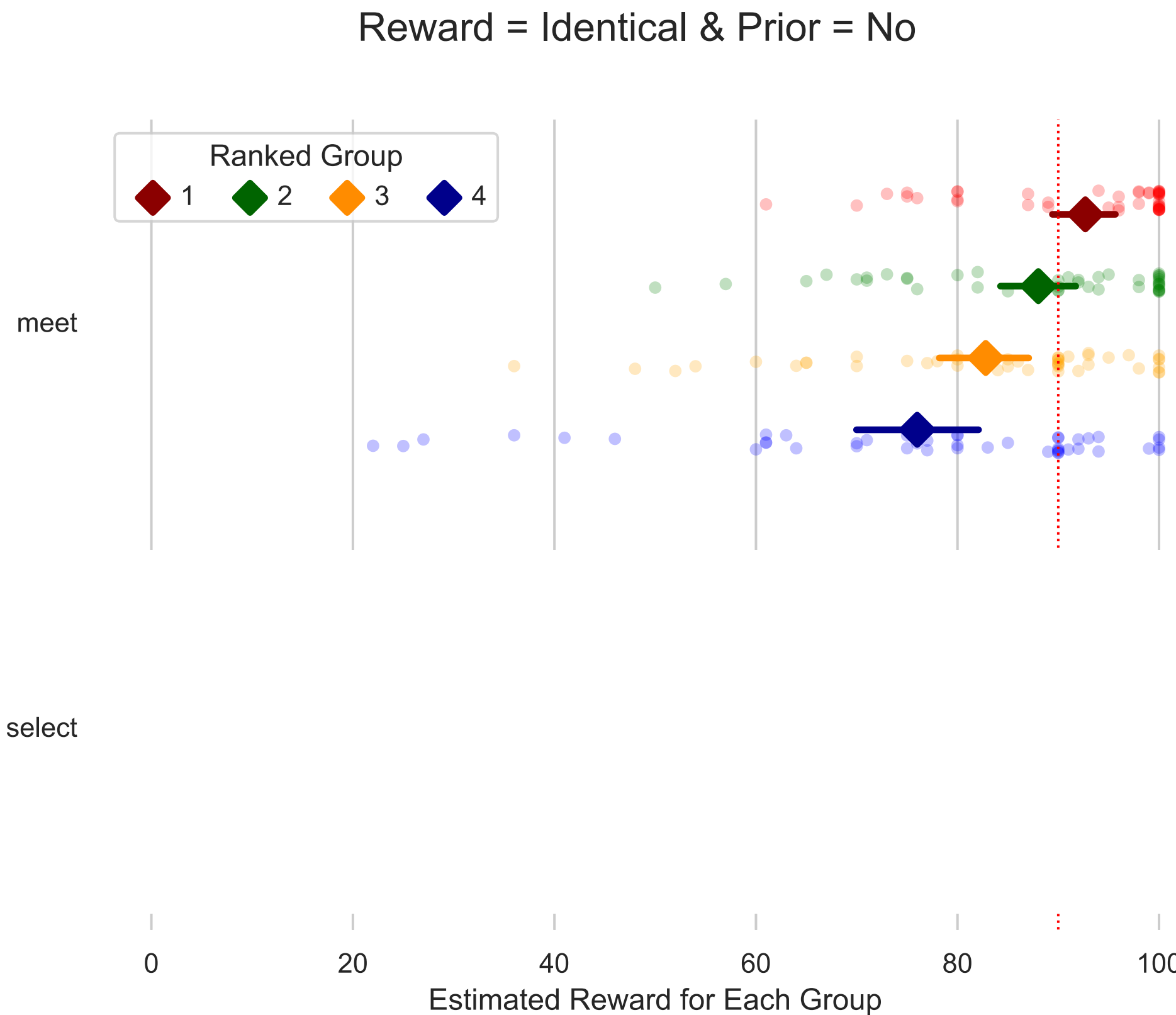
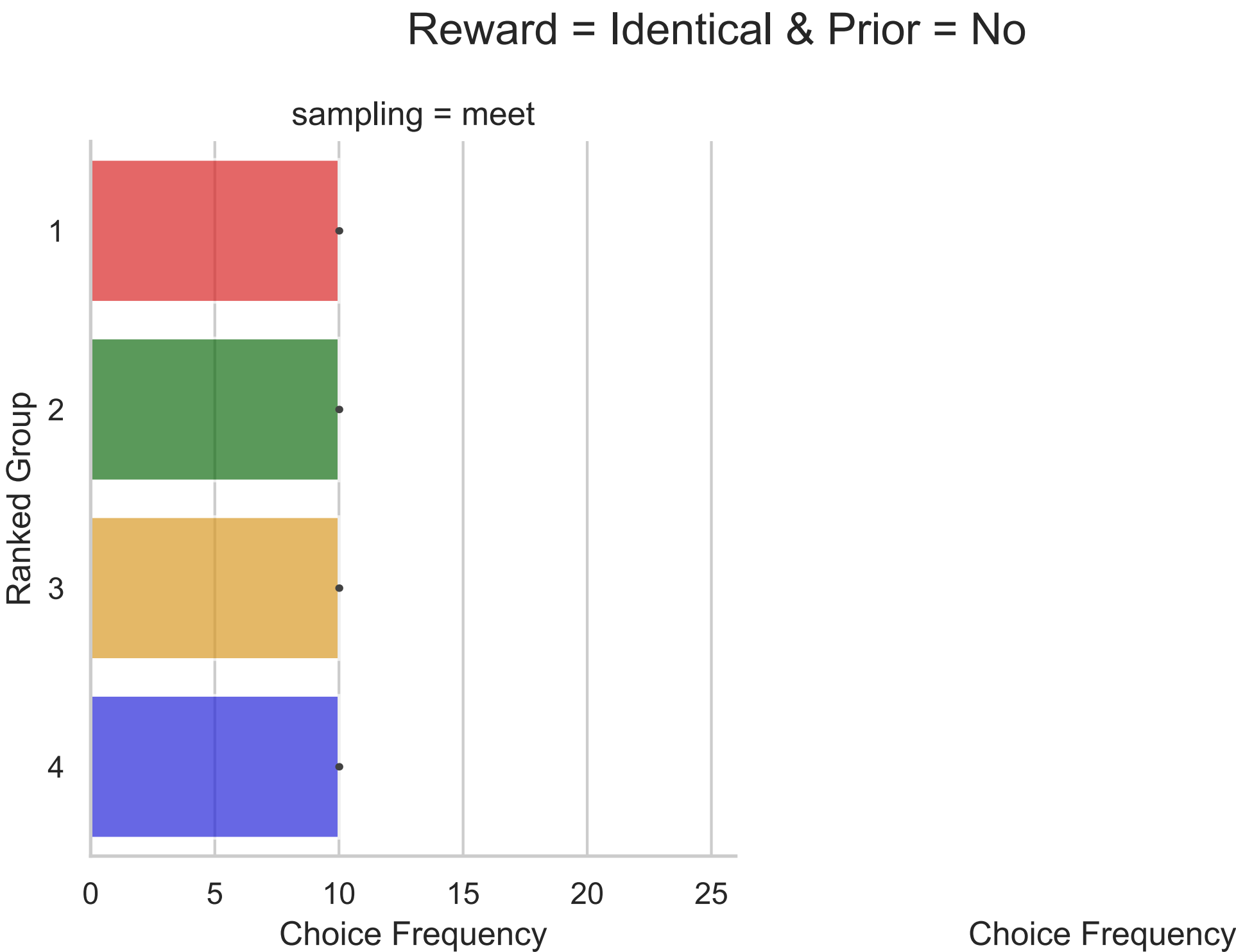
400 online participants in Study 1 (N = 2000 in Study 2)



Sketching the Mechanism: **Experiment**

Explore Toma City:

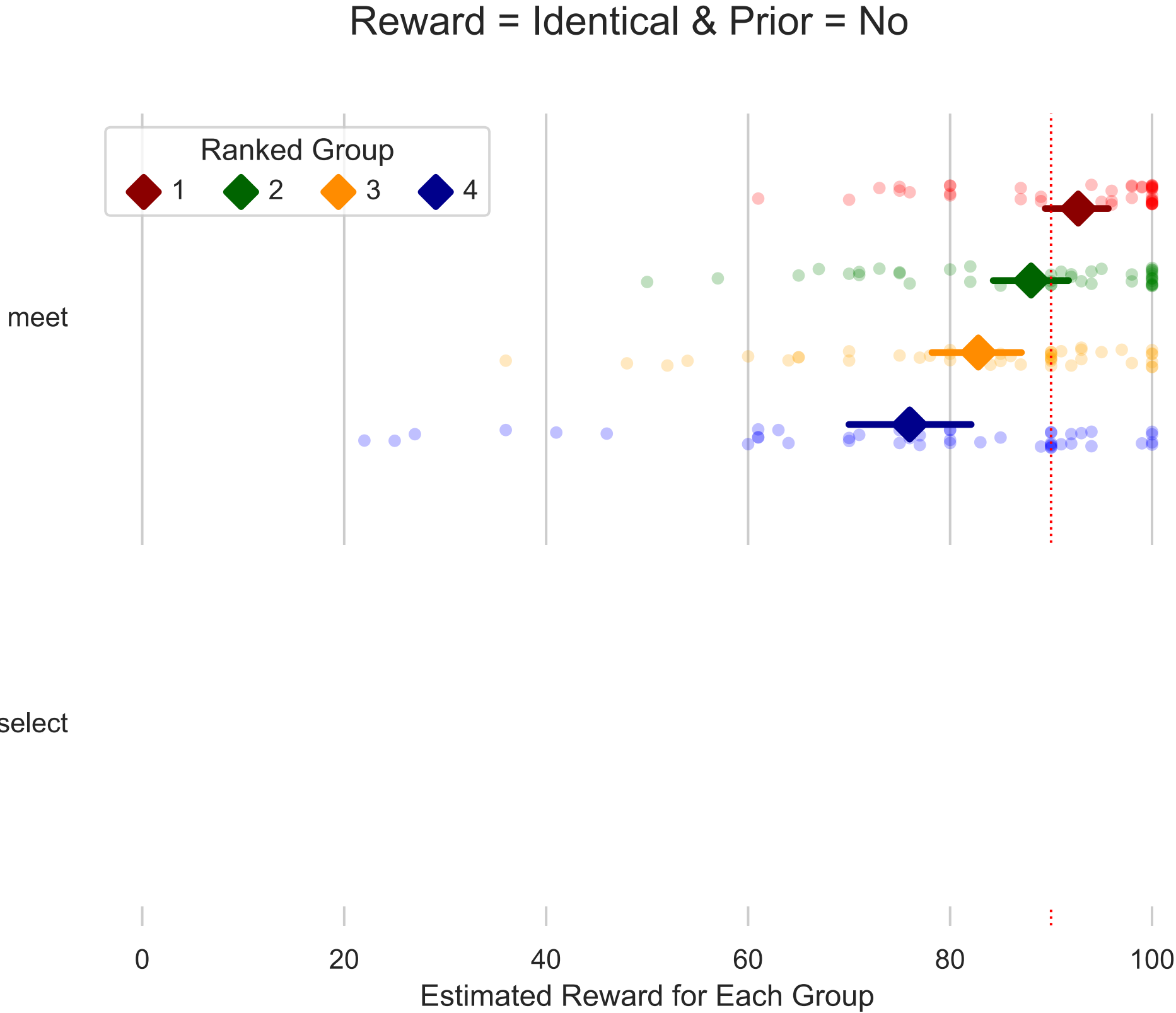
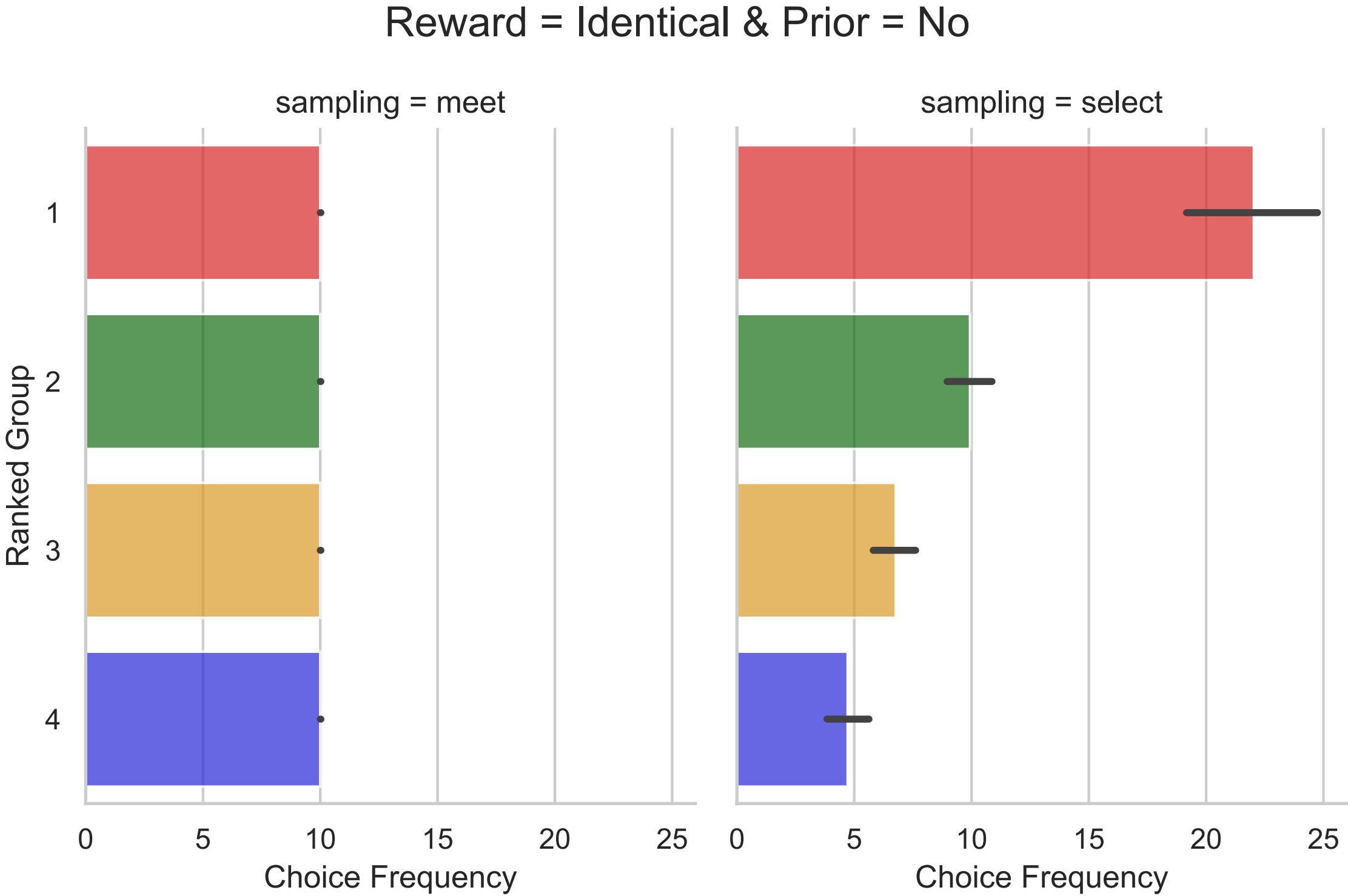
400 online participants in Study 1 (N = 2000 in Study 2)



Sketching the Mechanism: **Experiment**

Explore Toma City:

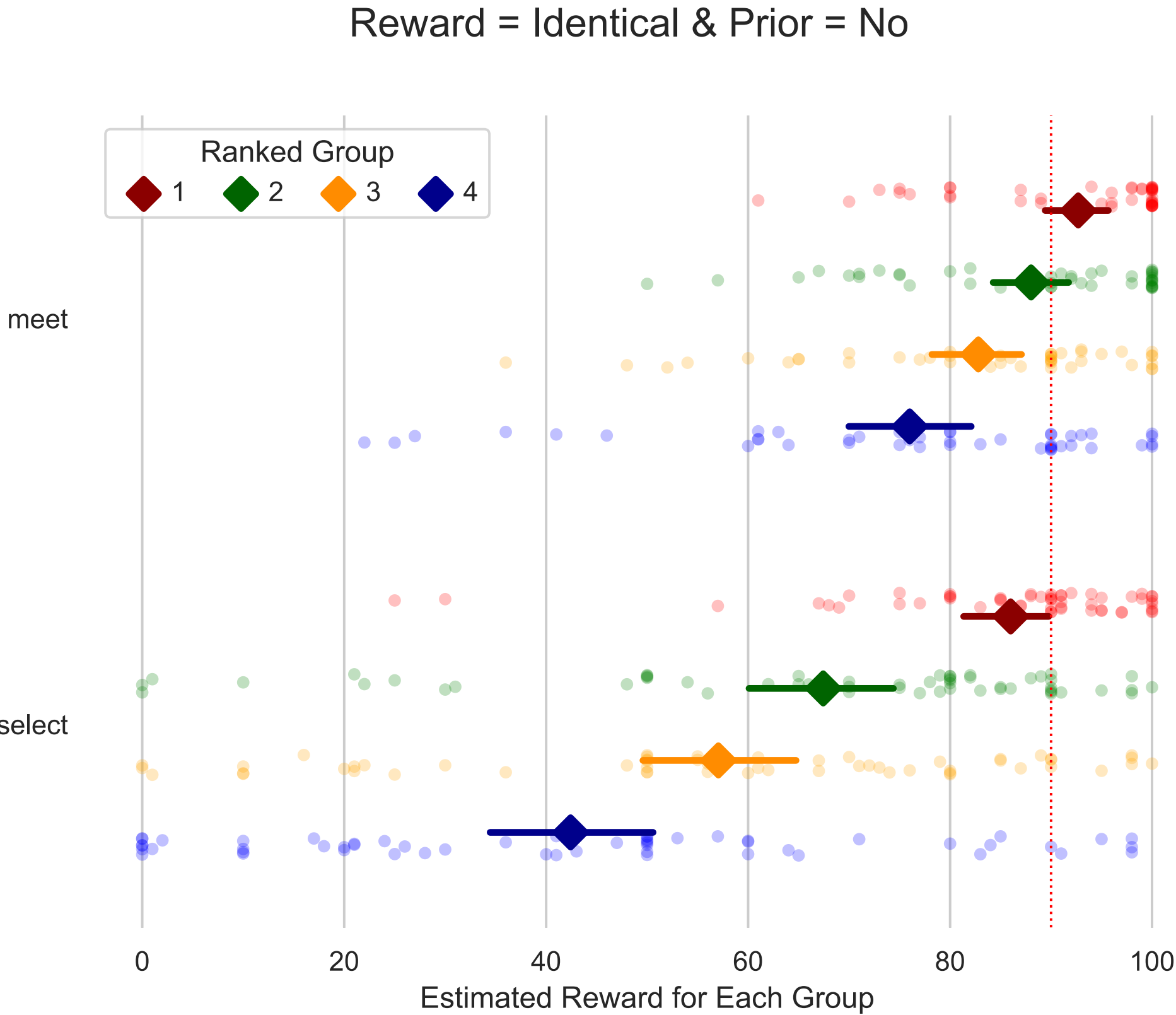
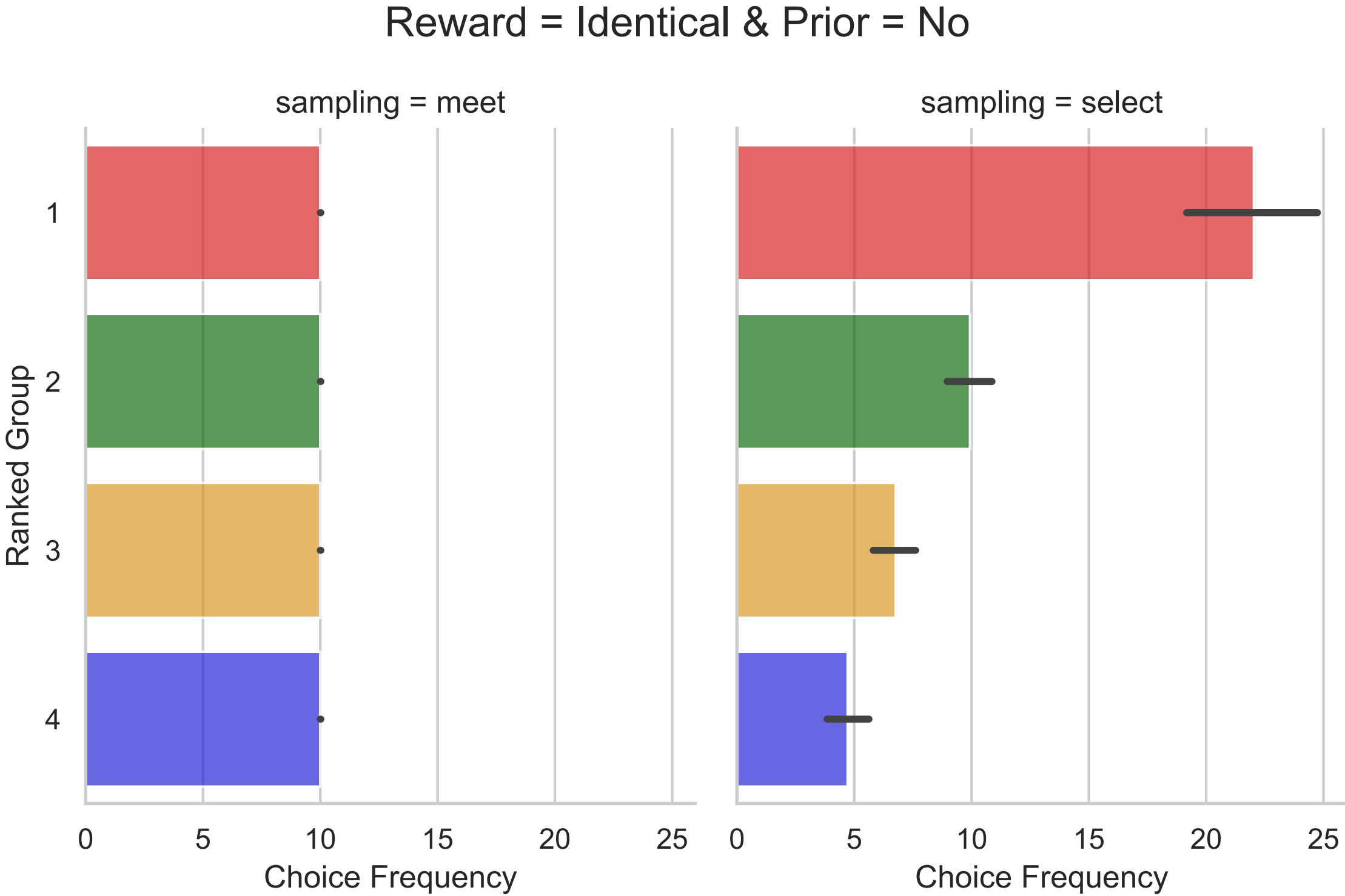
400 online participants in Study 1 (N = 2000 in Study 2)



Sketching the Mechanism: **Experiment**

Explore Toma City:

400 online participants in Study 1 (N = 2000 in Study 2)



Interim summary:

Perceived group differences emerge from agents making adaptive exploration:

- How: Make new decisions based on past [REDACTED] experiences.
- Why: Utility-maximizing [REDACTED].
- Tradeoff: Early positive experiences discourage [REDACTED] exploration.

Interim summary:

Perceived group differences emerge from agents making adaptive exploration:

- How: Make new decisions based on past (**selective**) experiences.
- Why: Utility-maximizing but (**not belief**) maximizing.
- Tradeoff: Early positive experiences discourage (**exhaustive**) exploration.

Enriching the Context

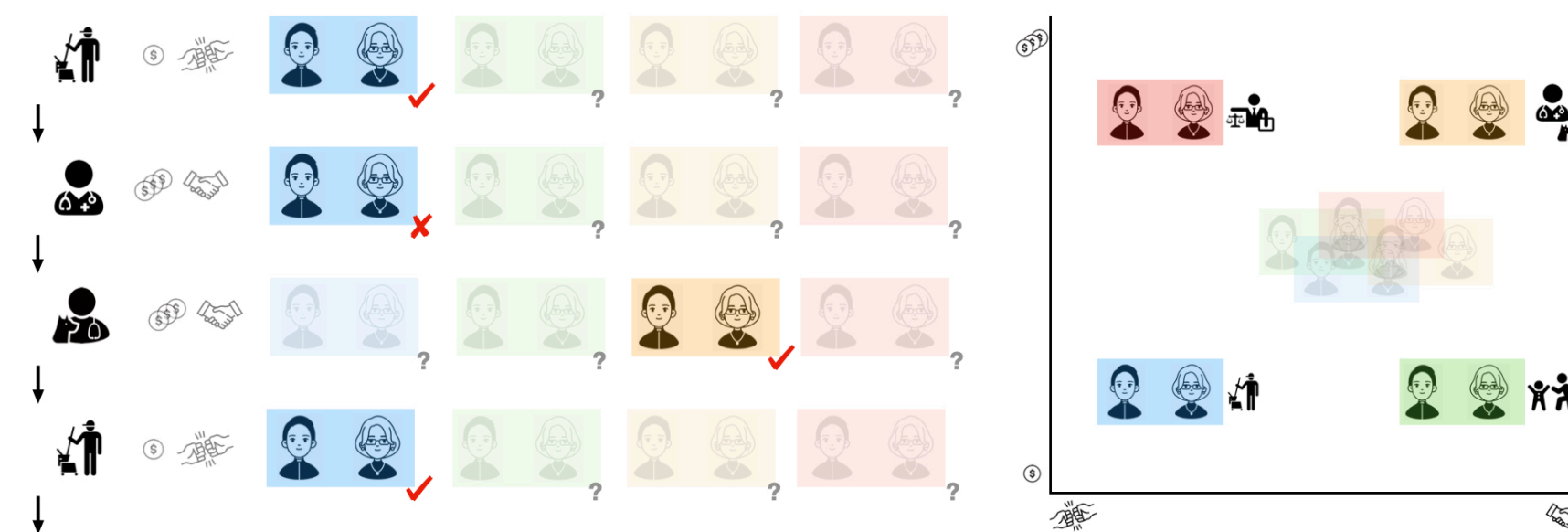


Tom Griffiths



Susan Fiske








Stereotypes entail **multi-dimensional** contents.



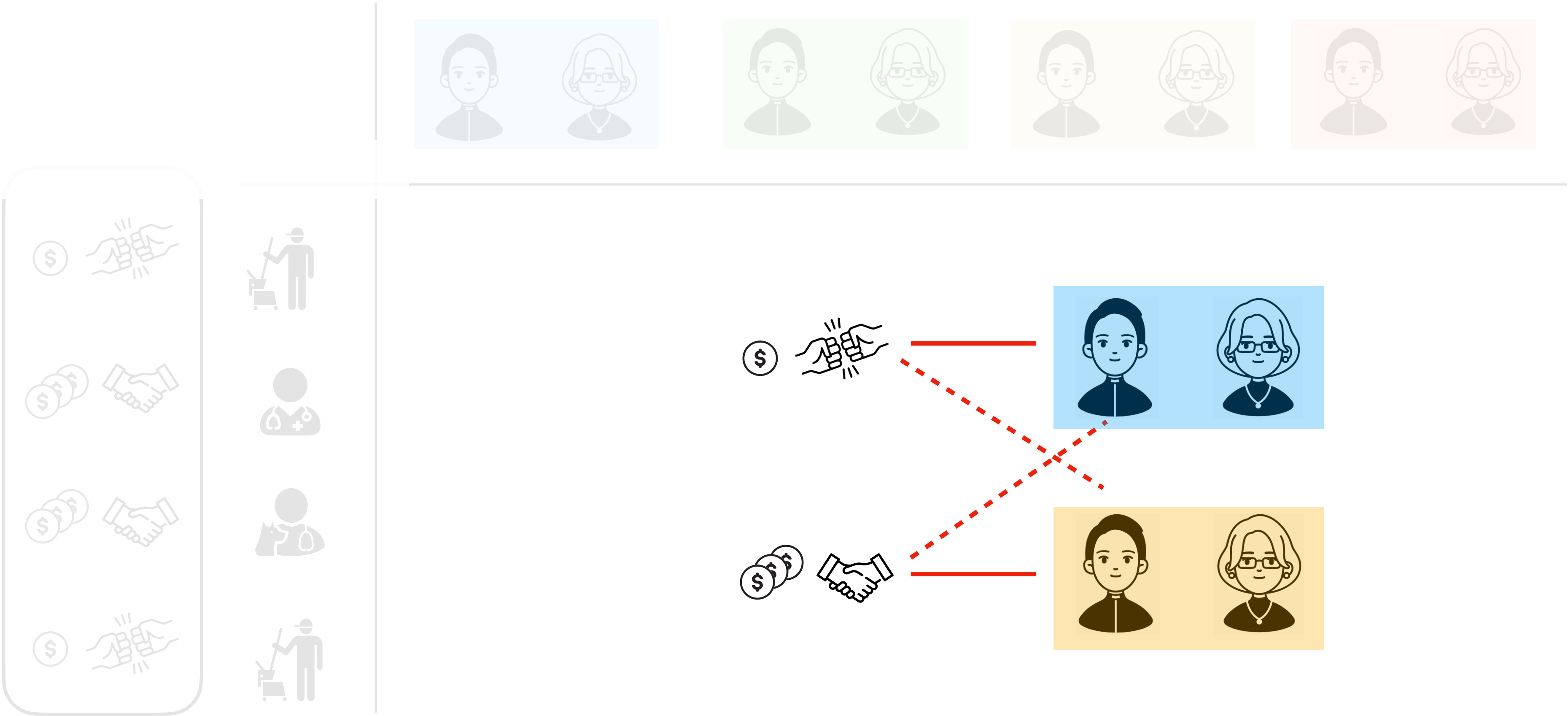
Is exploration ***per se*** the operating variable?

Hiring as Contextual Multi-Armed Bandit



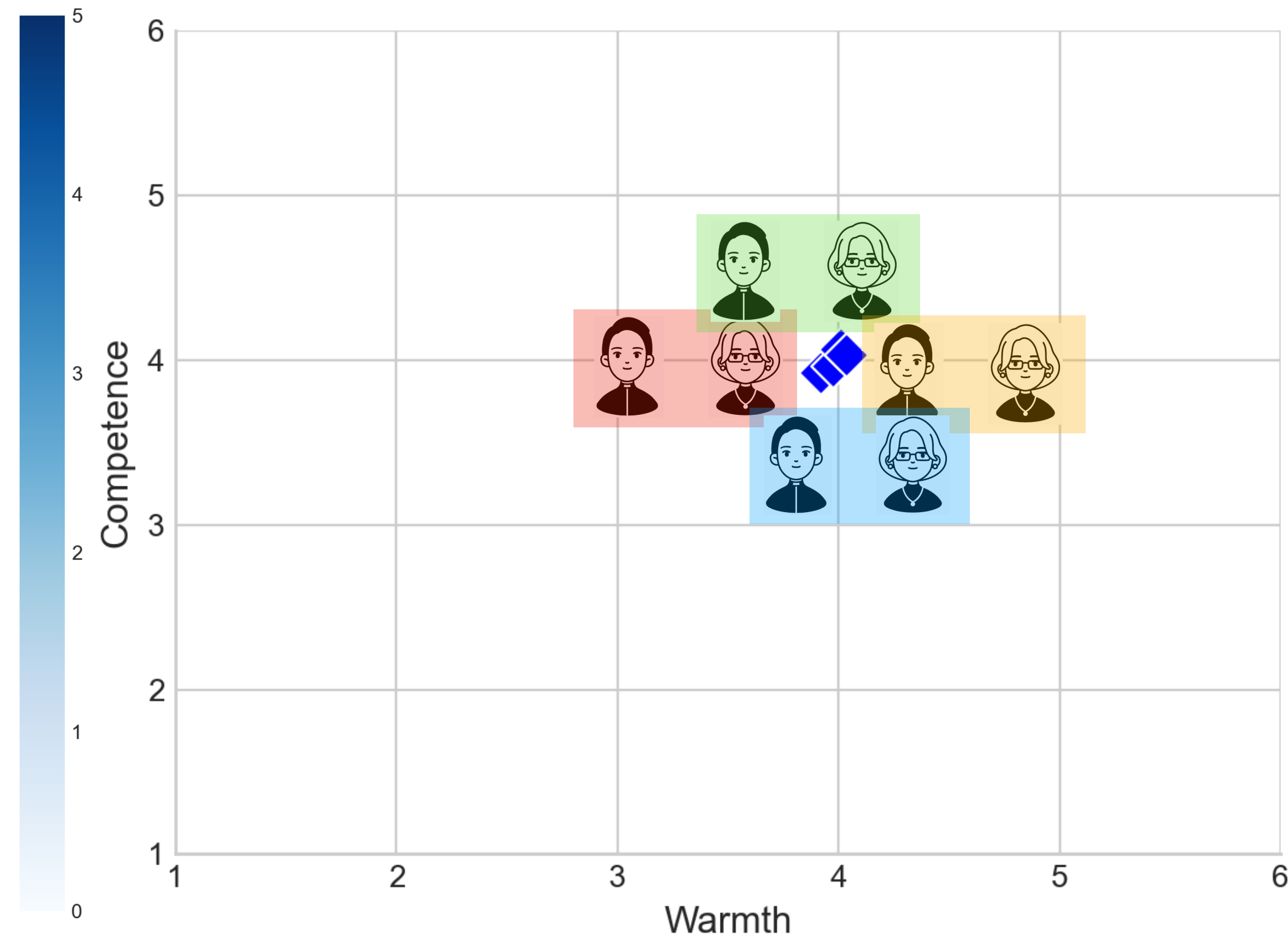
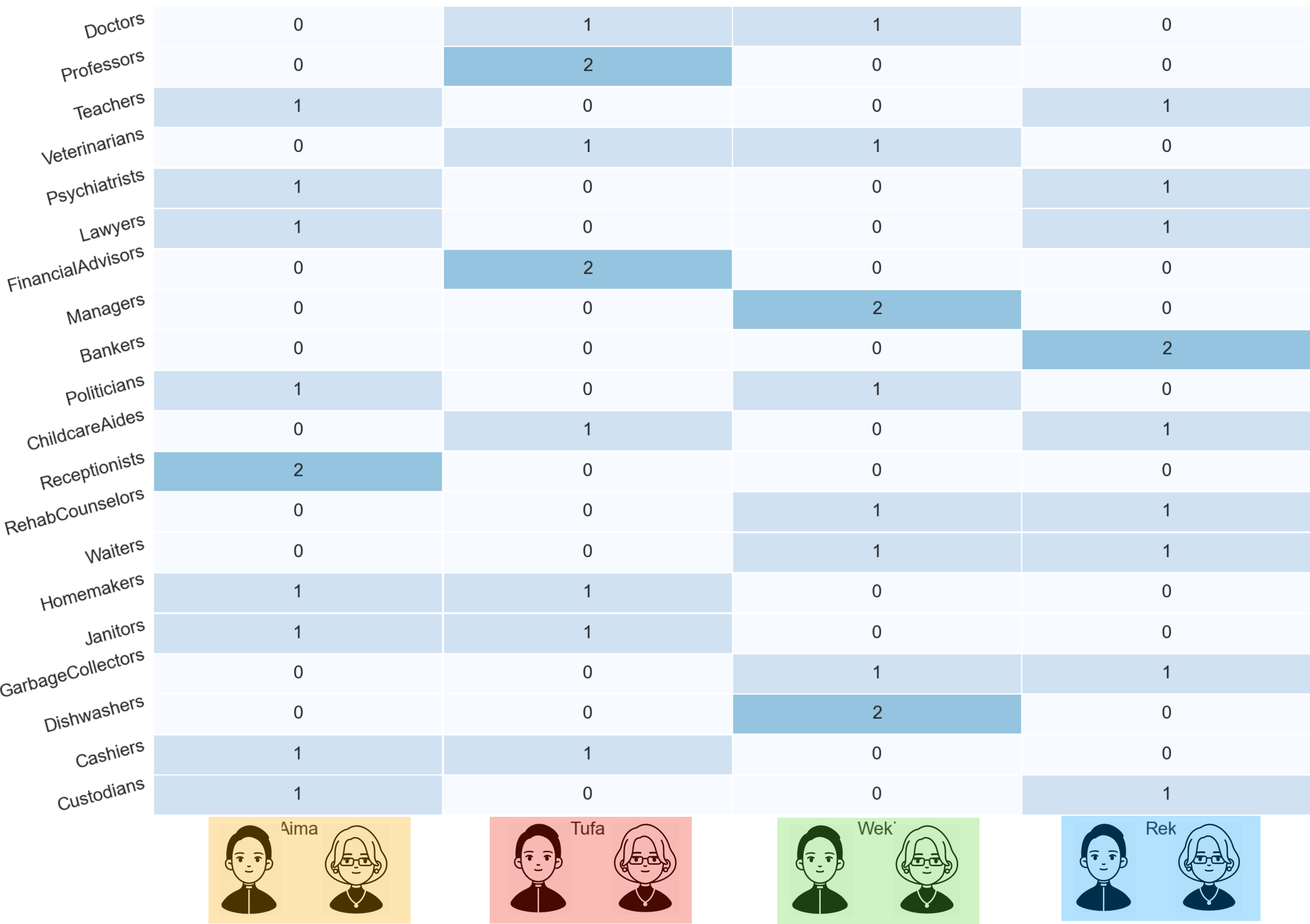
Hiring as **Contextual** Multi-Armed Bandit



Li, Chu, Langford, Schapire, 2010, WWW; Chapelle & Li, 2011, NeurIPS

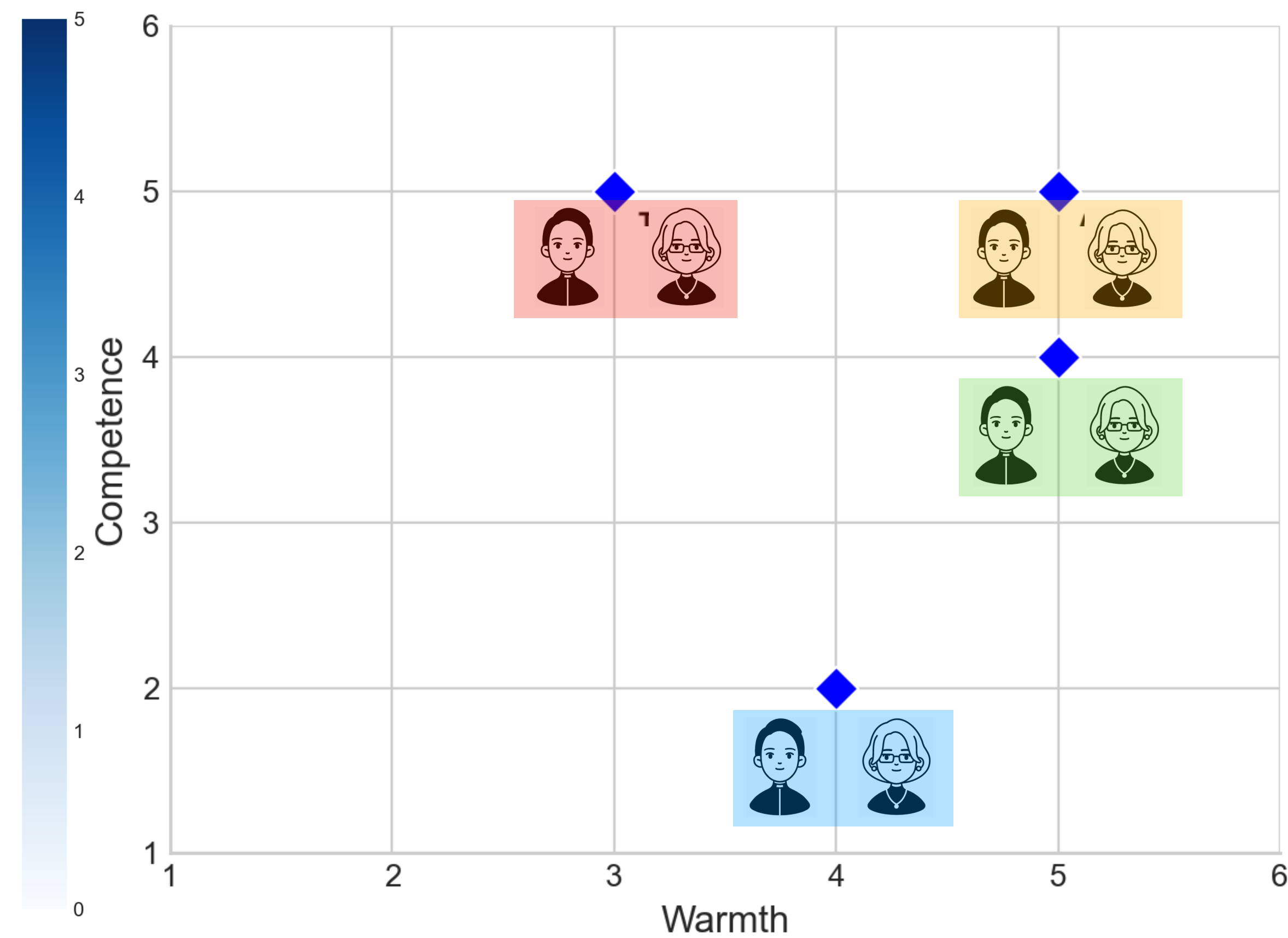
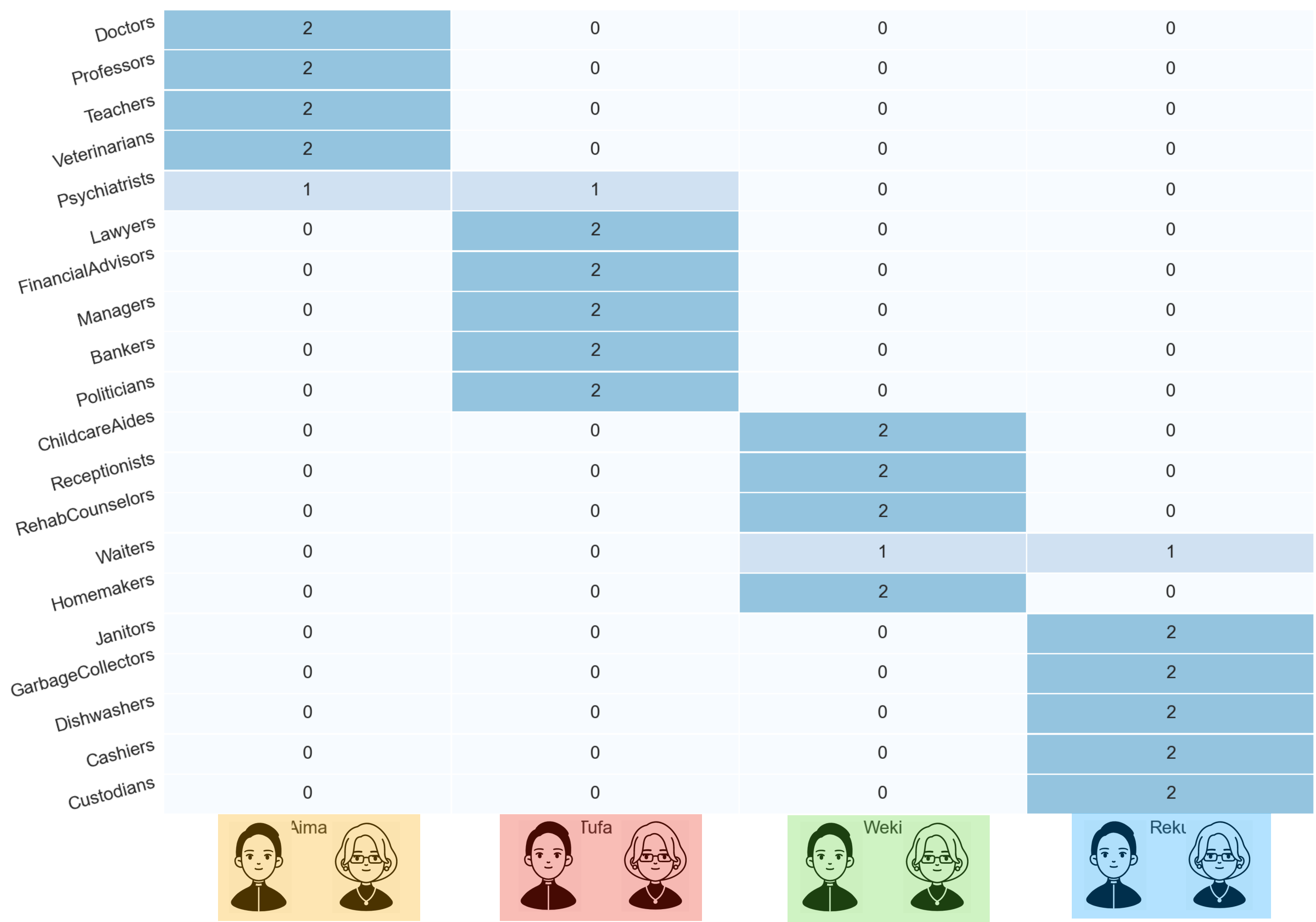
Enriching the Context: Experiments

Control group: random hiring



Enriching the Context: Experiments

Treatment group: exploratory hiring



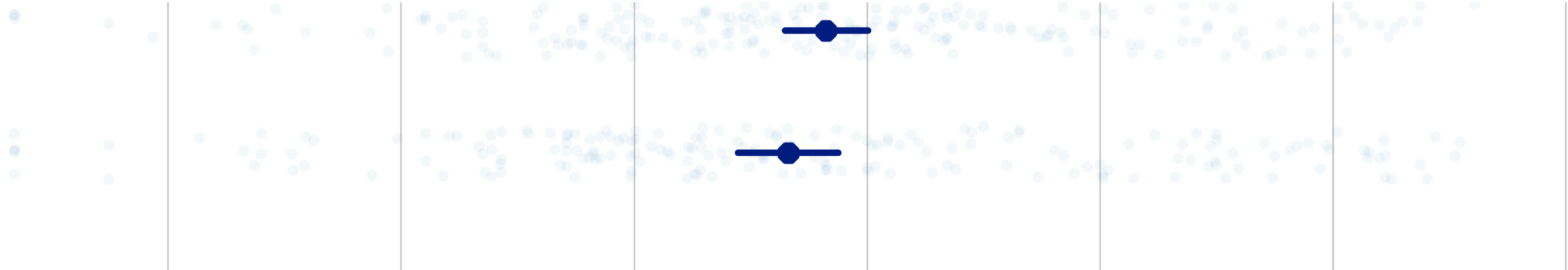
Enriching the Context: Experiments

Average treatment effect between exploratory vs. random hiring (N = 1300)

Default:

Exploratory hiring

Exploratory hiring (replication)



Ideal:

Random hiring

1.4

1.6

1.8

2.0

2.2

2.4

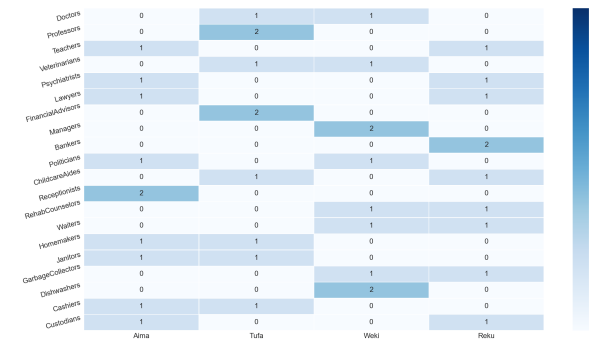
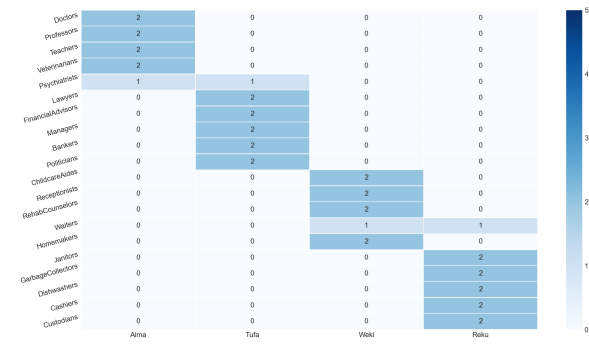
2.6

2.8

More Stratified

Decision Entropy

More Diversified



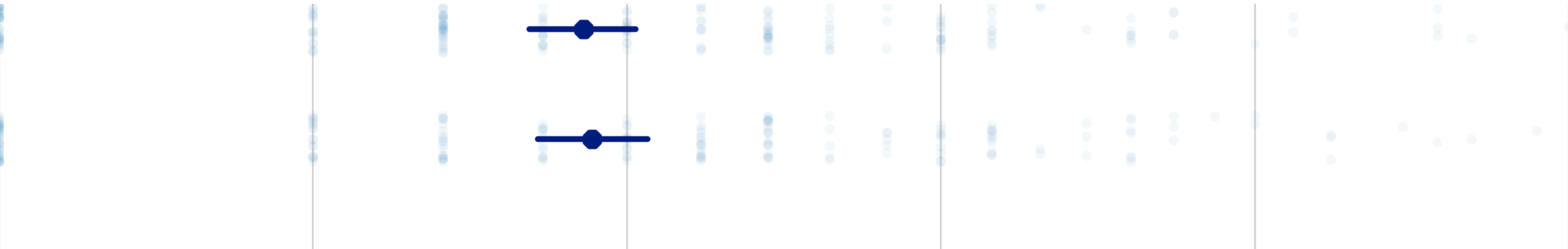
Enriching the Context: Experiments

Average treatment effect between exploratory vs. random hiring (N = 1300)

Default:

Exploratory hiring

Exploratory hiring (replication)



Ideal:

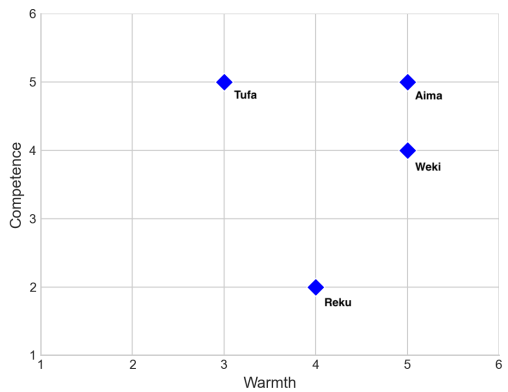
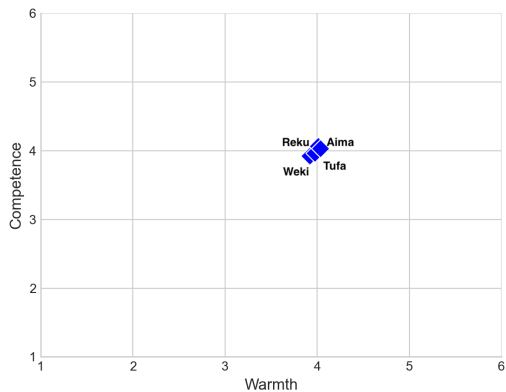
Random hiring



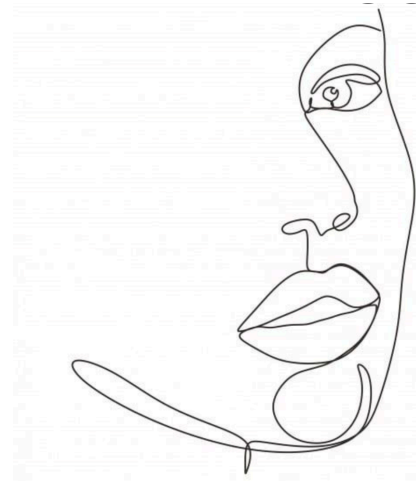
More Similar

← **Stereotype Dispersion** →

More Dissimilar



In hiring practices:



Motivational biases

- Identity
- Dominance

Cognitive biases

- Limited memory
- Selective attention

e.g., de-bias training

Enriching the Context: **Implications**

In hiring practices:



Sample biases

- Unequal group size

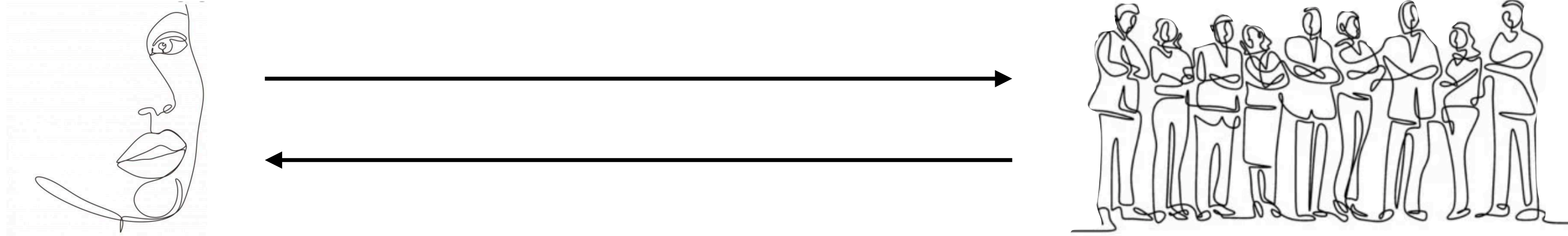
Group differences

- Gender

e.g., diversity recruitment

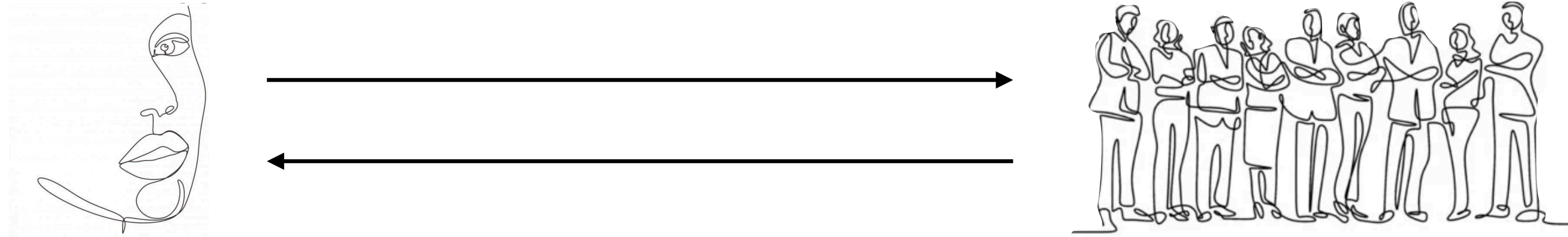
Enriching the Context: Implications

In hiring practices:



Enriching the Context: Implications

In hiring practices:



An alternative intervention:

Design a system that encourages continuous exploration.

Enriching the Context: Intervention

Average intervention effect between exploratory vs. random hiring (N = 1300)

Default:

Exploratory hiring

Exploratory hiring (replication)

Interventions:

Decrease expected reward

Add exploration bonus

Holdout groups at random

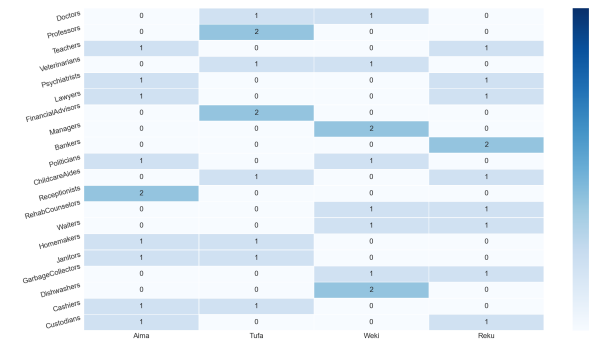
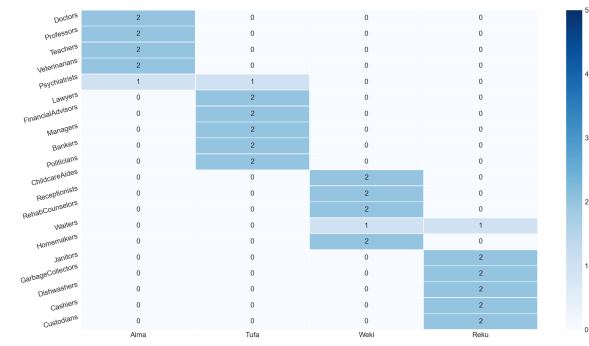
Ideal:

Random hiring

More Stratified

Decision Entropy

More Diversified



Enriching the Context: Intervention

Average intervention effect between exploratory vs. random hiring (N = 1300)

Default:

Exploratory hiring

Exploratory hiring (replication)

Interventions:

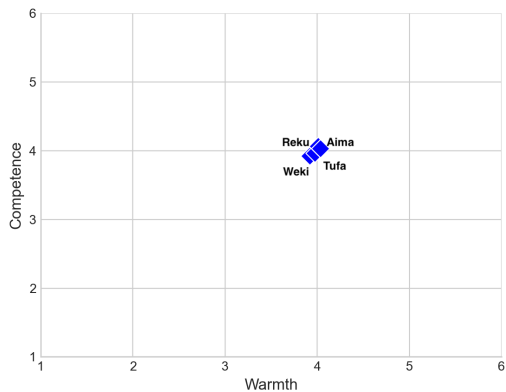
Decrease expected reward

Add exploration bonus

Holdout groups at random

Ideal:

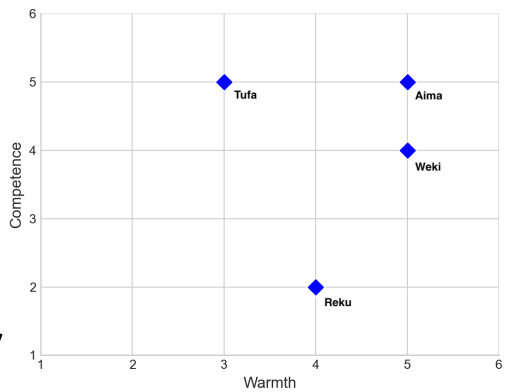
Random hiring



More Similar

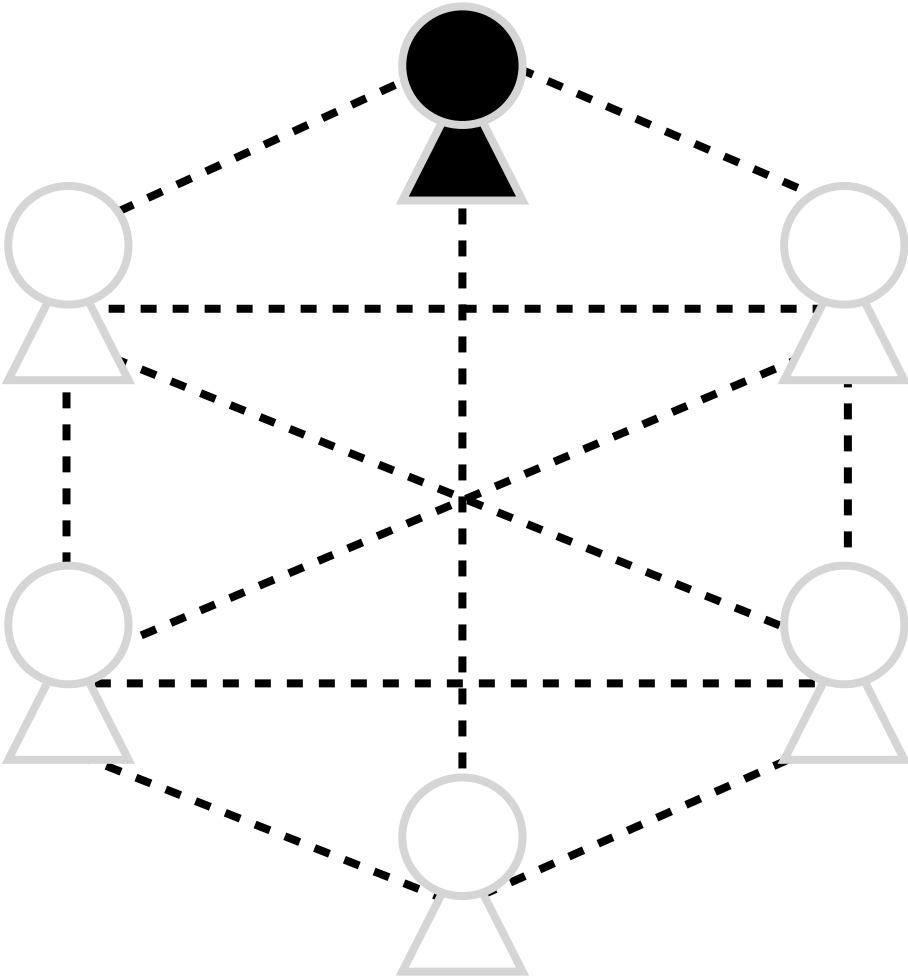
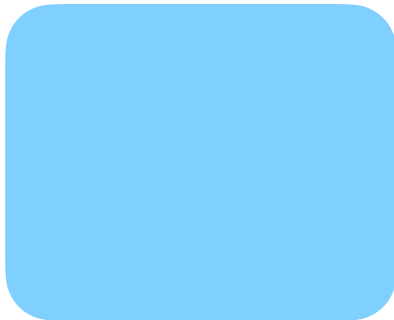
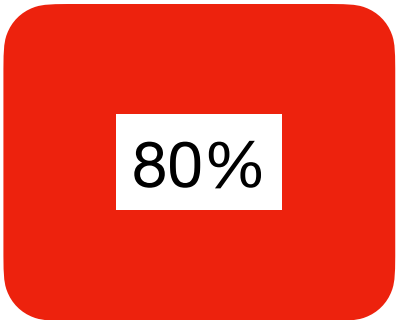
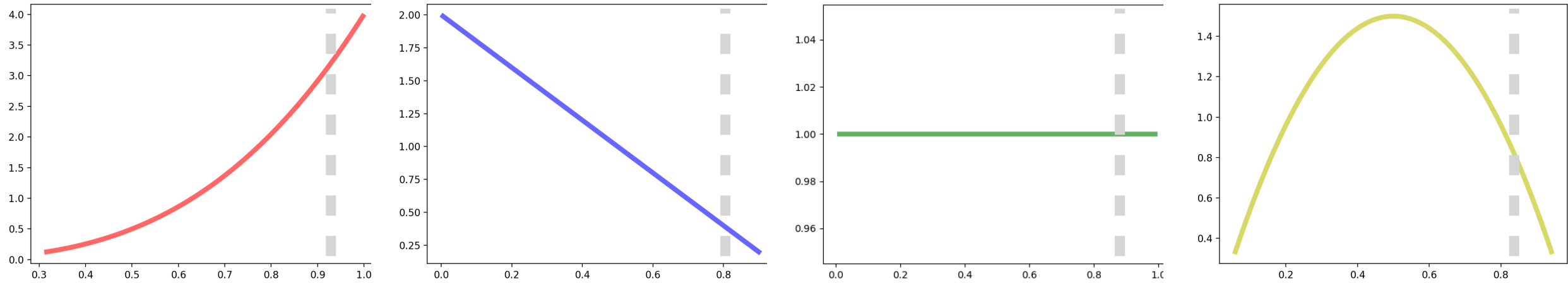
← **Stereotype Dispersion** →

More Dissimilar

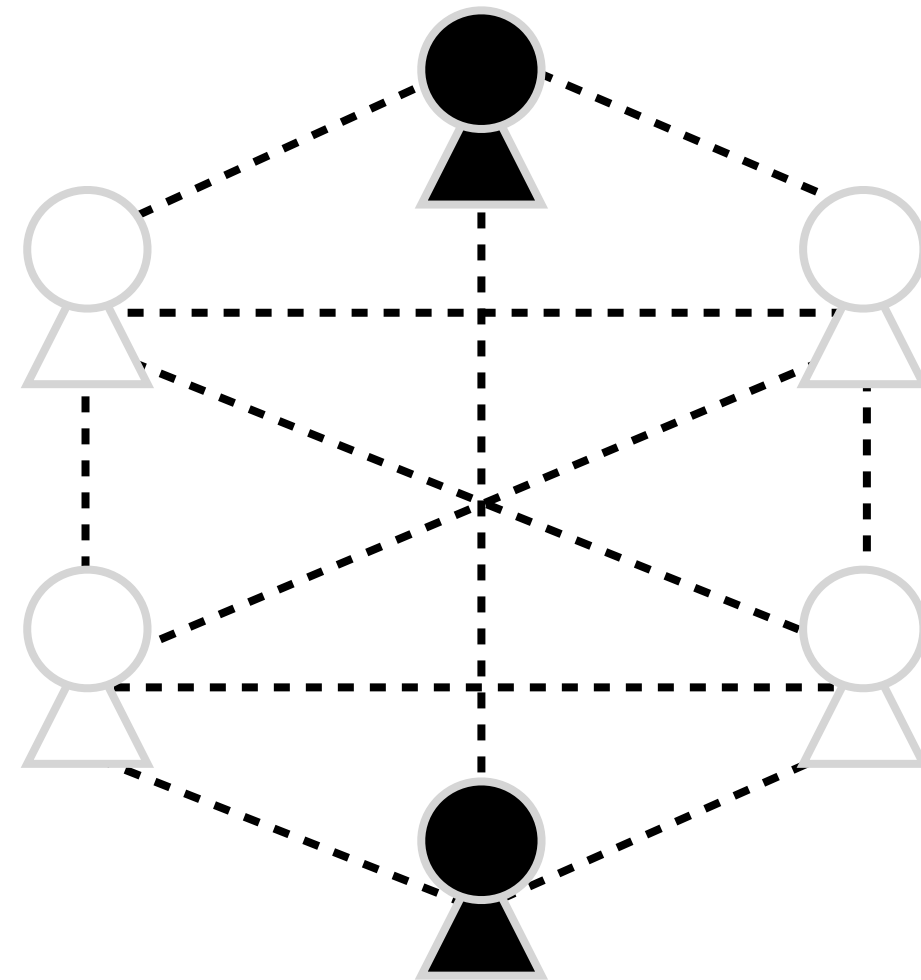
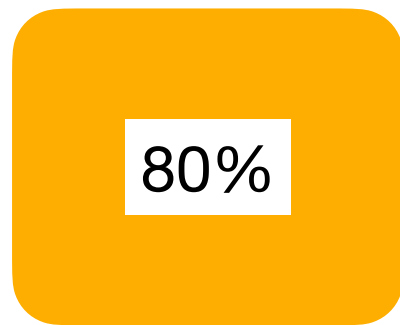
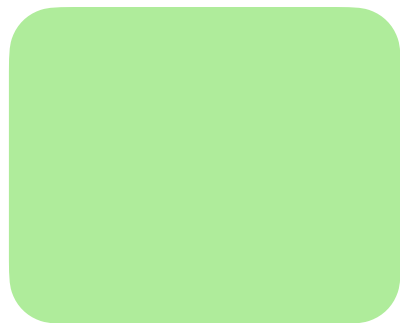
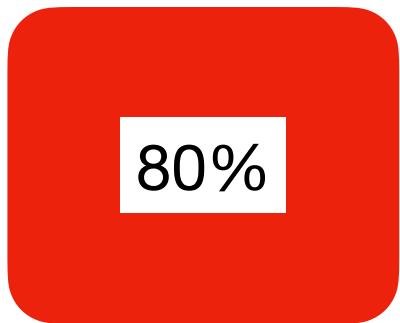
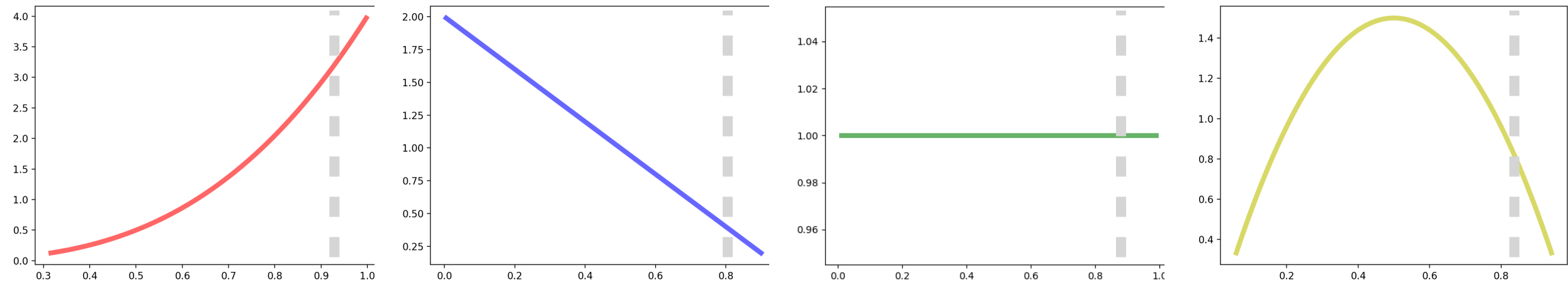


Here is a **collective** puzzle

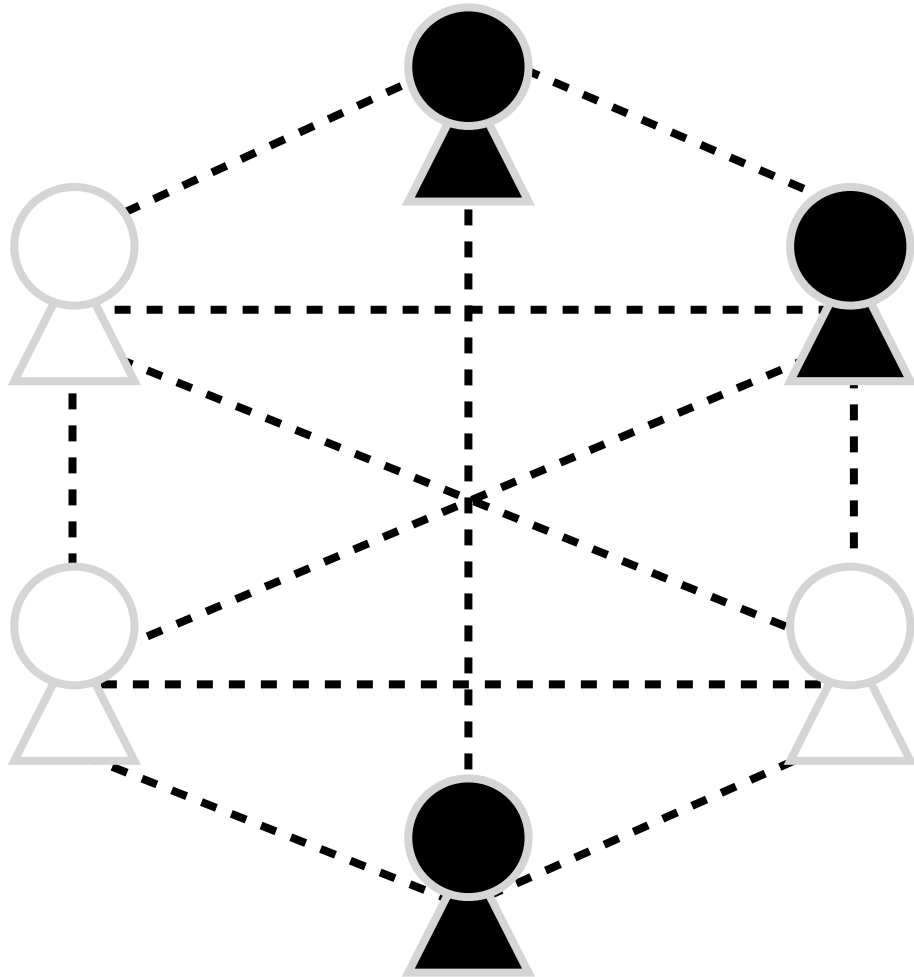
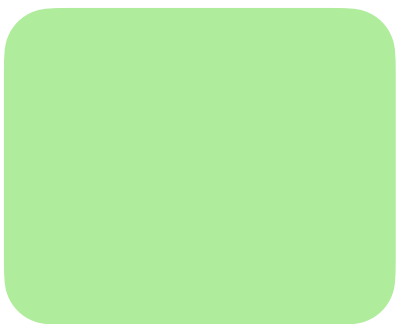
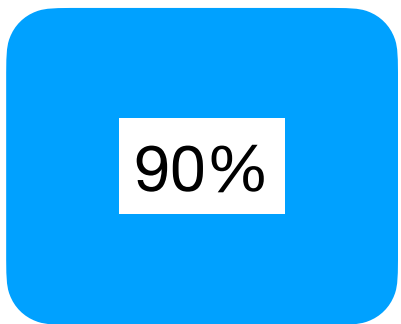
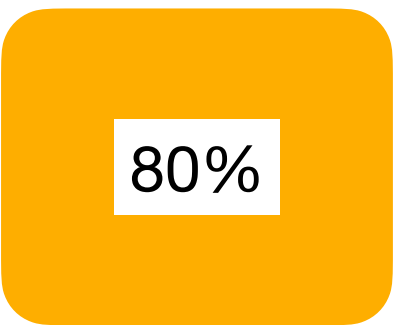
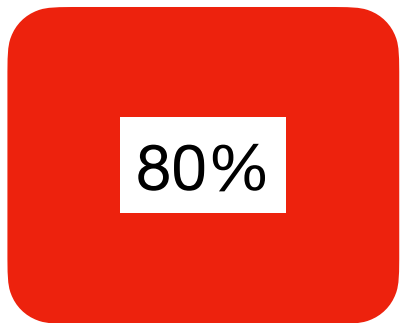
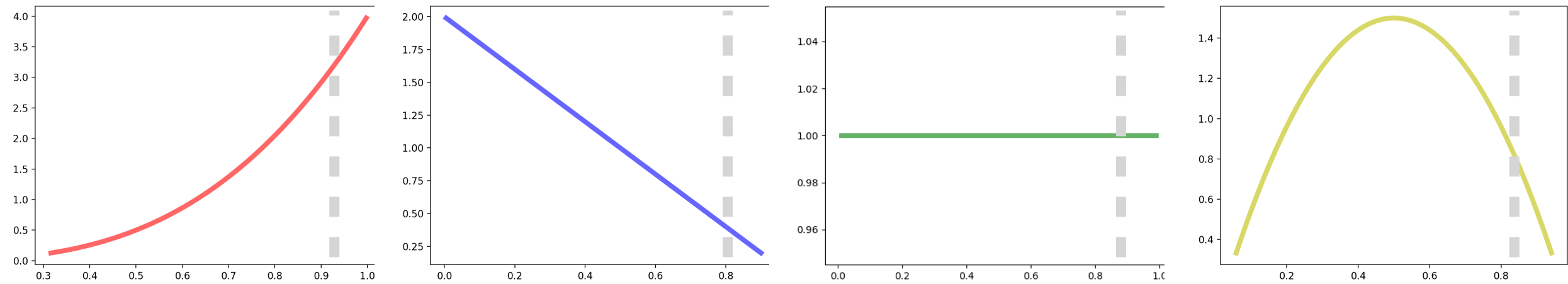
Here is a collective **puzzle**



Here is a collective puzzle

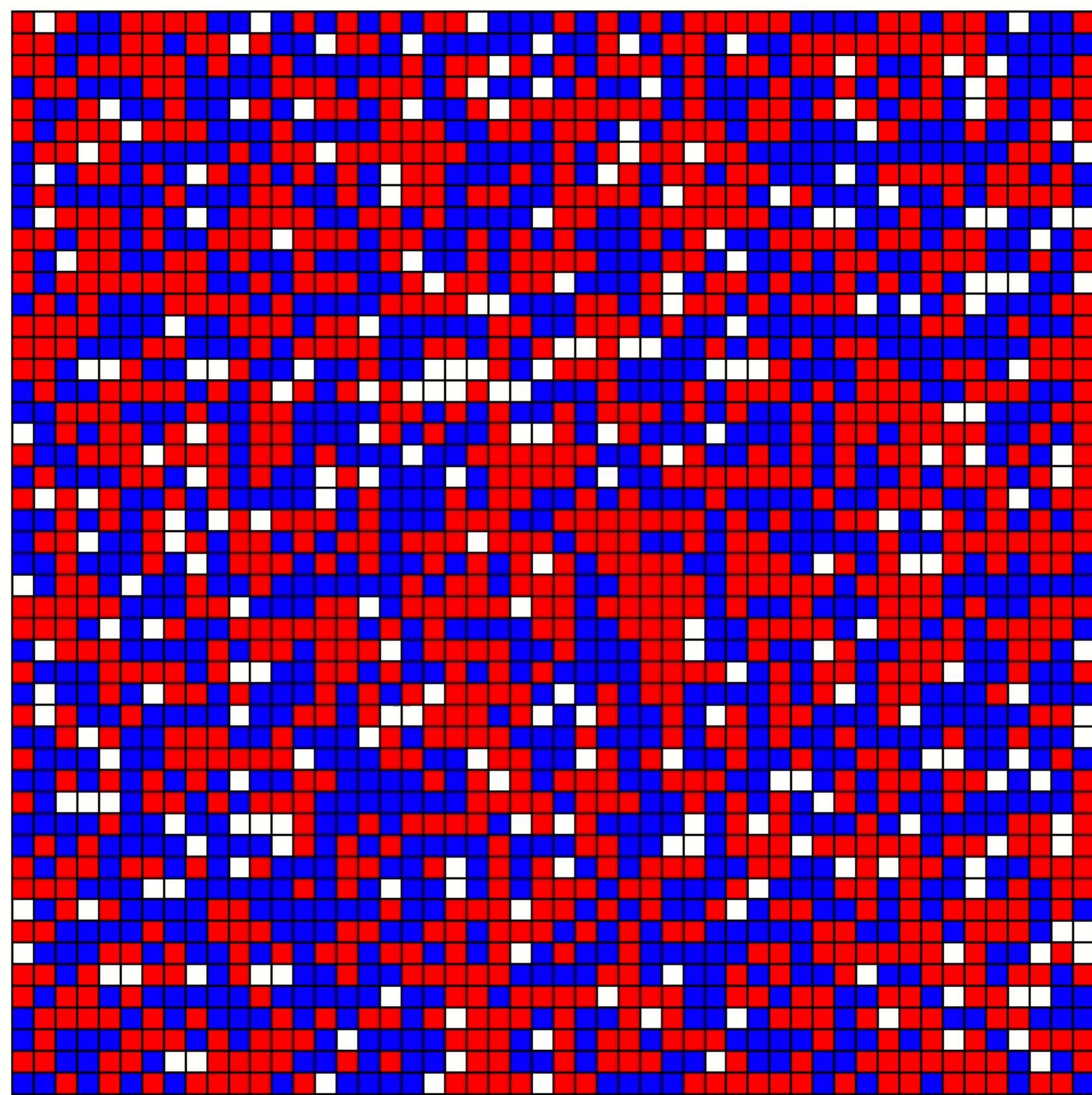


Here is a collective puzzle



Reverse-engineering the Society: Multi-agent

Here is a collective puzzle



Thomas Schelling
1921 - 2016



Reverse Engineering Mind and Society

Xuechunzi Bai

Assistant Professor of Psychology

University of Chicago