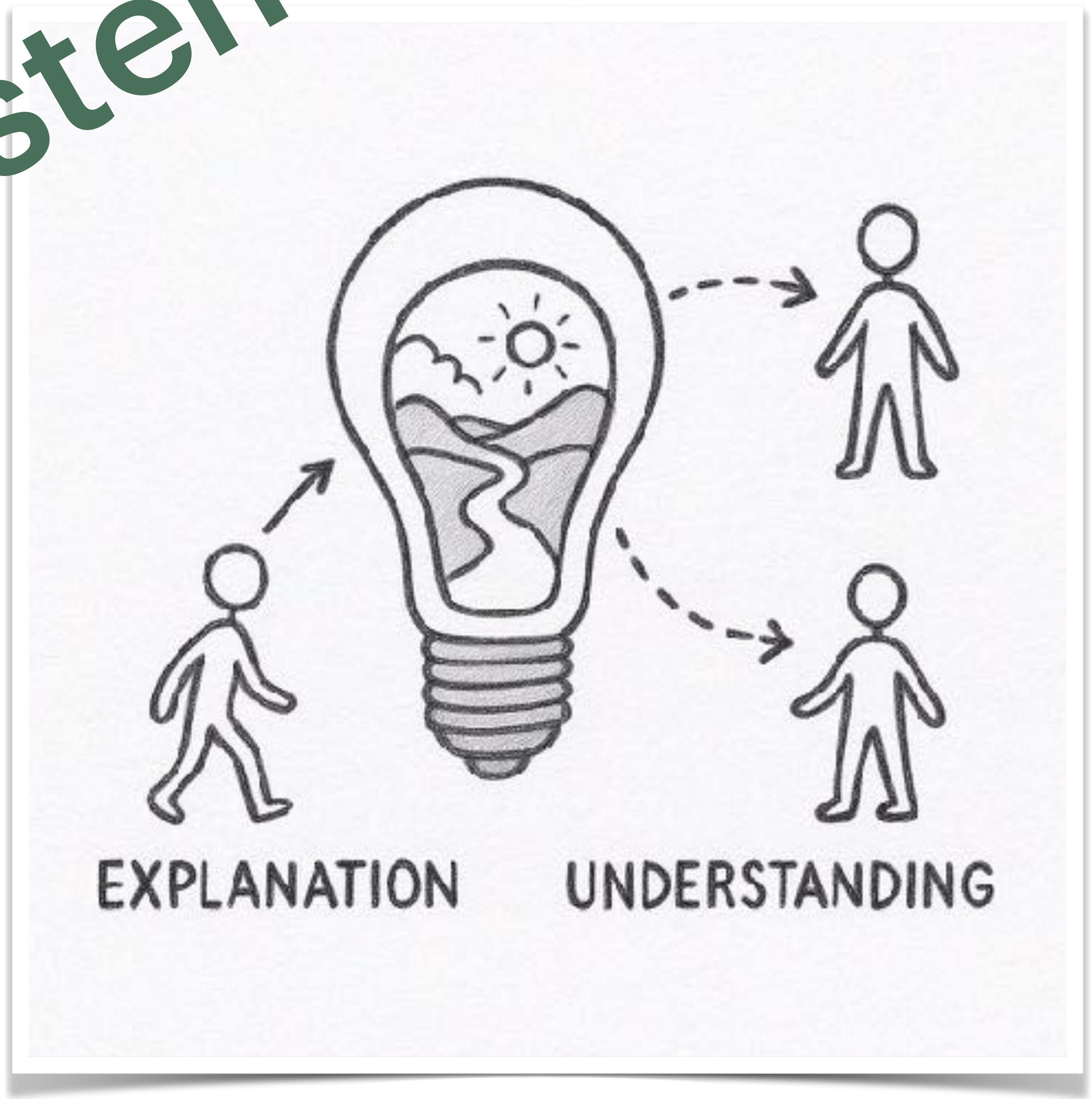
An Explanation Game: From Mechanistic Interpretability to Strategic Explanation Design

Krikamol Muandet
Rational Intelligence Lab
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany

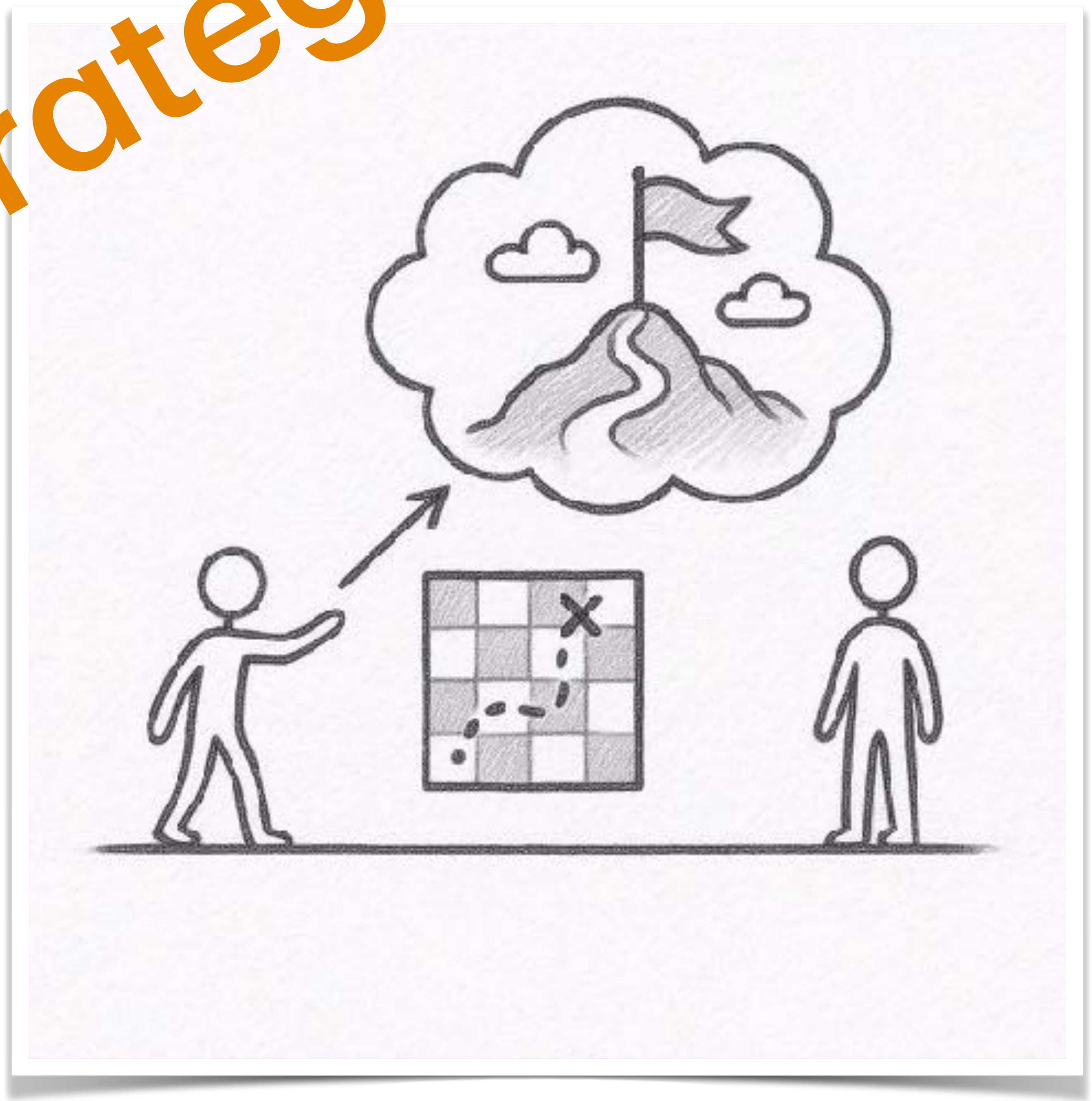
What Makes Great Explanation?

Epistemic



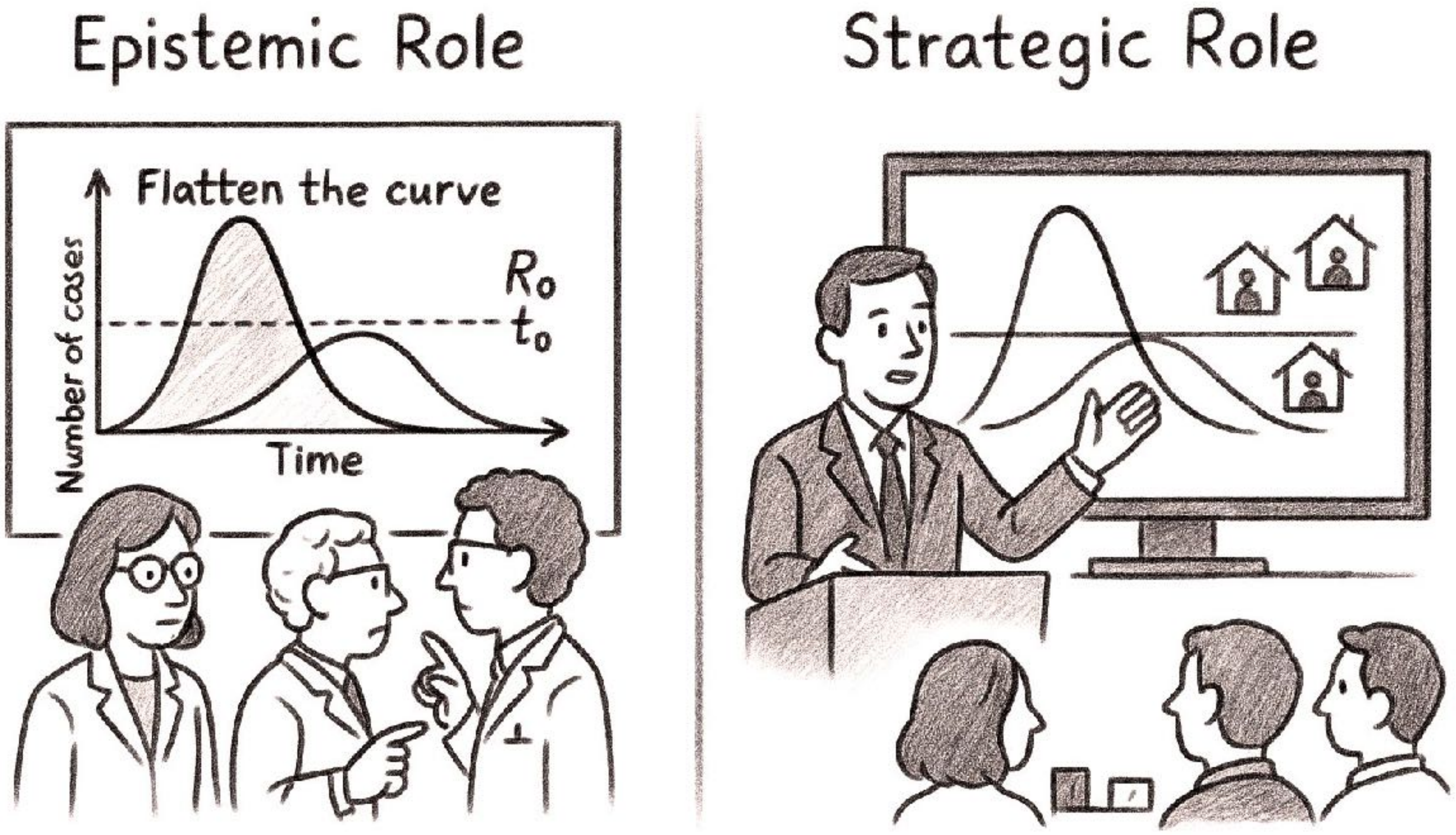
Explanation as understanding:
It turns information into understanding and knowledge.

Strategic



Explanation as influence:
It informs, convinces, and guides others toward desirable actions.

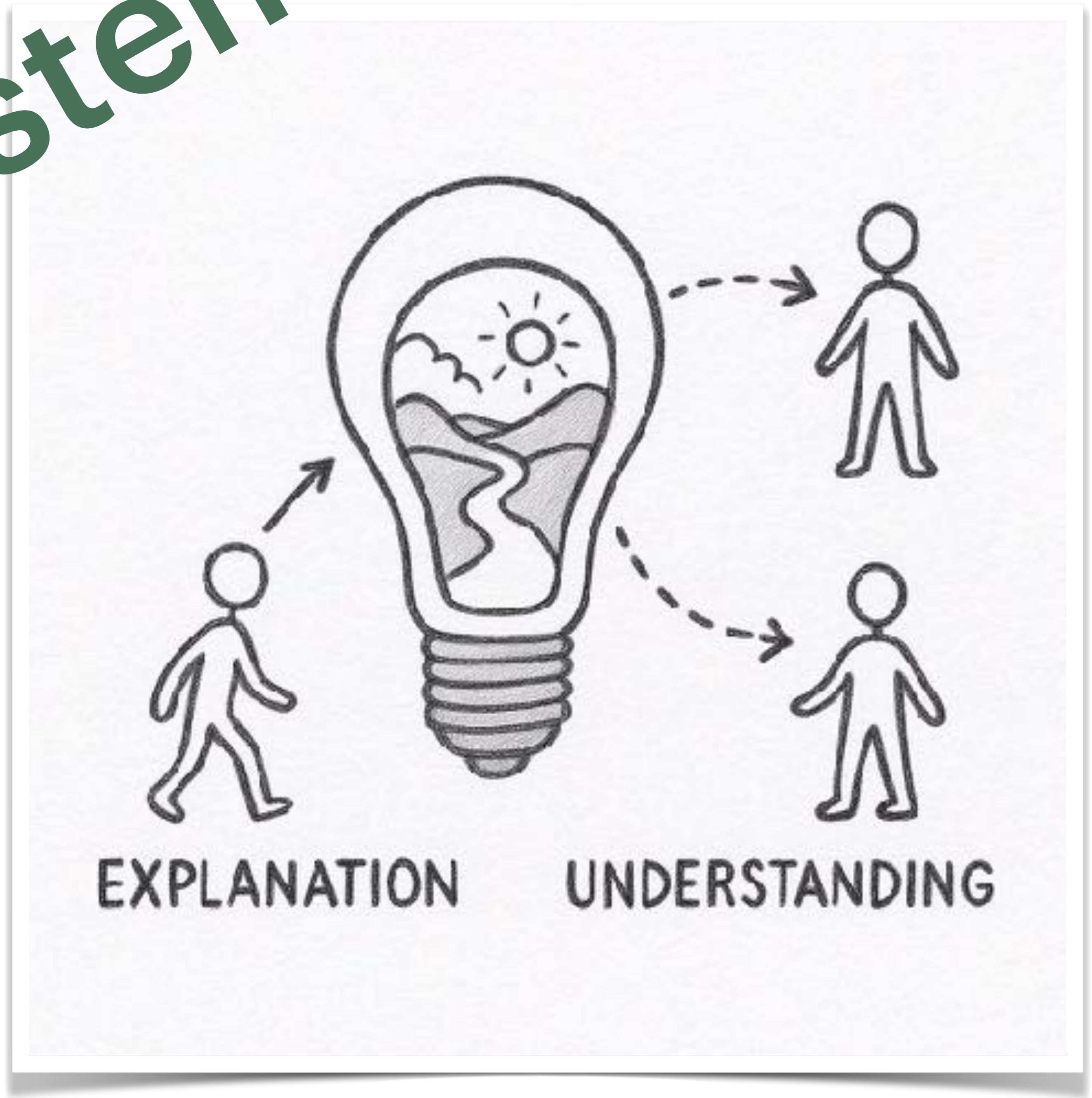
"Flattening the Curve"



Explanation is not a purely cognitive act but also a *socially strategic act of information design*.

What Makes Great Explanation?

Epistemic



Explanation as understanding:
It turns information into understanding and knowledge.

On the Relationship Between Explanation and Prediction: A Causal View


Amir-Hossein Karimi
Waterloo


Krikamol Muandet
CISPA


Simon Kornblith
Anthropic


Bernhard Schölkopf
MPI-IS


Been Kim
Google DeepMind

On the Relationship Between Explanation and Prediction: A Causal View

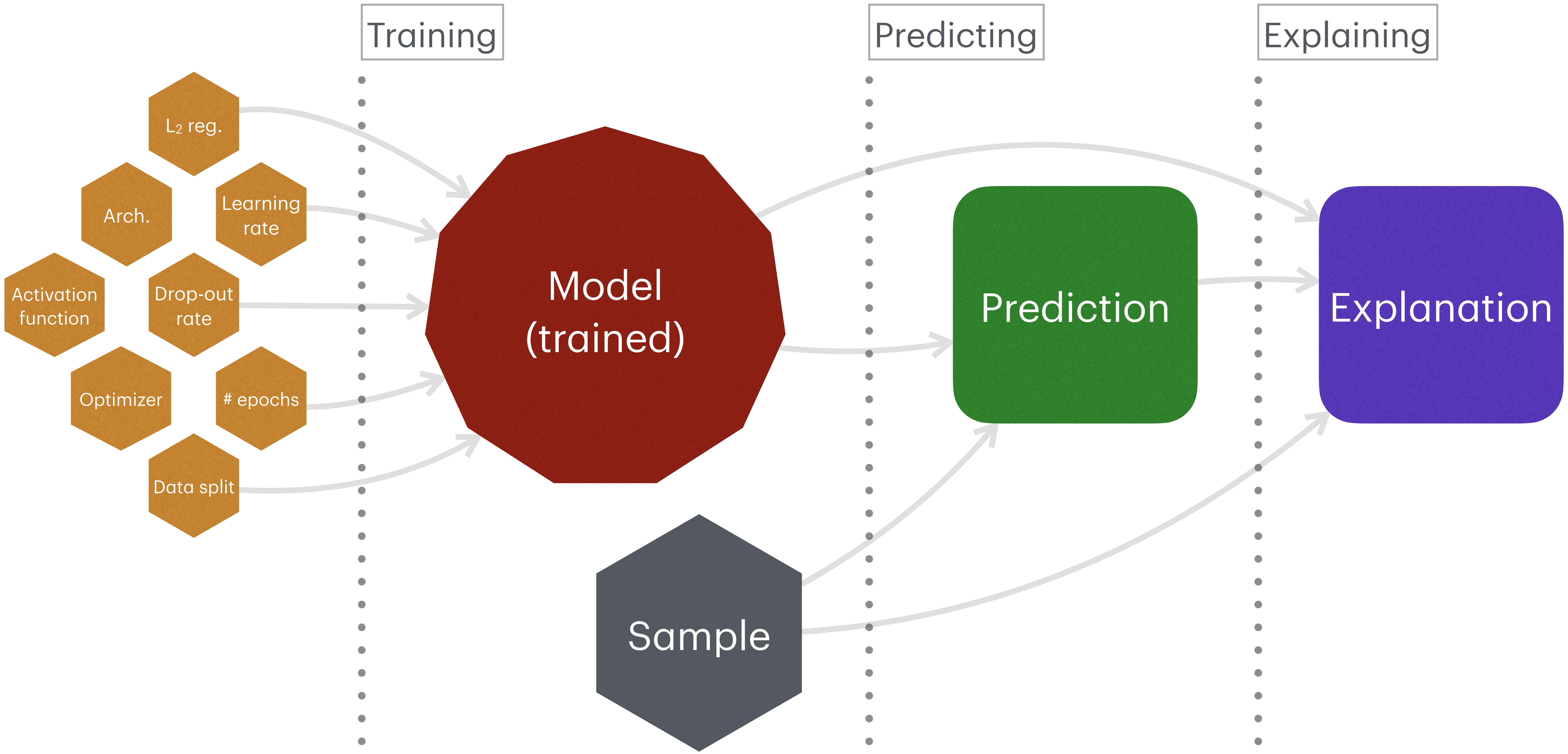
Amir-Hossein Karimi [1 2 3] **Krikamol Muandet** [4] **Simon Kornblith** [3] **Bernhard Schölkopf** [1] **Been Kim** [3]

Abstract

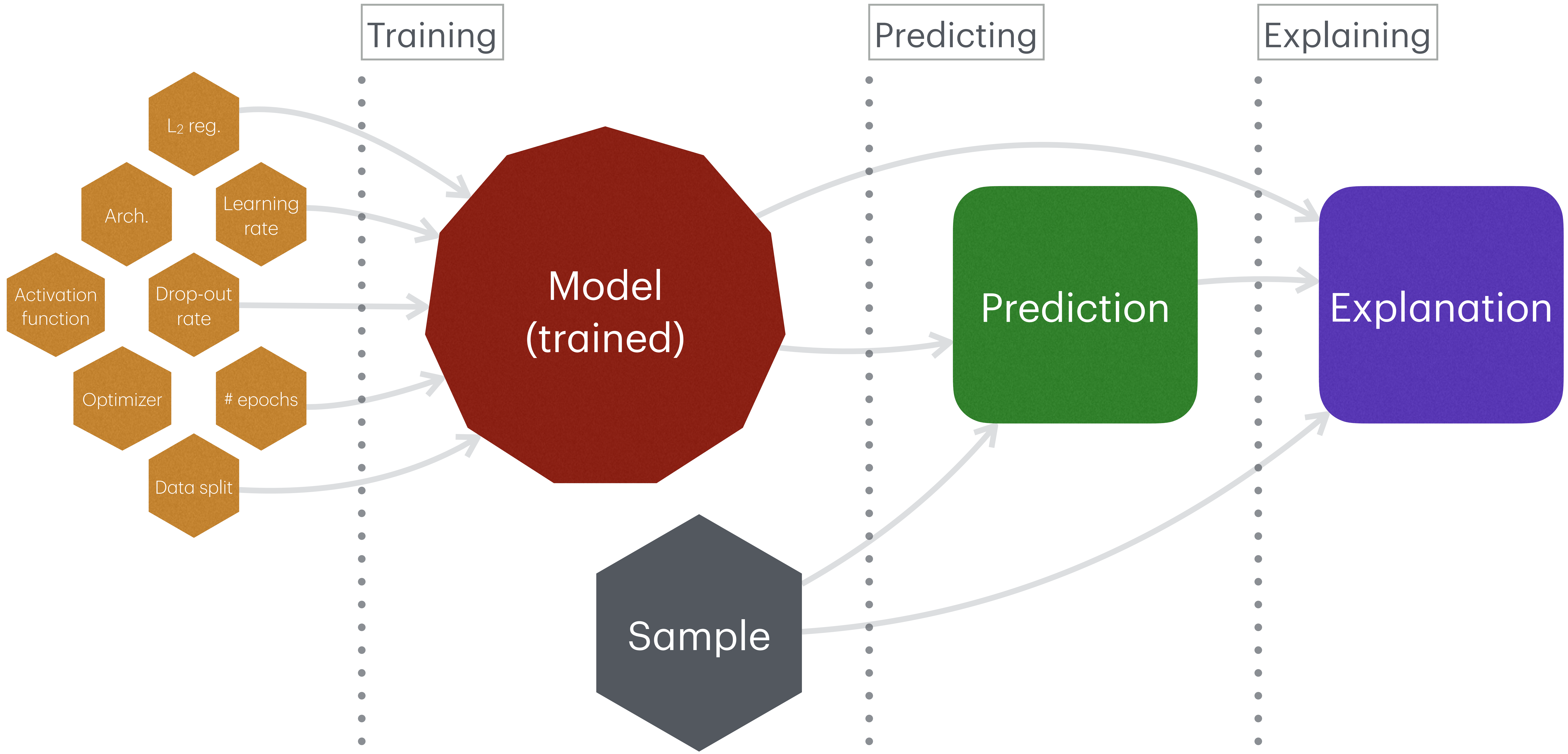
Being able to provide explanations for a model's decision has become a central requirement for the development, deployment, and adoption of machine learning models. However, we are yet to understand what explanation methods can and

to influence the model's decision (Koh et al., 2020; Bau et al., 2020; Meng et al., 2022), but also to ensure that models comply with regulatory requirements (Parliament & of the European Union, 2016). However, Existing tools for interpretability have however elicited criticisms, often highlighting computational or qualitative user-study-based

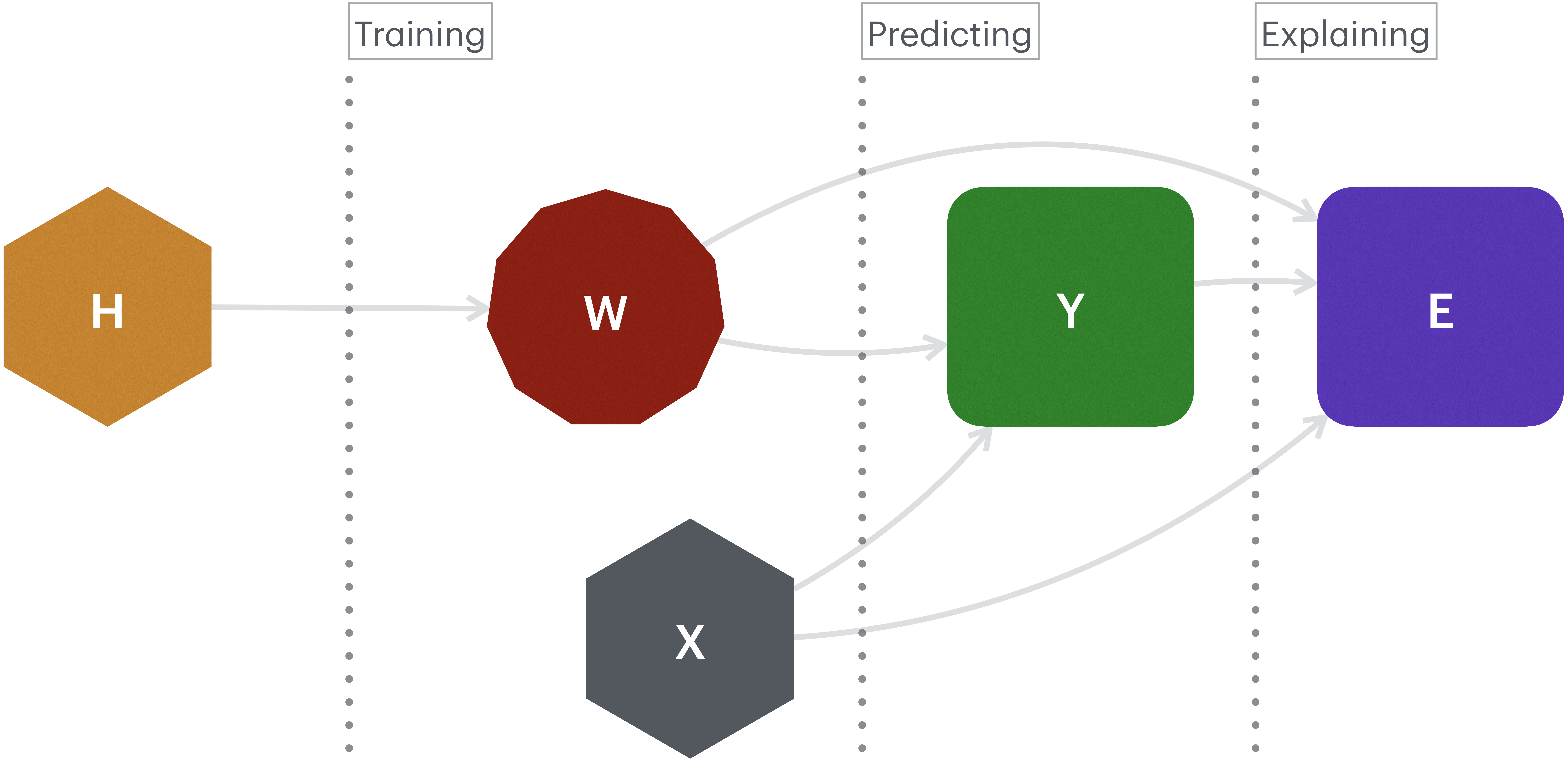
Explanation Generation Process



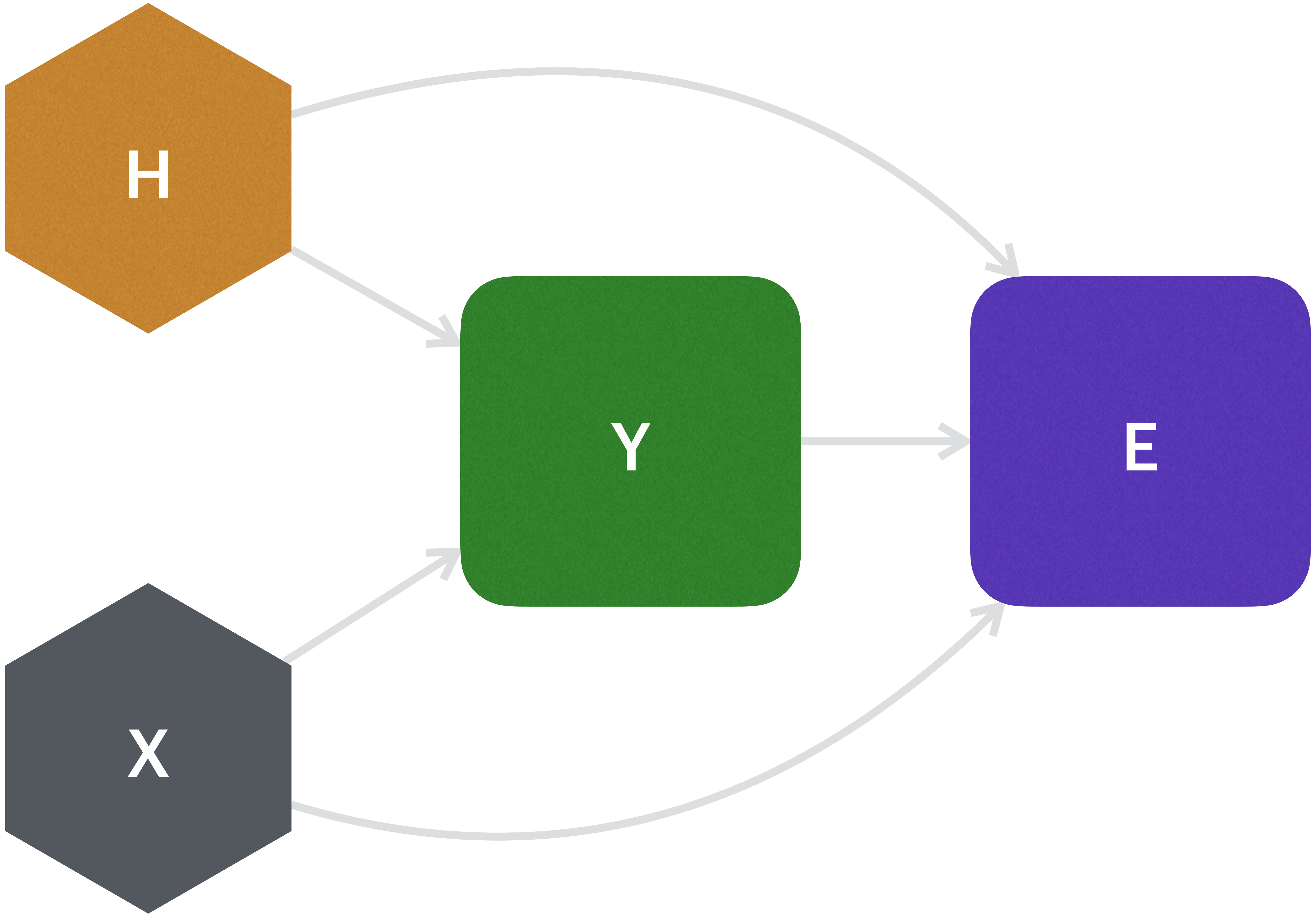
Explanation Generation Process



Explanation Generation Process

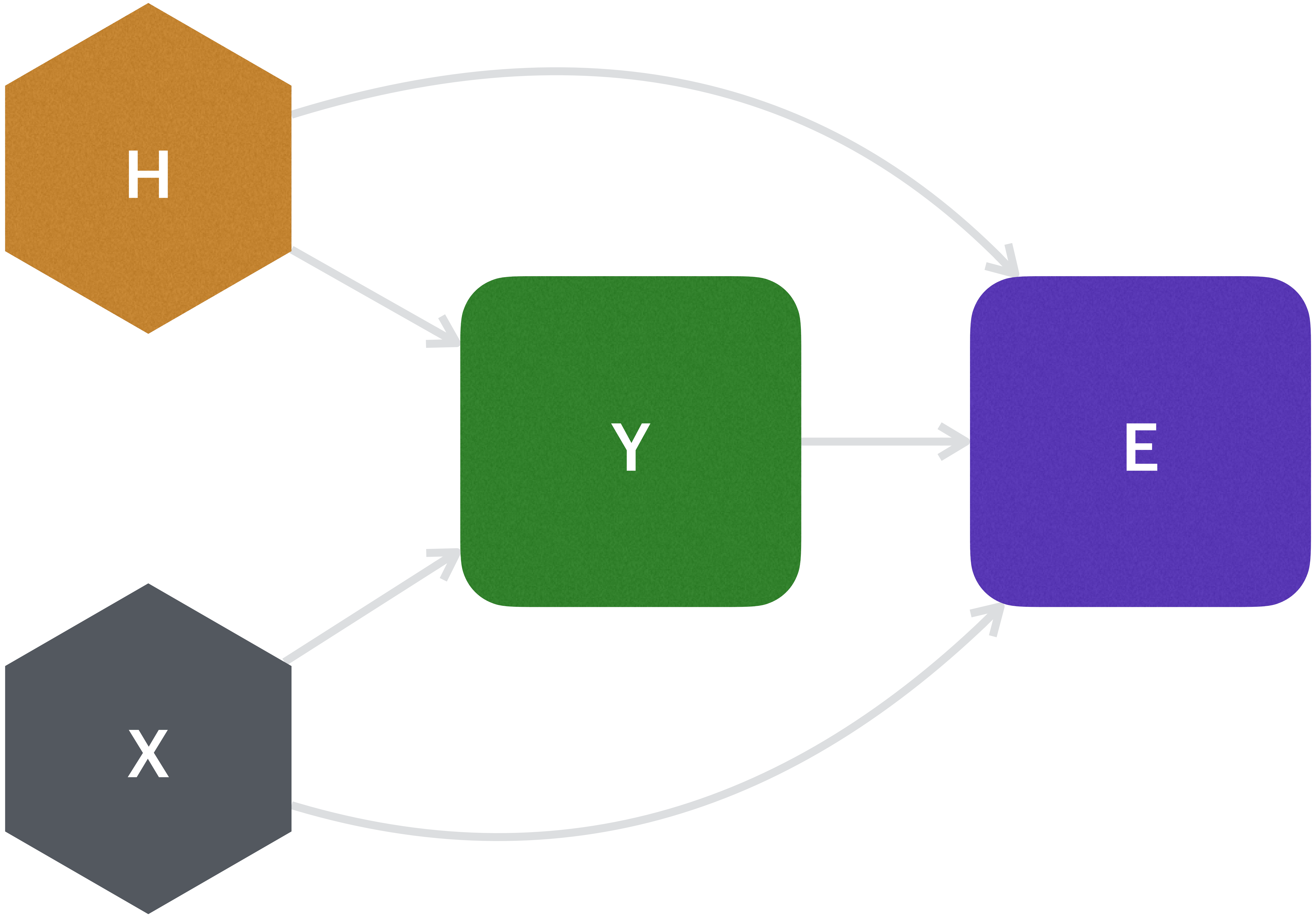


Hyperparameters as Treatments



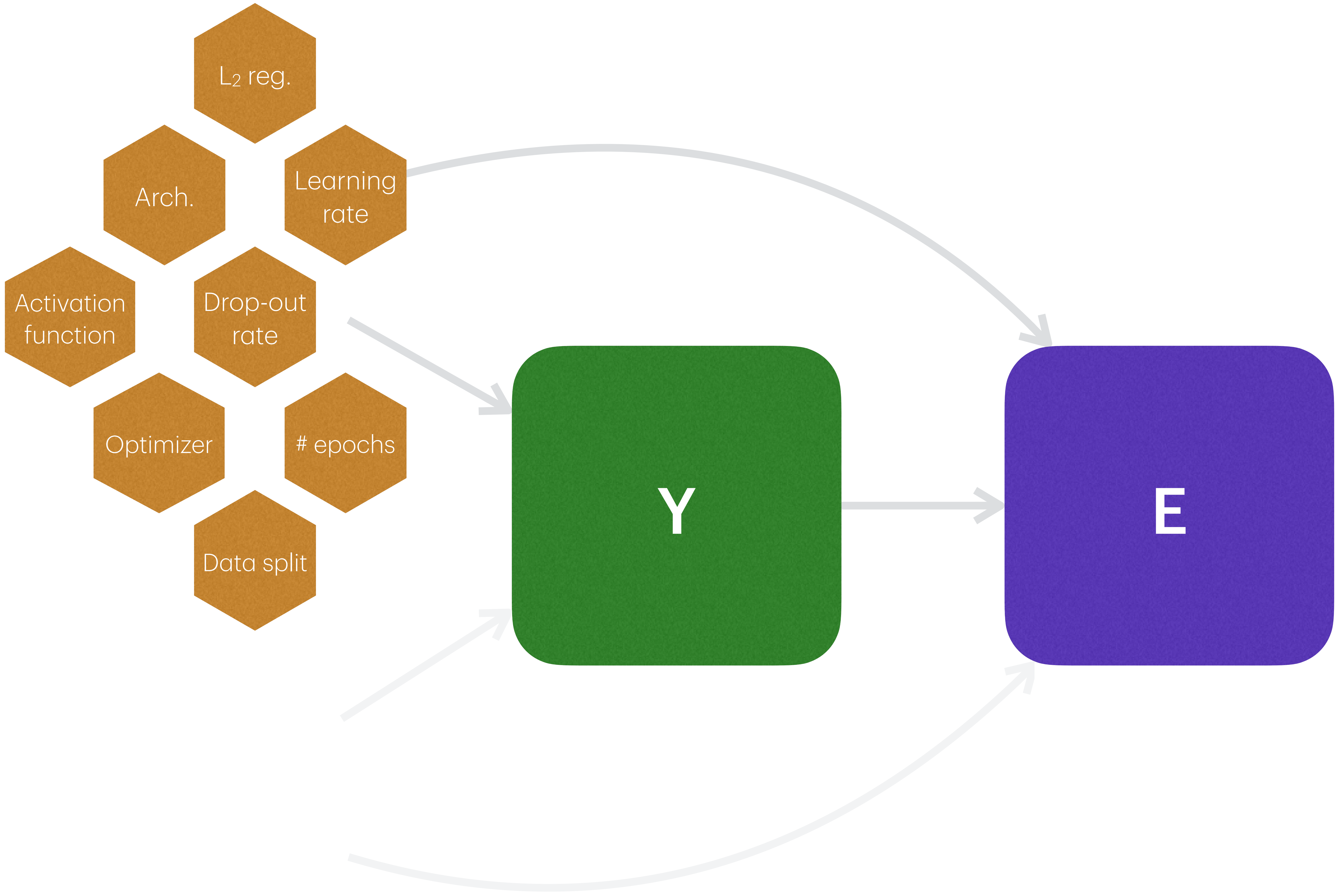
What is the effect of the hyperparameters on the resulting prediction/explanation?

Hyperparameters as Treatments



What does the **prediction/ explanation** for X = x look like, if the **hyperparameters** take on value H = h rather than H = h’, *all else being equal?*

Extended Treatment Effects



What does the prediction/explanation for $X = x$ look like, if the hyperparameters take on value $H = h$ rather than $H = h'$, *all else being equal?*

$Y_{h=1} - Y_{h=0}$
single binary treatment

$\mathbb{E}_{m\neq n} [Y_{h=n} - Y_{h=m}]$
single non-binary treatment

$\mathbb{E}_{h\setminus i} [\mathbb{E}_{m\neq n} [Y_{h_i=n,\, h\setminus i} - Y_{h_i=m,\, h\setminus i}]]$
multiple non-binary treatment

$\mathbb{E}_{h\setminus i} [\mathbb{E}_{m\neq n} [\| \varphi(Y_{h_i=n,\, h\setminus i}) - \varphi(Y_{h_i=m,\, h\setminus i}) \|_G]]$
multiple non-binary treatments
& a non-binary target

Model Zoo & Explanations

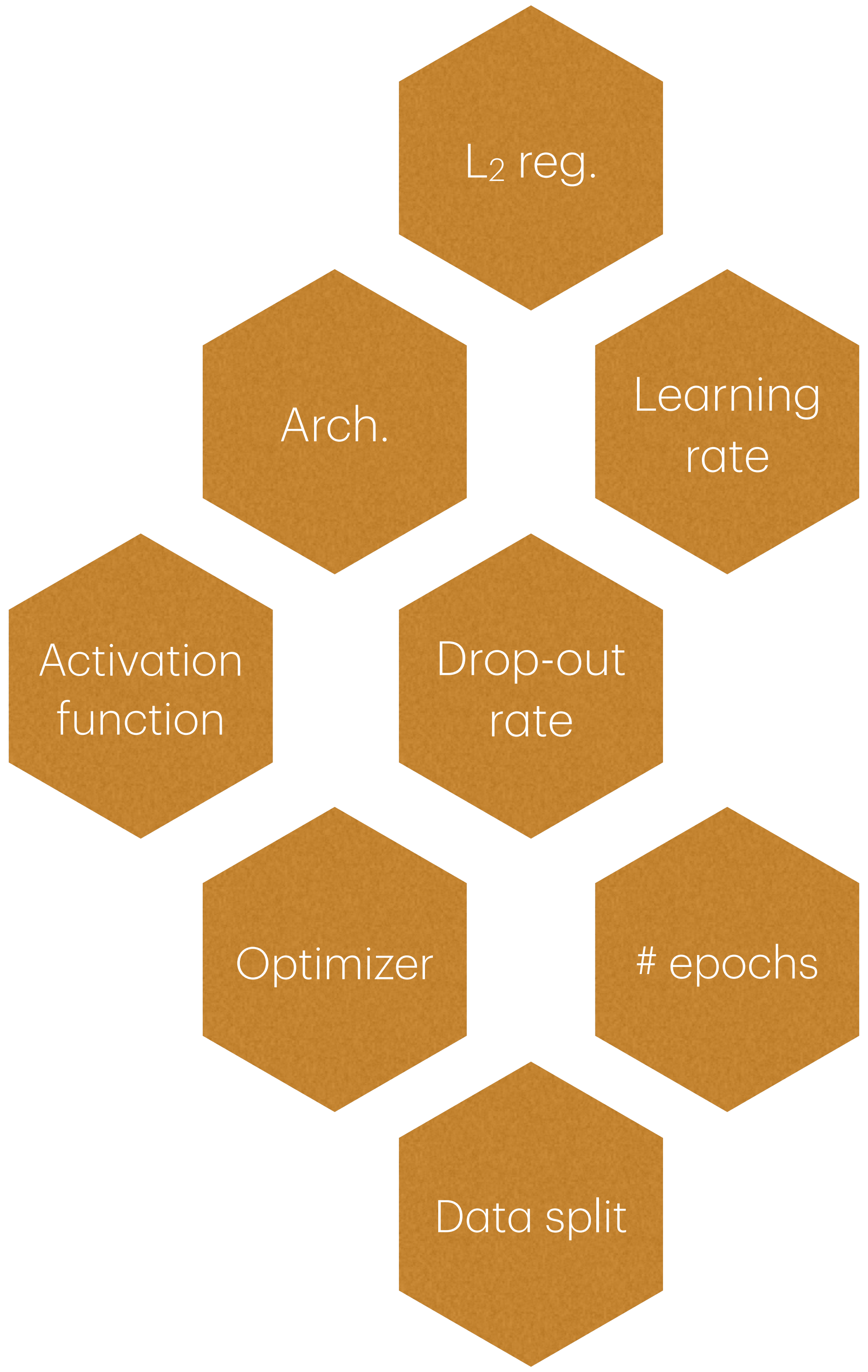
30,000 pre-trained models:
3 layer CNNs (4,970 parameters);
trained to convergence (max 86 epochs)

4 datasets:
MNIST, FASHION, SVHN, CIFAR10

8 hyparparameters:
drawn “independently at random” from pre-specified ranges

Fixed architecture. Fixed random seed.

4+1 saliency-based explanations:
Gradient, SmoothGrad, Integrated Gradients, Grad-CAM
Reference explanation: “identity”, i.e., E = Y —> ITE_E = ITE_Y



[Unterthiner et al. 2020]

Most types of H influence Y (and E) in a similar way

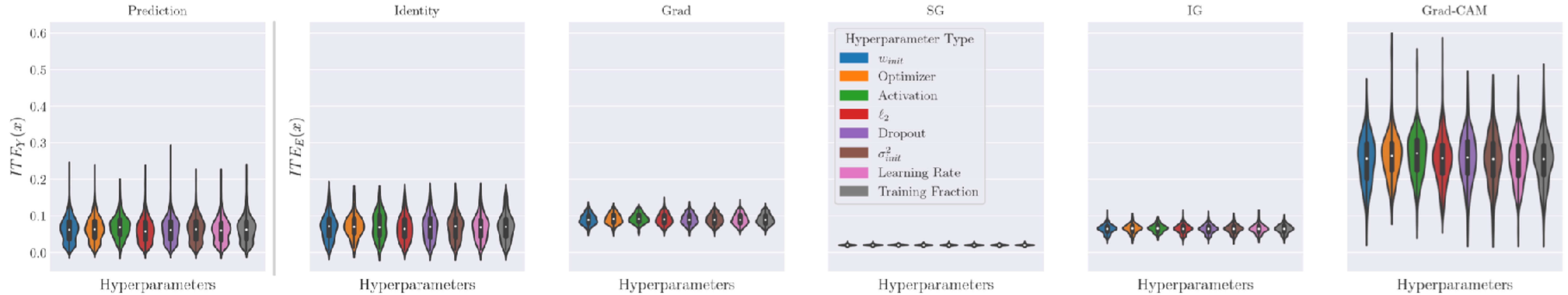


Figure 3: Comparison of ITE_Y and ITE_E for CIFAR10 shows that different types of H influence E and Y in a similar way.

H influences Y (and E) differently across performance buckets

Performance buckets:

- 0 - 20 pctl.
- 20 - 40 pctl.
- 40 - 60 pctl.
- 60 - 80 pctl.
- 80 - 90 pctl.
- 90 - 95 pctl.
- 95 - 99 pctl.
- 99 - 100 pctl.

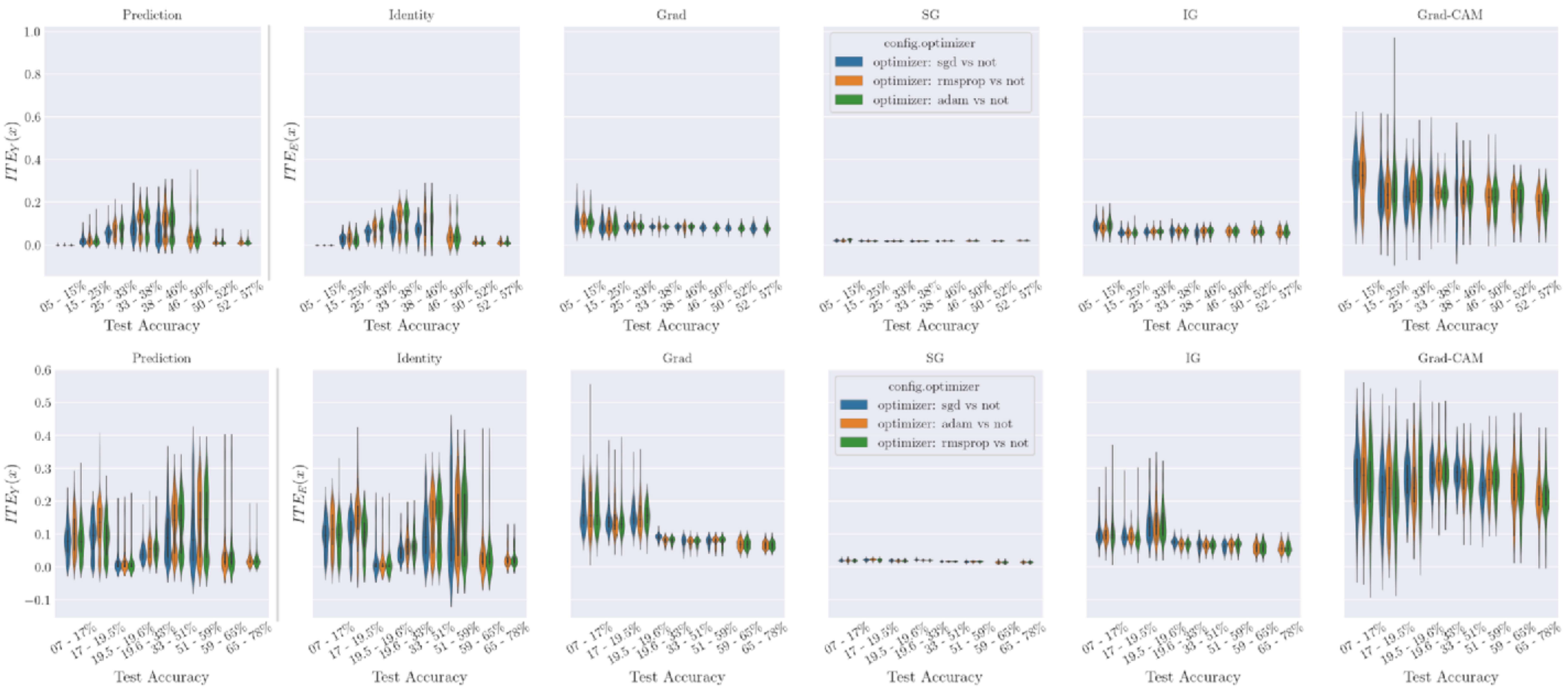


Figure 4: Comparison of ITE values of $h_{\text{optimizer}}$ on Y (left) and E (right) for models across different performance buckets, showing the discrepancy in the effect of H on Y vs. that on E (top: CIFAR10; bottom: SVHN). Interestingly, there is a difference of ITE_E across accuracy buckets, and more importantly, none of the explainability methods resemble ITE_Y.

Explanations may still be explaining something other than the prediction

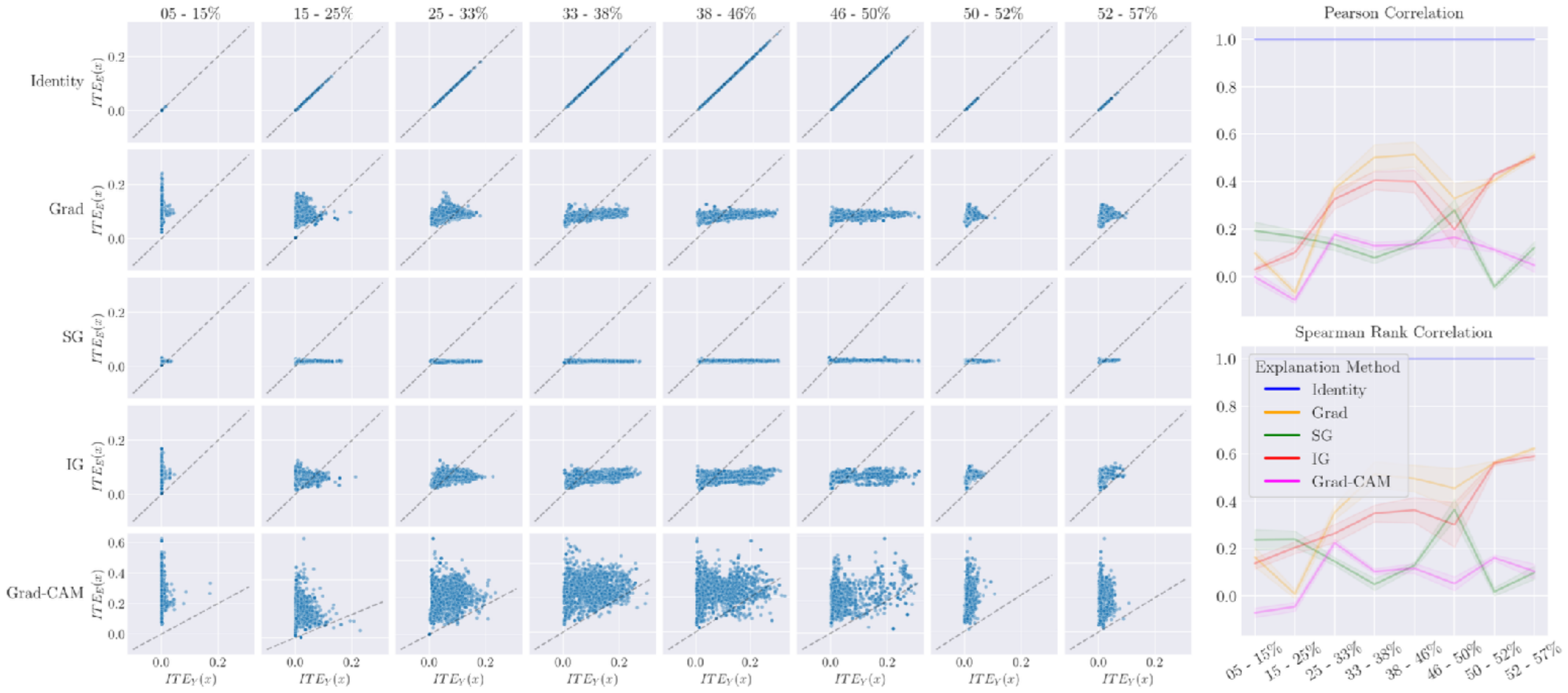
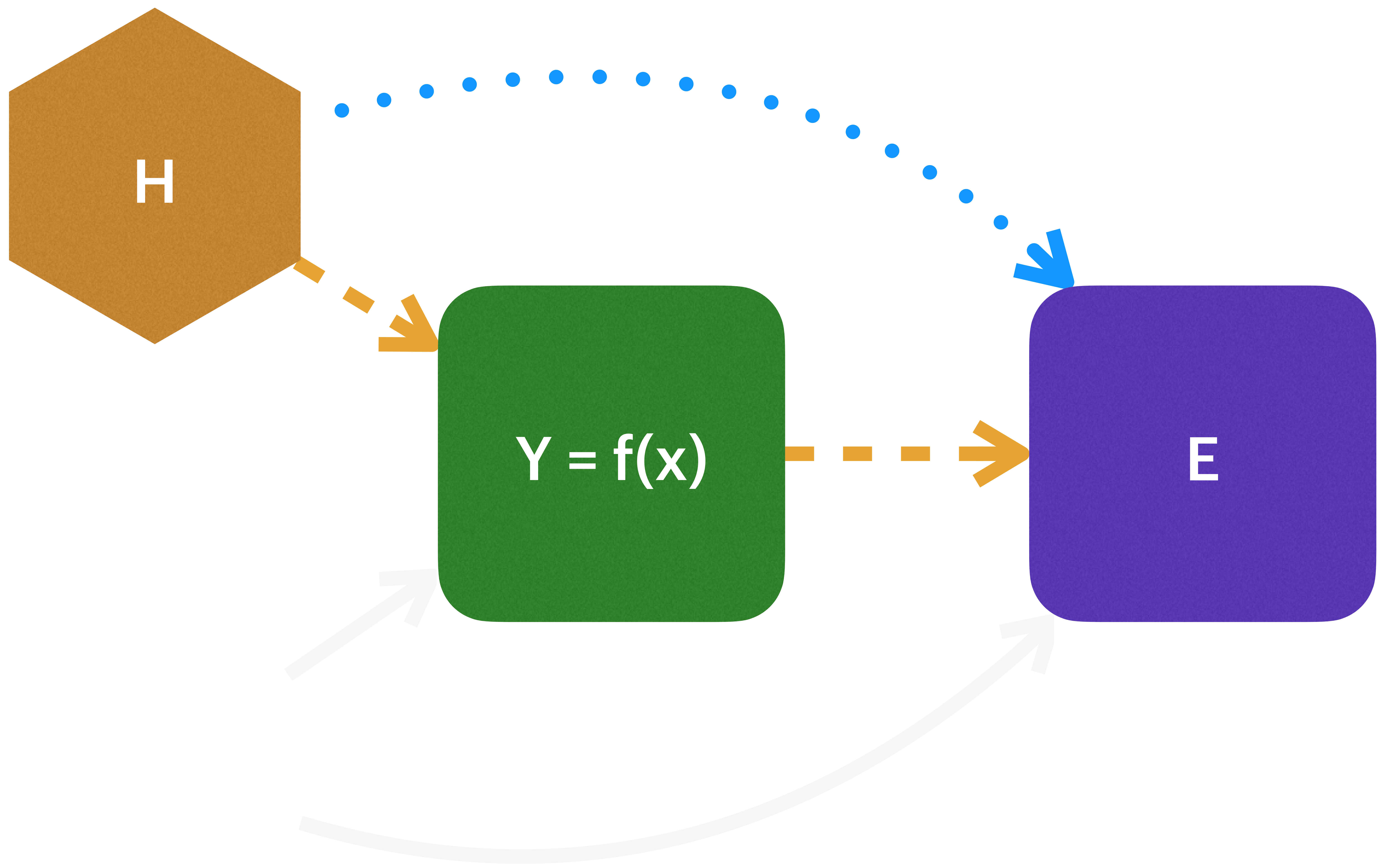


Figure 5: (left) Each column is a subset of models at each accuracy bucket, each row is a different explanation method. Whereas low-performing CIFAR10 models (first column) show little change in predictions as their explanations differ, top-performing models show the reverse of this trend. (right) Correlation measures of the scatter plots on the left show a decreased correlation in the top 1% models.

Direct vs Indirect Effects

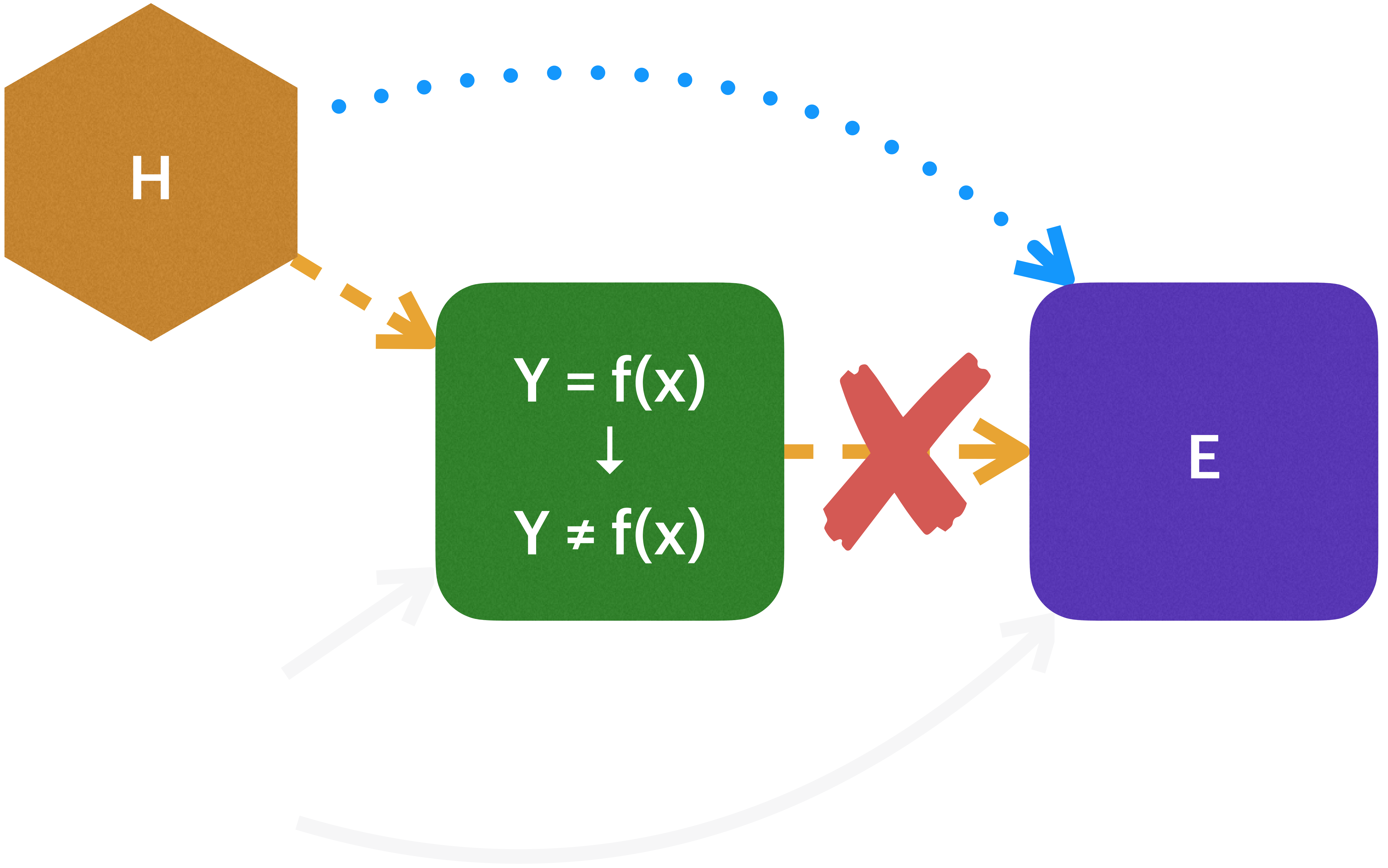


ITE_E measures the **total** effect:

* **direct effect**
* **indirect effect**

How to tease them apart?

Direct vs Indirect Effects



$\mathrm{ITE}_{\mathbb{E}}$ measures the **total** effect:

* **direct effect**
* **indirect effect**

How to tease them apart?

We can sever the flow of dependence from **H** to **E** by randomising **Y**

* **total effect**: $\mathrm{ITE}_{\mathbb{E},\ y=f(x)}$
* **direct effect**: $\mathrm{ITE}_{\mathbb{E},\ y\neq f(x)}$
* **indirect effect**: Δ above

Explanations from the highest performing models may be comparatively less reliable

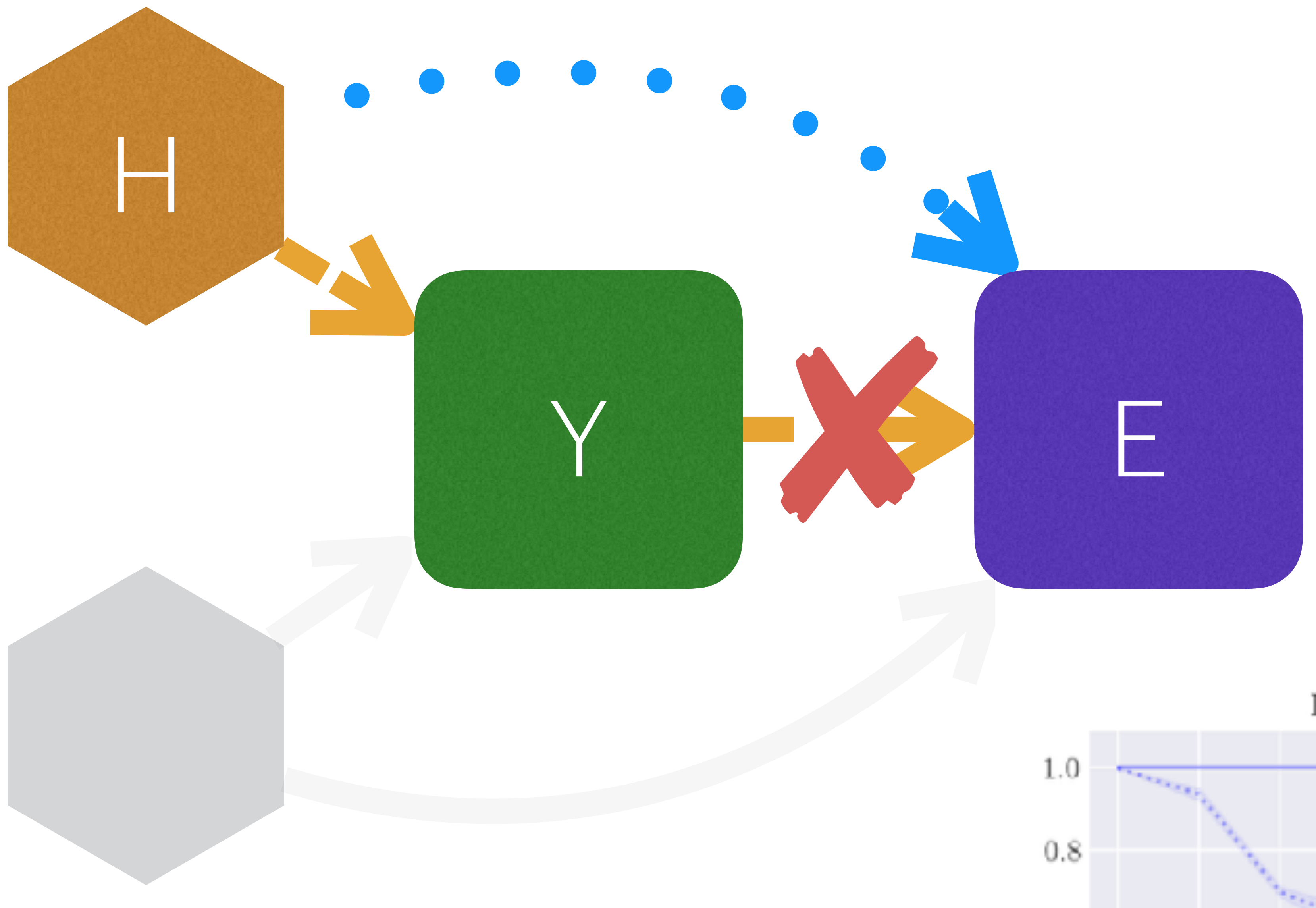
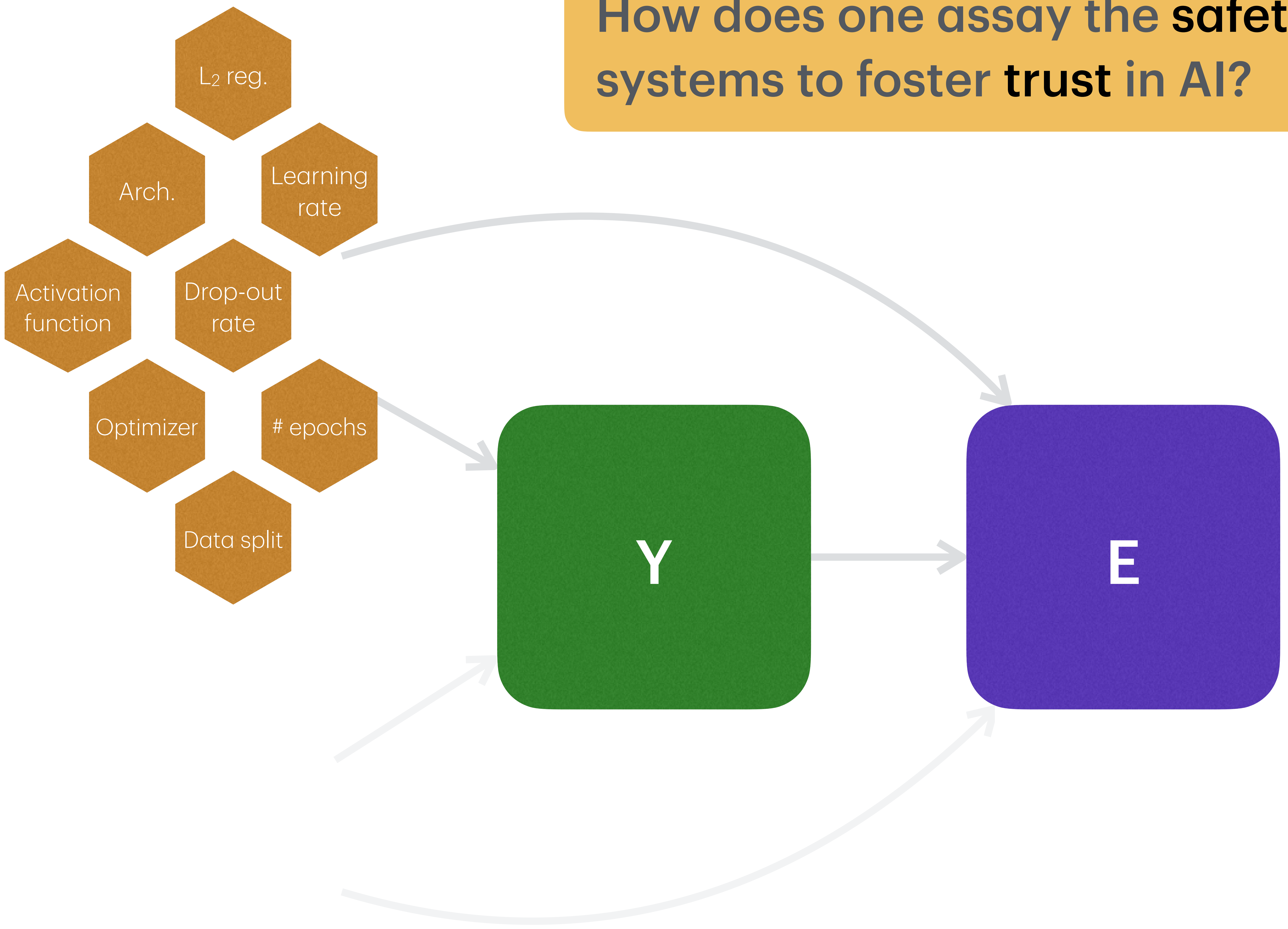


Figure 6: Pearson correlation between ITE_Y and ITE_E in total and direct effect (first column). The second column is the difference between total and direct effect, where higher values mean that the influence of H on E flows more through Y (ideal). The third column plots the difference of delta correlations between ideal case (Identity) and each method. In other words, it indicates how far each method moves away from ideal case, as a model performs better.

How does one assay the safety, factuality, and ethics of AI systems to foster trust in AI?



Common answer: use **explanations**

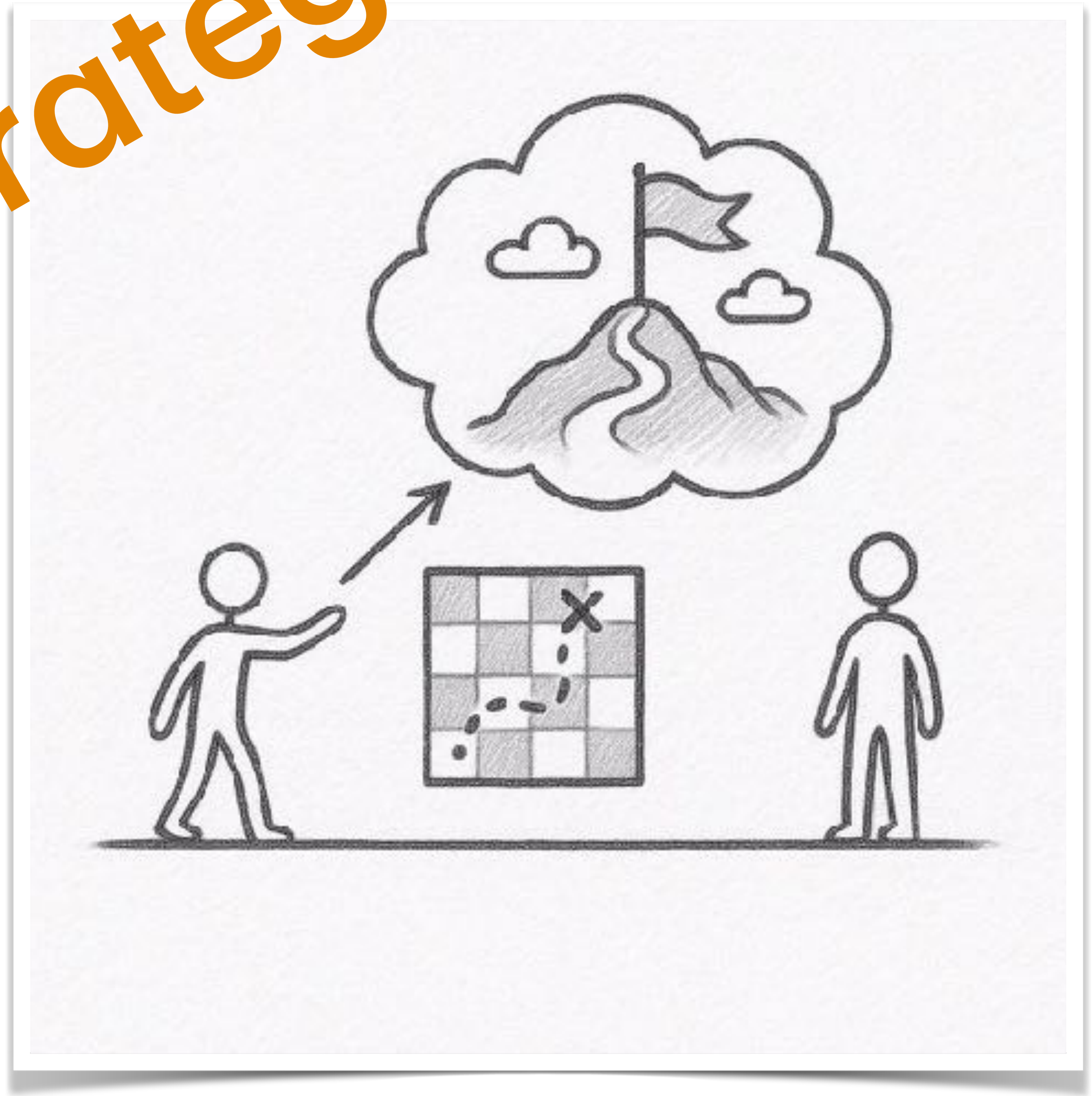
- **Cautionary tale**: explanations may still be explaining something other than the prediction
- We propose a **causally-grounded quantitative metric** to study the relationship between predication and explanation

Future work:

- Extension **beyond saliency map** e.g., SHAP, LIME, recourse, etc.
- Creating a OSS tool to measure causal effect of Y on E for any given black-box model

What Makes Great Explanation?

Strategic



Explanation as influence:
It informs, convinces, and guides others toward desirable actions.

Causal Strategic Learning with Competitive Selection

Kiet Vo
CISPA

Muneeb Aadil
UCL

Siu Lun Chau
NTU

Krikamol Muandet
CISPA

AAAI

Causal Strategic Learning with Competitive Selection

A Preprint

Kiet Q. H. Vo[1,2], Muneeb Aadil[1,2], Siu Lun Chau[1], Krikamol Muandet[1]

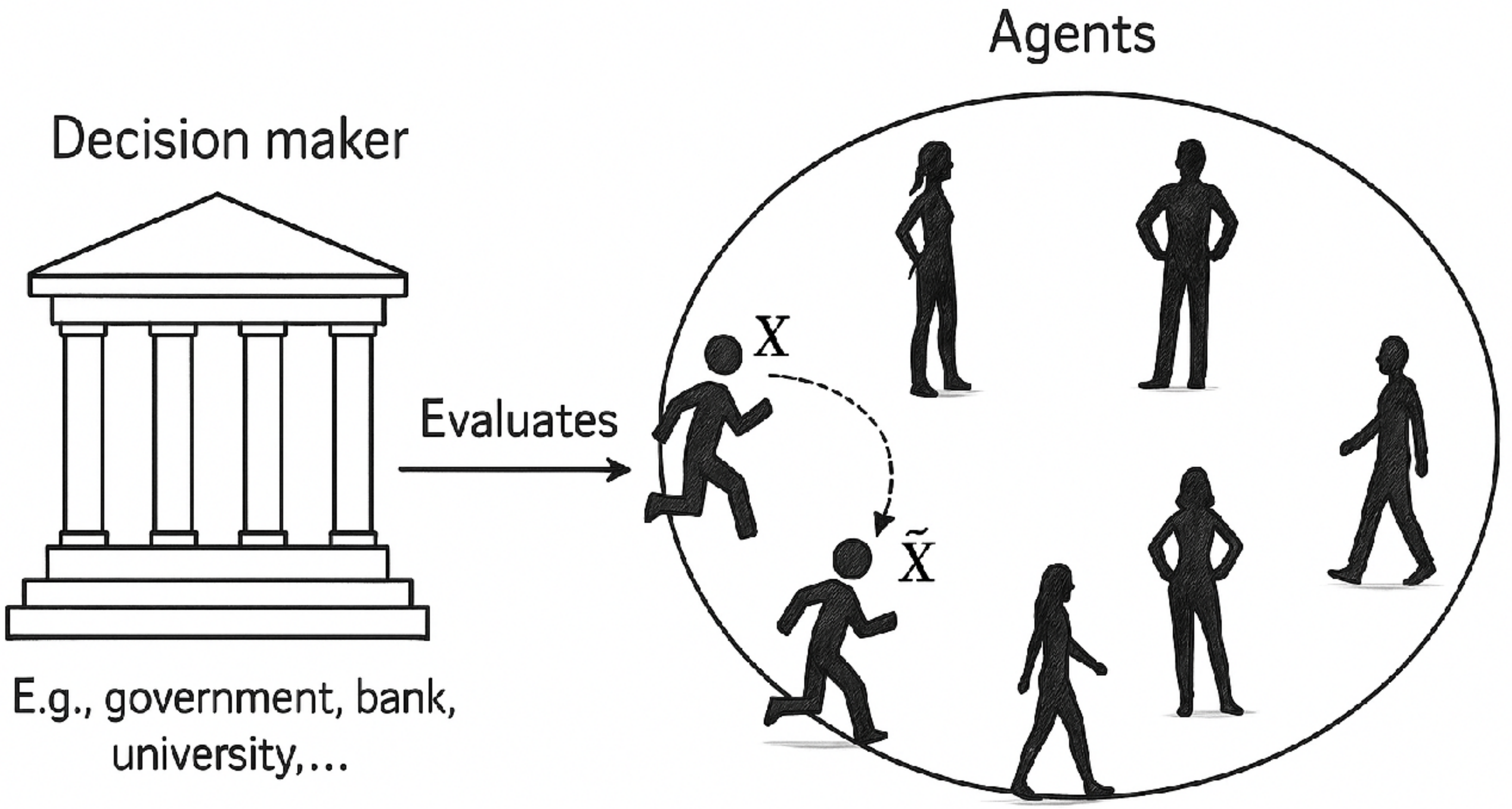
[1]CISPA Helmholtz Center for Information Security, Saarbrücken, Germany
[2]Saarland University, Saarbrücken, Germany

February 6, 2024

Abstract

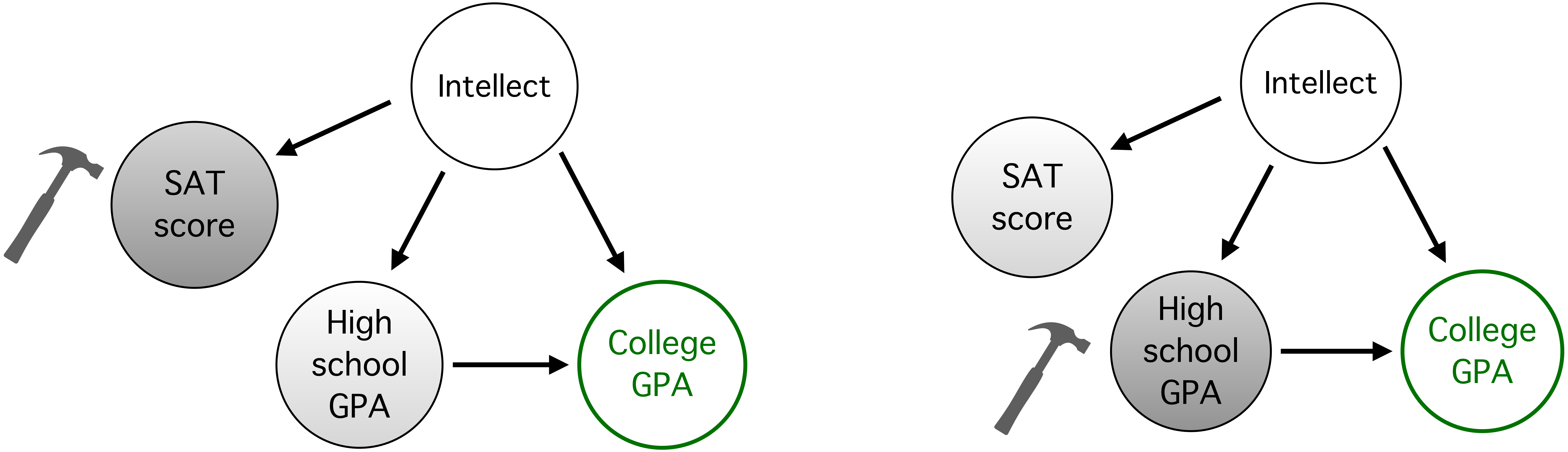
We study the problem of agent selection in causal strategic learning under multiple decision

Transparency Invites Strategic Behaviour



Gaming vs. Improvement

- Causal modelling is necessary to incentivise improvement (Miller et al., 2020).



College Admissions (Harris et al., 2022).

Prior Work vs. Ours

Prior work

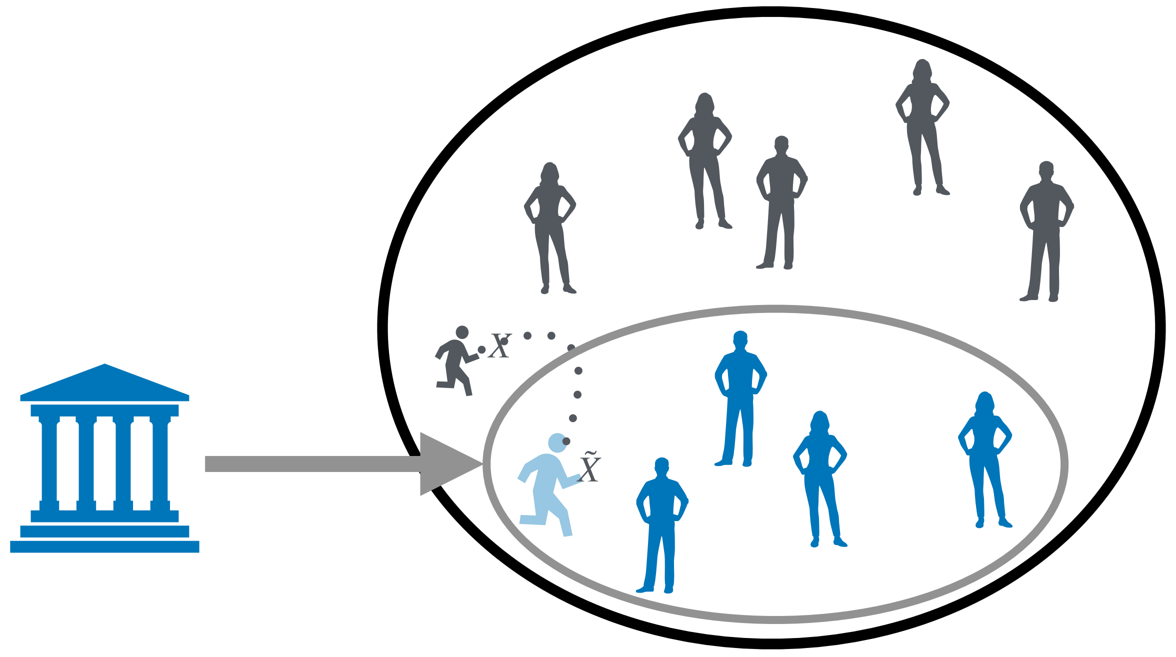
- Agent evaluation
- Only a single decision maker

Our settings

- Agent evaluation, then selection
- Multiple decision makers.

Problem Formulation

Single decision maker vs. multiple agents

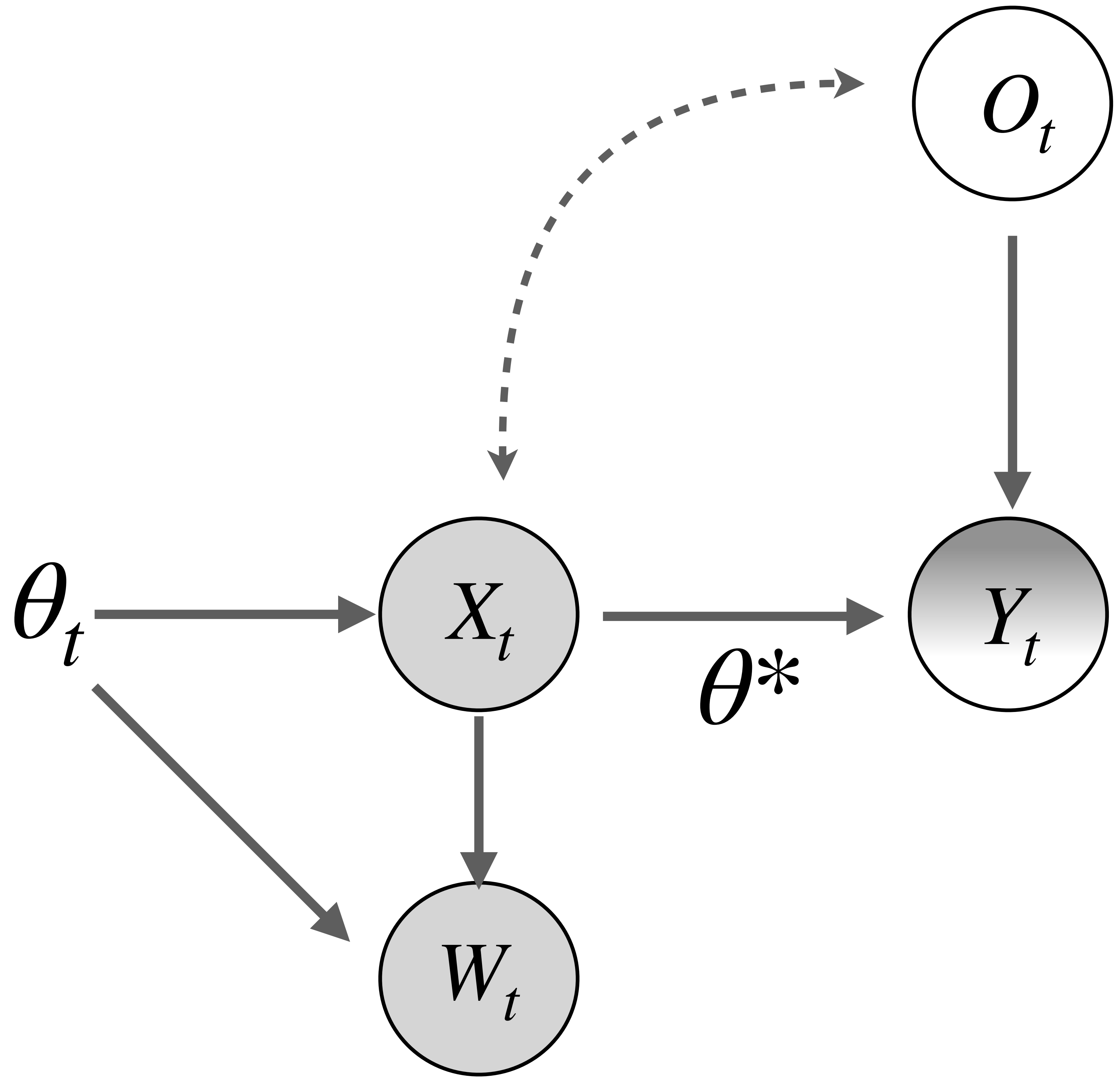


1. **Agents' model:**
   - Initial covariates & noise: $B_t, O_t \sim P_{B,O}$,
   - Updated covariates: $X_t := B_t + M\theta_t$,
   - Outcome: $Y_t := X_t^\top \theta^* + O_t$.
2. **DM's selection rule:**

$$\delta_{\theta_t} : \mathbf{x} \mapsto p(W_t = 1 \,|\, X_t = \mathbf{x}; \theta_t).$$

3. **DM's objective:**

$$\max_{\theta_t} \mathcal{Q}(\theta_t) = \max_{\theta_t} \mathbb{E}[Y_t \,|\, W_t = 1; \theta_t] \, .$$



Implicit Tradeoff

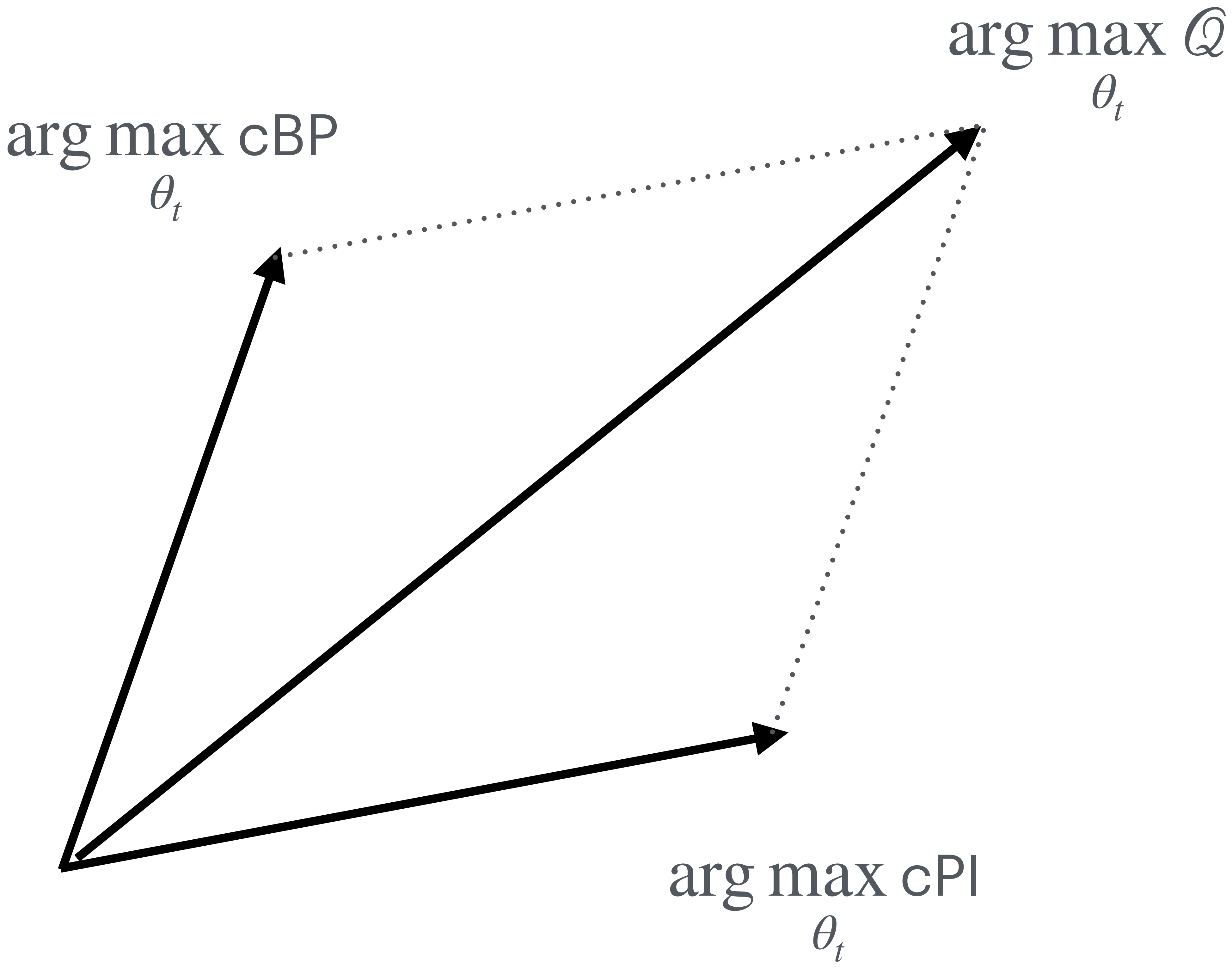
- Linearity assumptions help convey the idea that the optimal decision rule is a trade-off:

$$
\begin{aligned}
\mathcal{Q}(\theta_t) &= \mathbb{E}[Y_t \,|\, W_t = 1; \theta_t] \\
&= \mathbb{E}[(B_t + M\theta_t)^\top \theta^* + O_t \,|\, W_t = 1; \theta_t] \\
&= \mathbb{E}[B_t^\top \theta^* + O_t \,|\, W_t = 1; \theta_t] + \theta_t^\top M \theta^* \\
&= \text{cBP}(\theta_t).+.\text{cPI}(\theta_t)
\end{aligned}
$$

1. **Bounded Optimum**: If $\text{cBP}(\theta_t) = \alpha^\top \theta_t + \beta$, we have:

$$
\underset{\theta_t}{\arg\max}\ \mathcal{Q}(\theta_t) = \frac{\alpha + M\theta^*}{\|\alpha + M\theta^*\|_2} =: \theta^{AO}.
$$

3. **Maximum Improvement**: If $\alpha = (k-1)M\theta^*$ for some $k > 0$, then the maximisers of $\mathcal{Q}(\theta_t)$ and $\text{cPI}(\theta_t)$ coincide.



Implicit Tradeoff

- Linearity assumptions help convey the idea that the optimal decision rule is a trade-off:

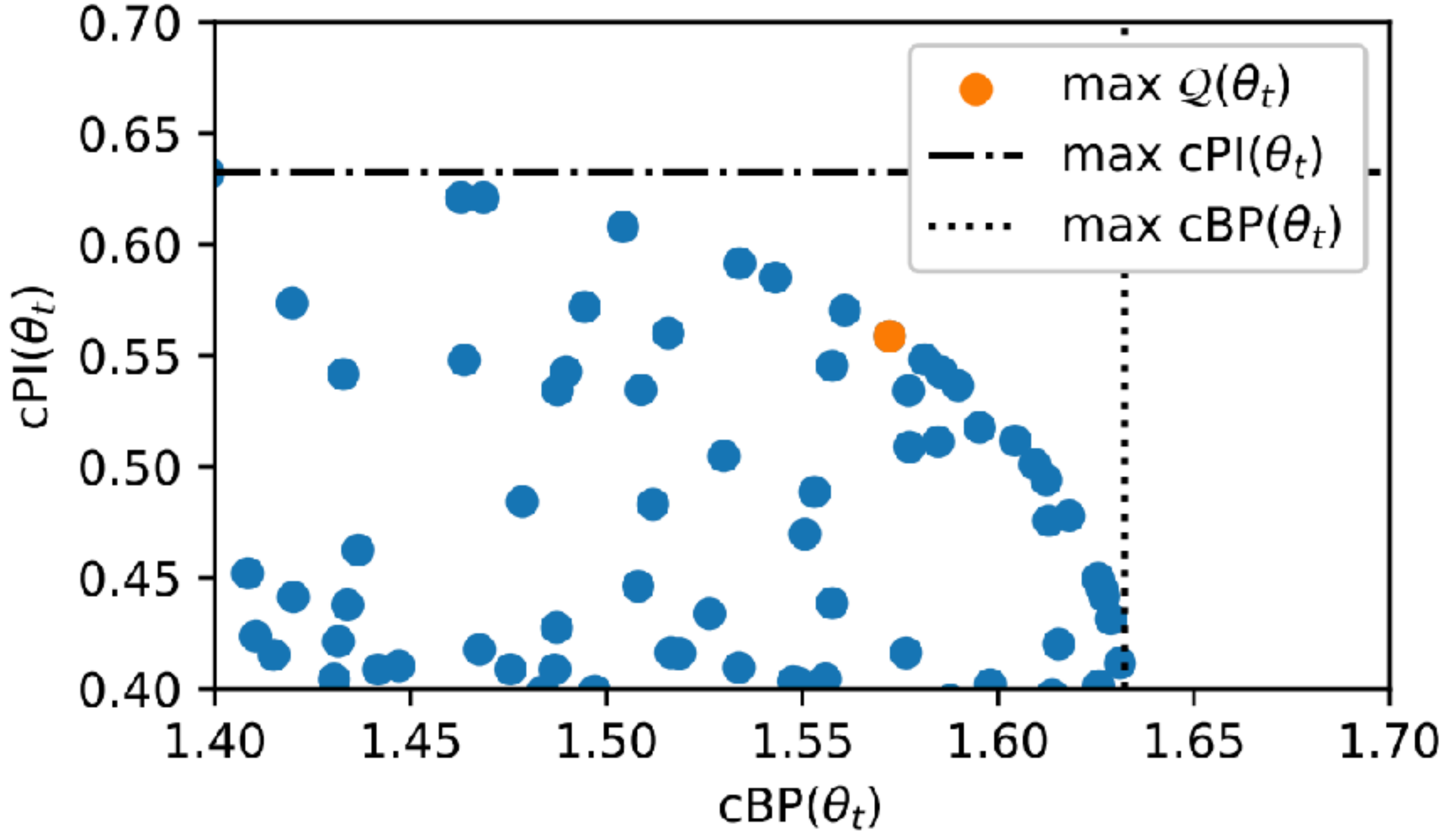
$$\mathcal{Q}(\theta_t) = \mathbb{E}[Y_t \mid W_t = 1; \theta_t]$$
$$= \mathbb{E}[(B_t + M\theta_t)^\top \theta^* + O_t \mid W_t = 1; \theta_t]$$
$$= \mathbb{E}[B_t^\top \theta^* + O_t \mid W_t = 1; \theta_t] + \theta_t^\top M \theta^*$$
$$= \text{cBP}(\theta_t) + \text{cPI}(\theta_t)$$

1. **Bounded Optimum**: If $\text{cBP}(\theta_t) = \alpha^\top \theta_t + \beta$, we have:

$$\arg\max_{\theta_t} \mathcal{Q}(\theta_t) = \frac{\alpha + M\theta^*}{\|\alpha + M\theta^*\|_2} =: \theta^{AO}.$$

3. **Maximum Improvement**: If $\alpha = (k-1)M\theta^*$ for some $k > 0$, then the maximisers of $\mathcal{Q}(\theta_t)$ and $\text{cPI}(\theta_t)$ coincide.

Illustration for certain δ

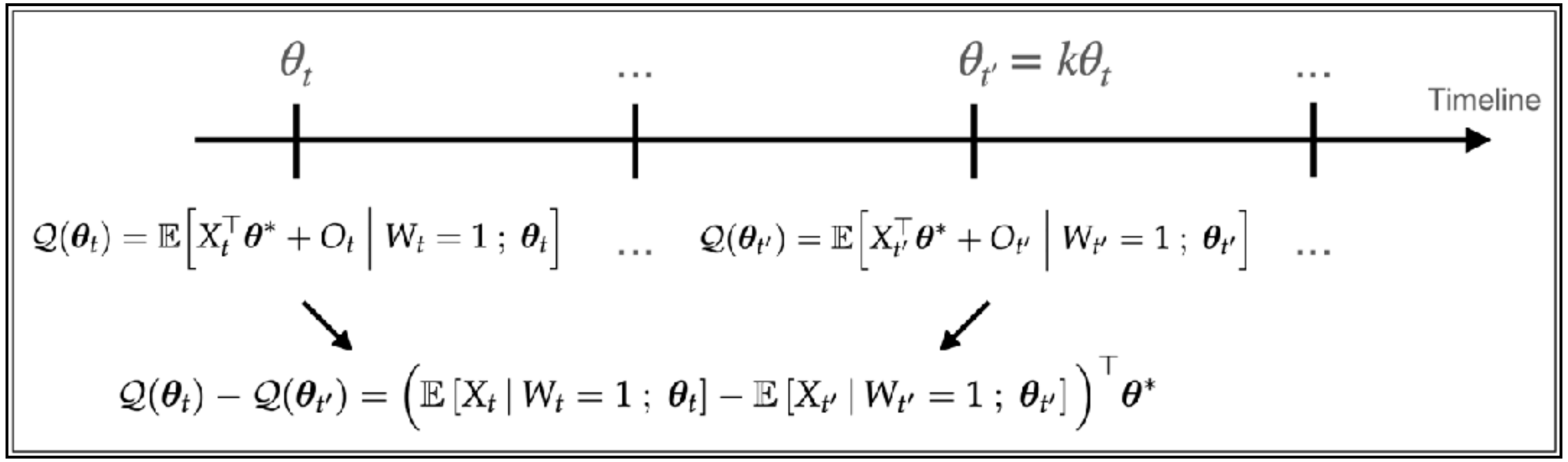


Identifying Causal Parameters

With ranking selection:

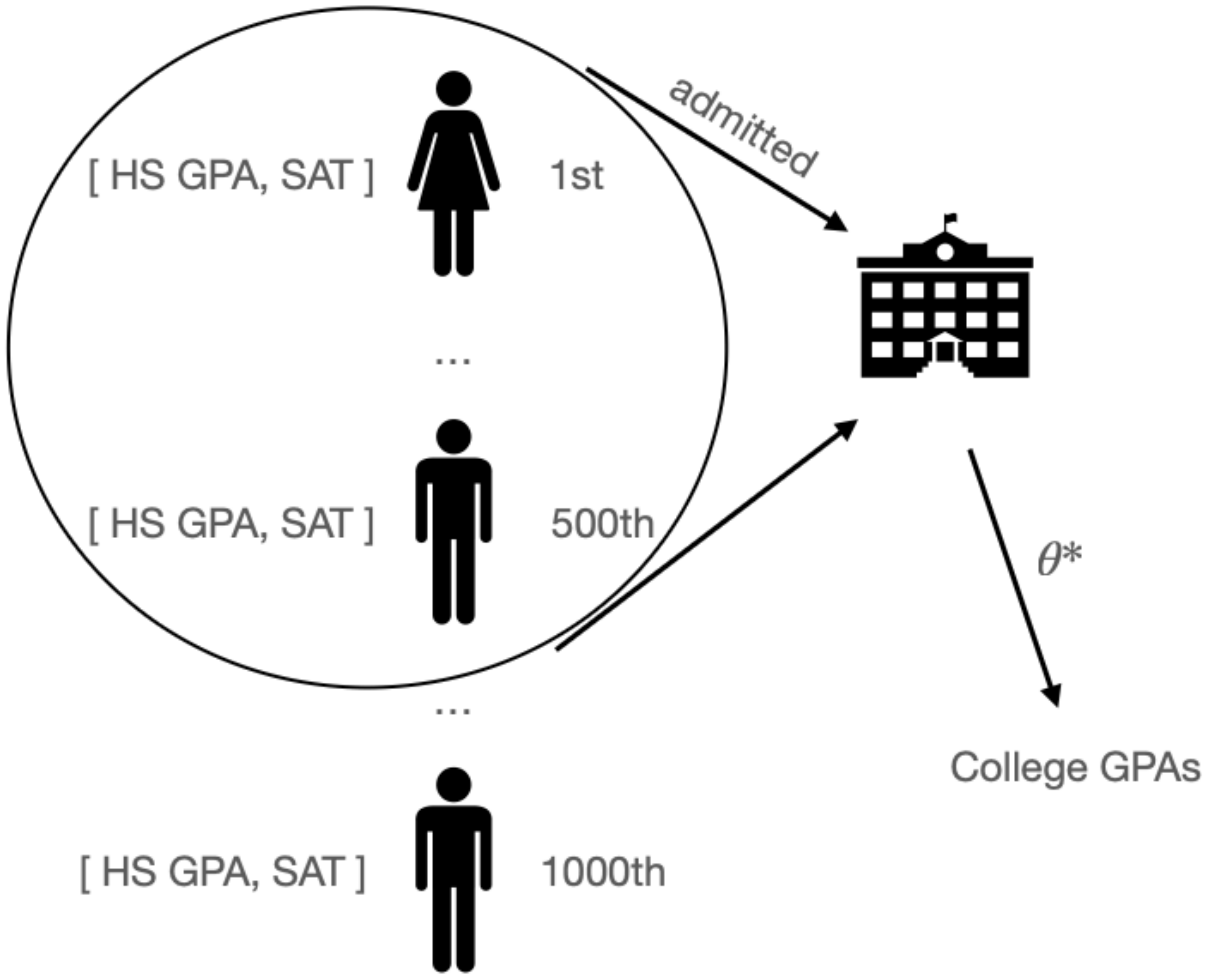
$$\delta_{\theta_t}(\mathbf{x}) = p\left(X_t^\top \theta_t \le \mathbf{x}^\top \theta_t\right) = \mathrm{CDF}_{X_t^\top \theta_t}\left(\mathbf{x}^\top \theta_t\right),$$

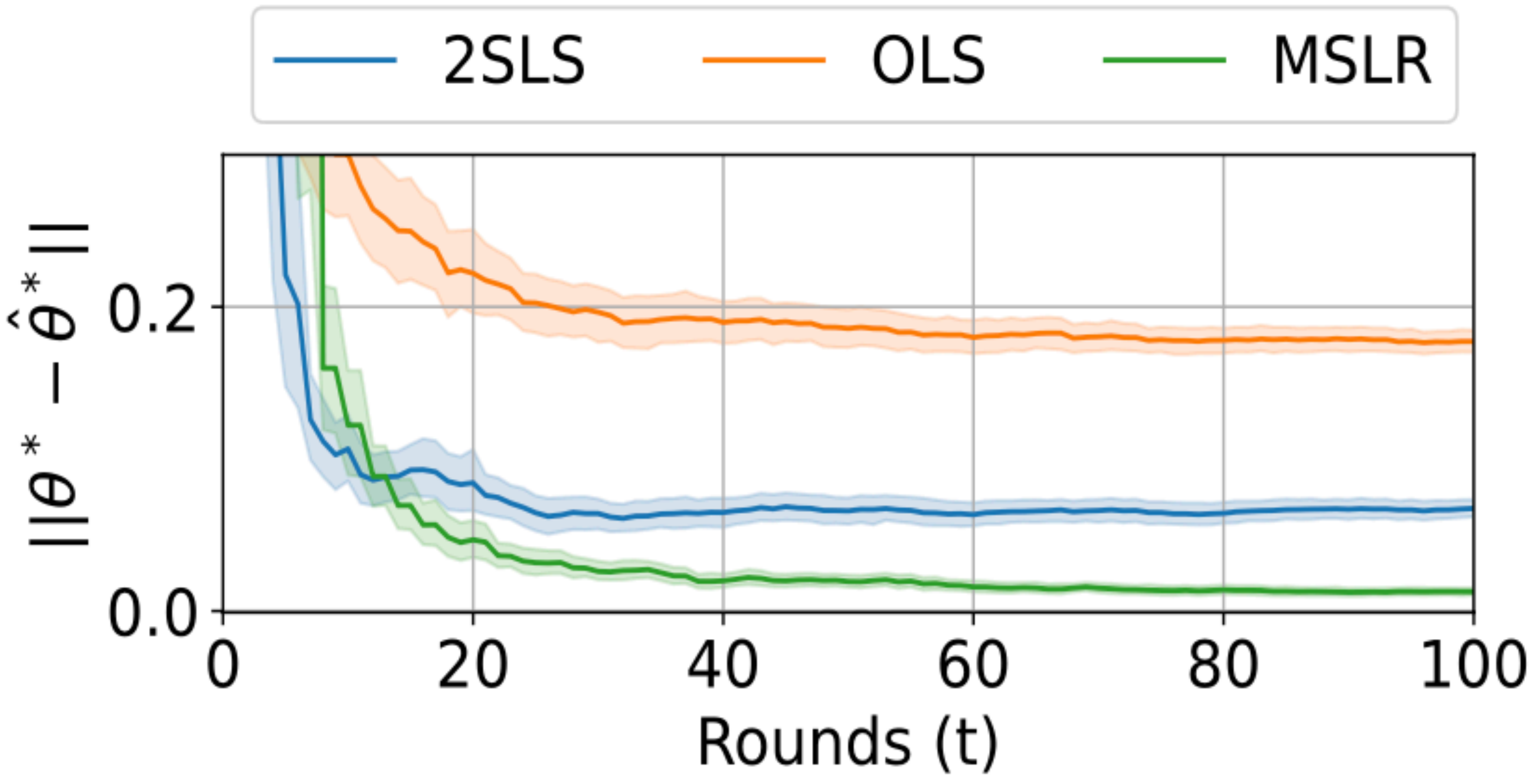
and when:



- **Our algorithm**: Mean-shift Linear Regression (MSLR)

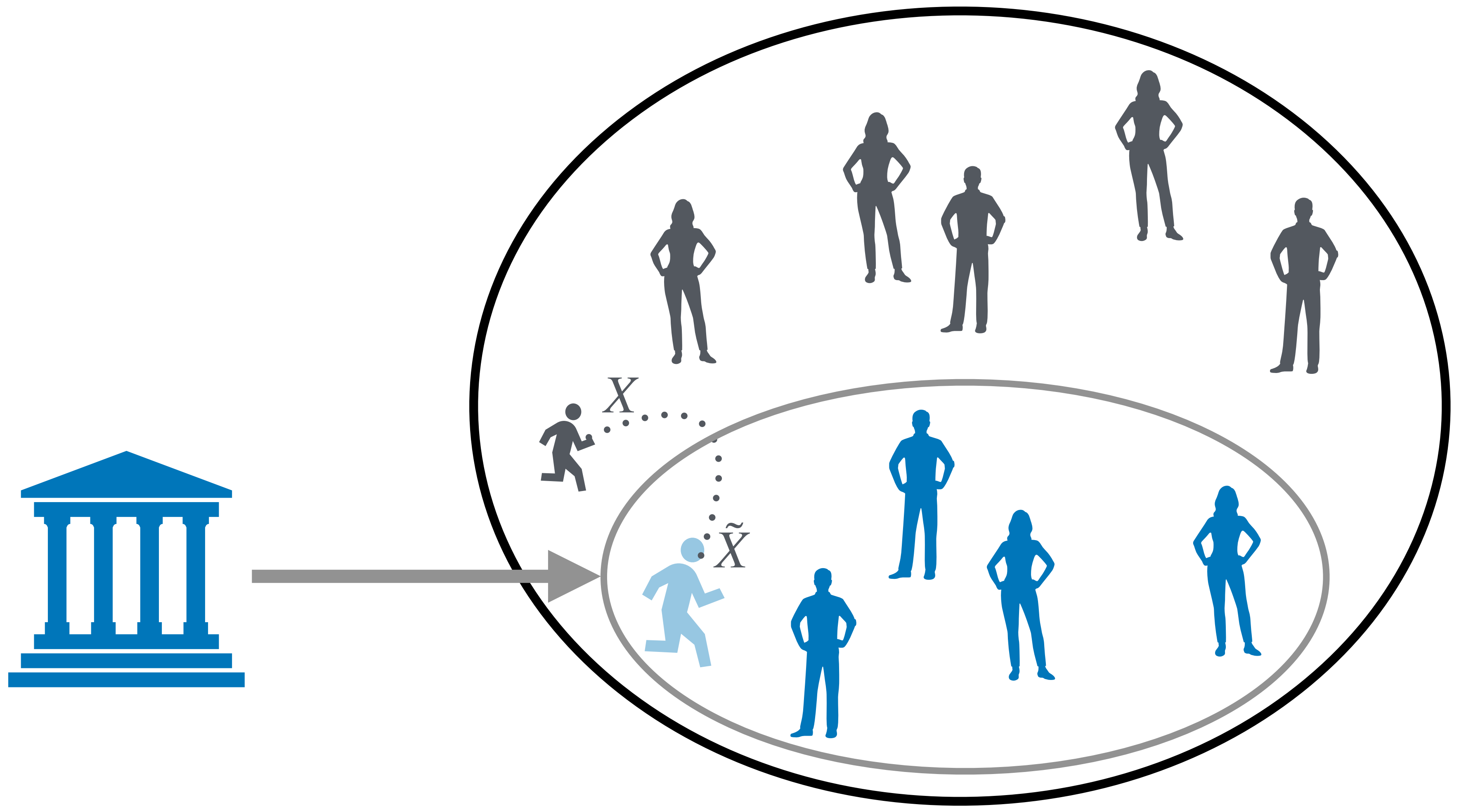
Identifying Causal Parameters





Takeaway

Single decision maker vs. multiple agents

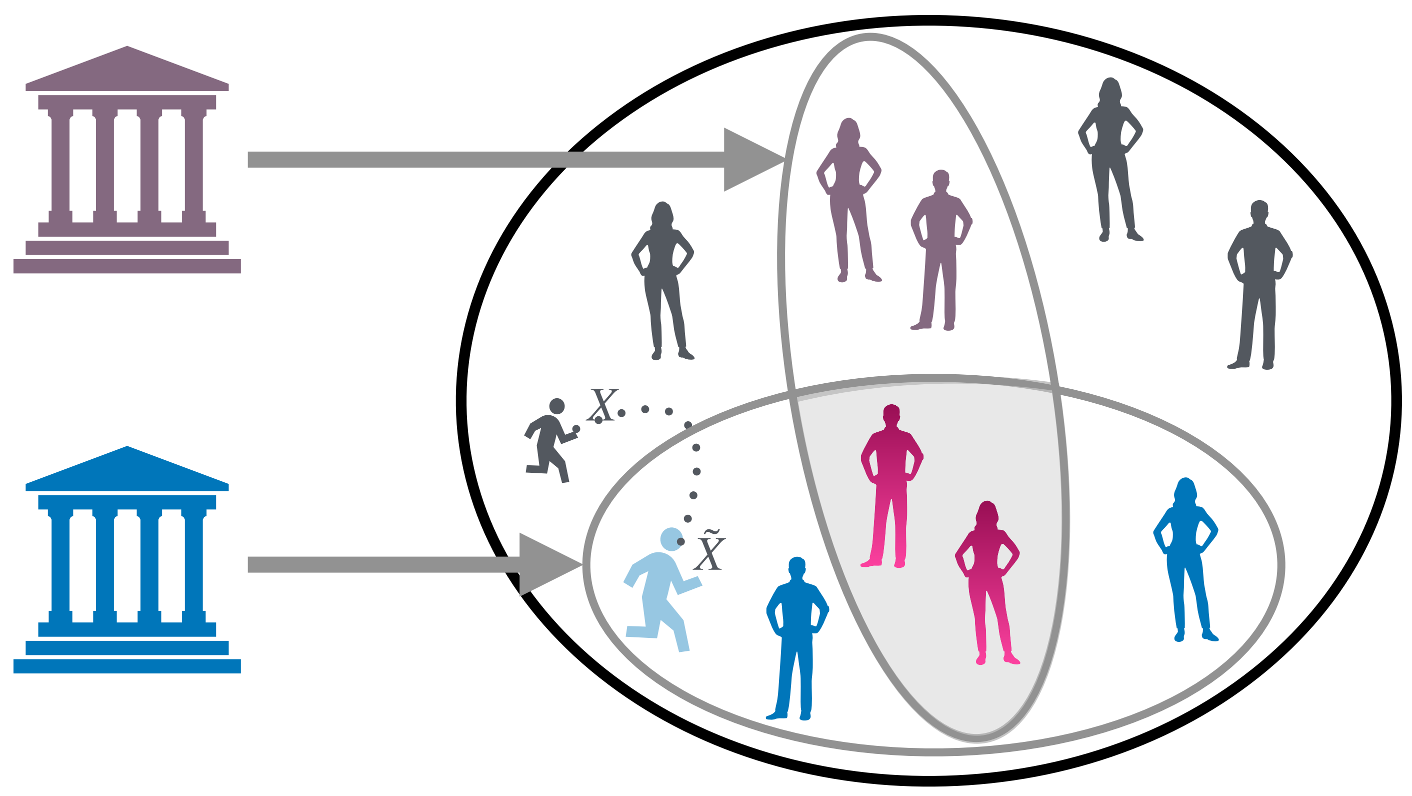


- With ***agent selection***:
  - There is a trade-off between choosing the best candidates & incentivising them.
  - There is selection bias in observational data.

- Causal modelling is important for ***incentivisation***, but we also need a **regulator**
  - to align the two objectives: **choosing vs. incentivising**.

Problem Formulation

Multiple decision makers vs. multiple agents

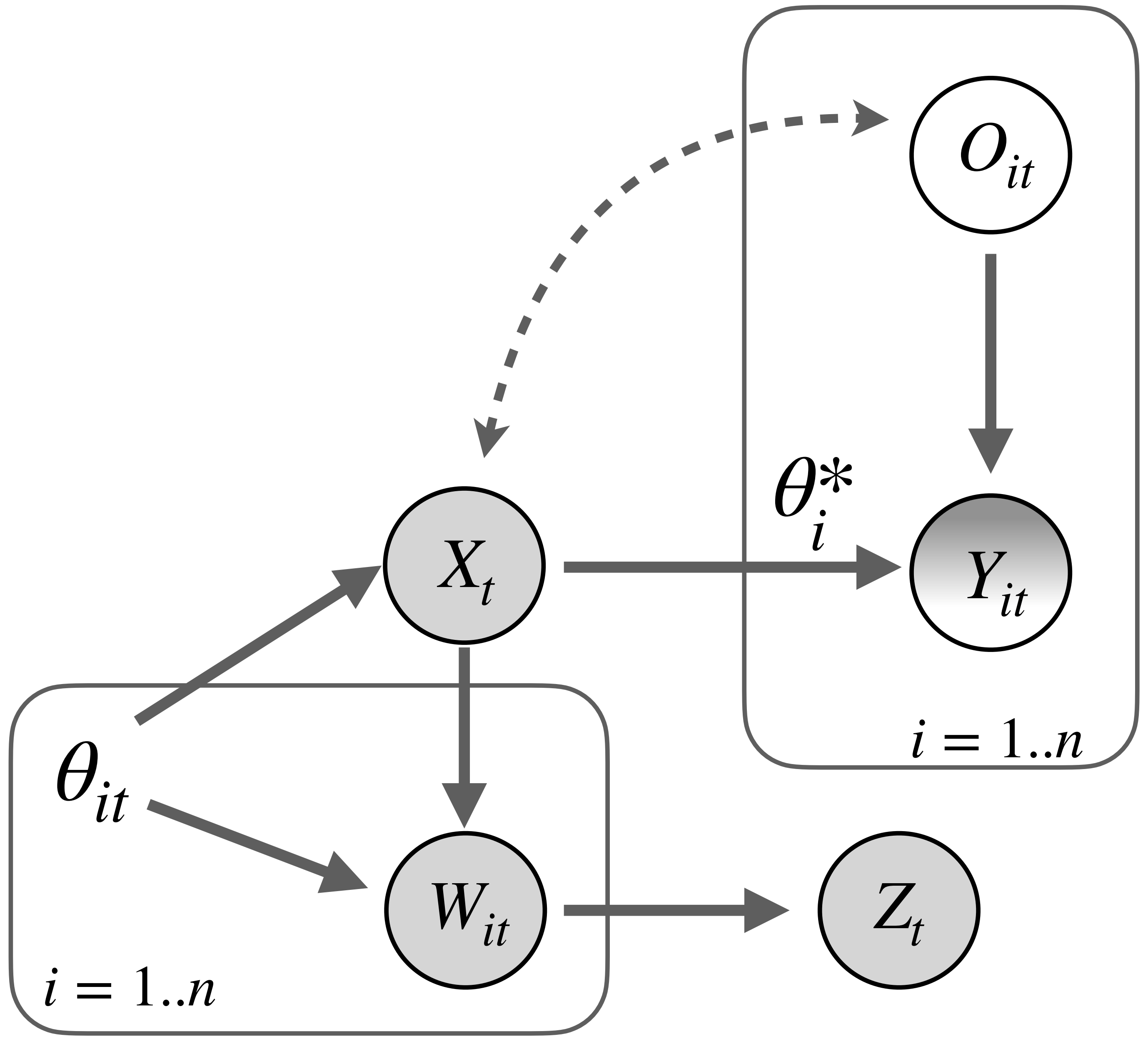


- **Agents' model:**
  - Initial covariates & noise: $B_t, O_t \sim P_{B,O}$,
  - Updated covariates: $X_t := B_t + M\left(\sum_{i=1}^{n} \gamma_i \theta_{it}\right)$,
  - Outcome: $Y_{it} := X_t^\top \theta_i^* + O_{it}$.
- **A DM's selection rule:**

$$\delta_{\theta_{it}} : \mathbf{x} \mapsto p(W_{it} = 1 \,|\, X_t = \mathbf{x}; \theta_{it}).$$

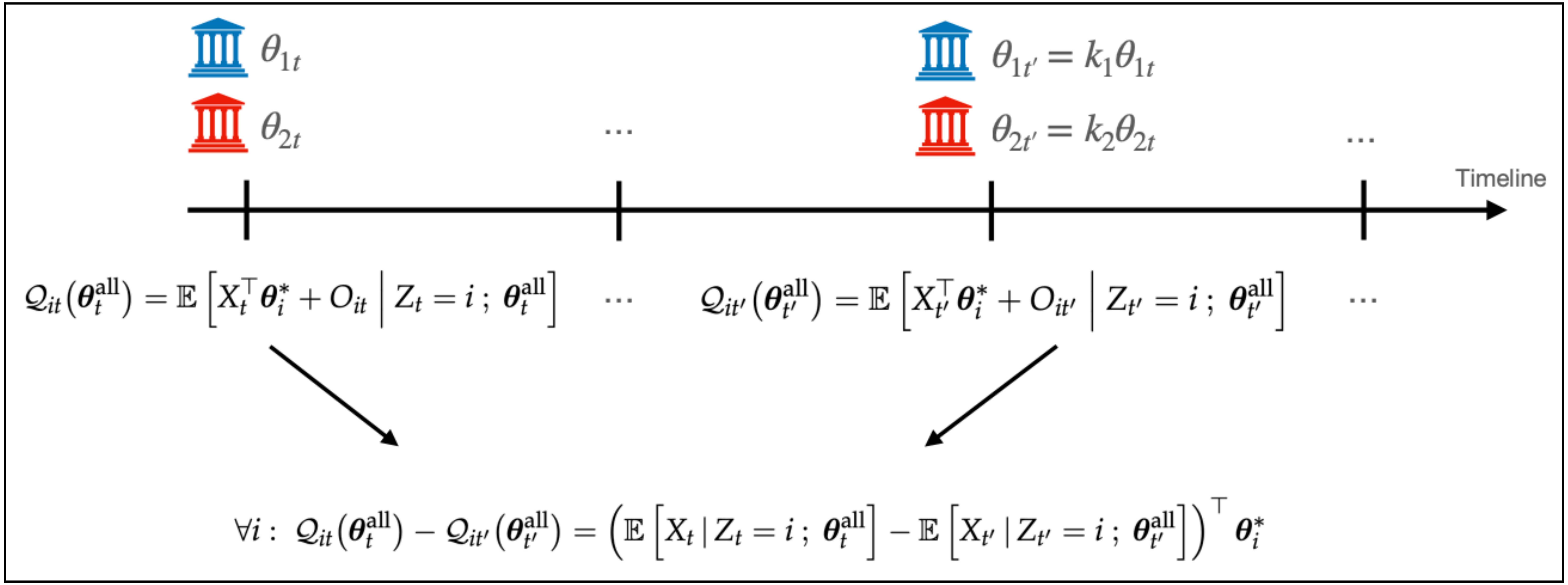
- **A DM's objective:**

$$\max_{\theta_{it}} \mathcal{Q}_i\left(\left\{\theta_{it}, \theta_t^{-i}\right\}\right) = \max_{\theta_{it}} \mathbb{E}[Y_{it} \,|\, Z_t = i; \theta_t^{\text{all}}] \,.$$



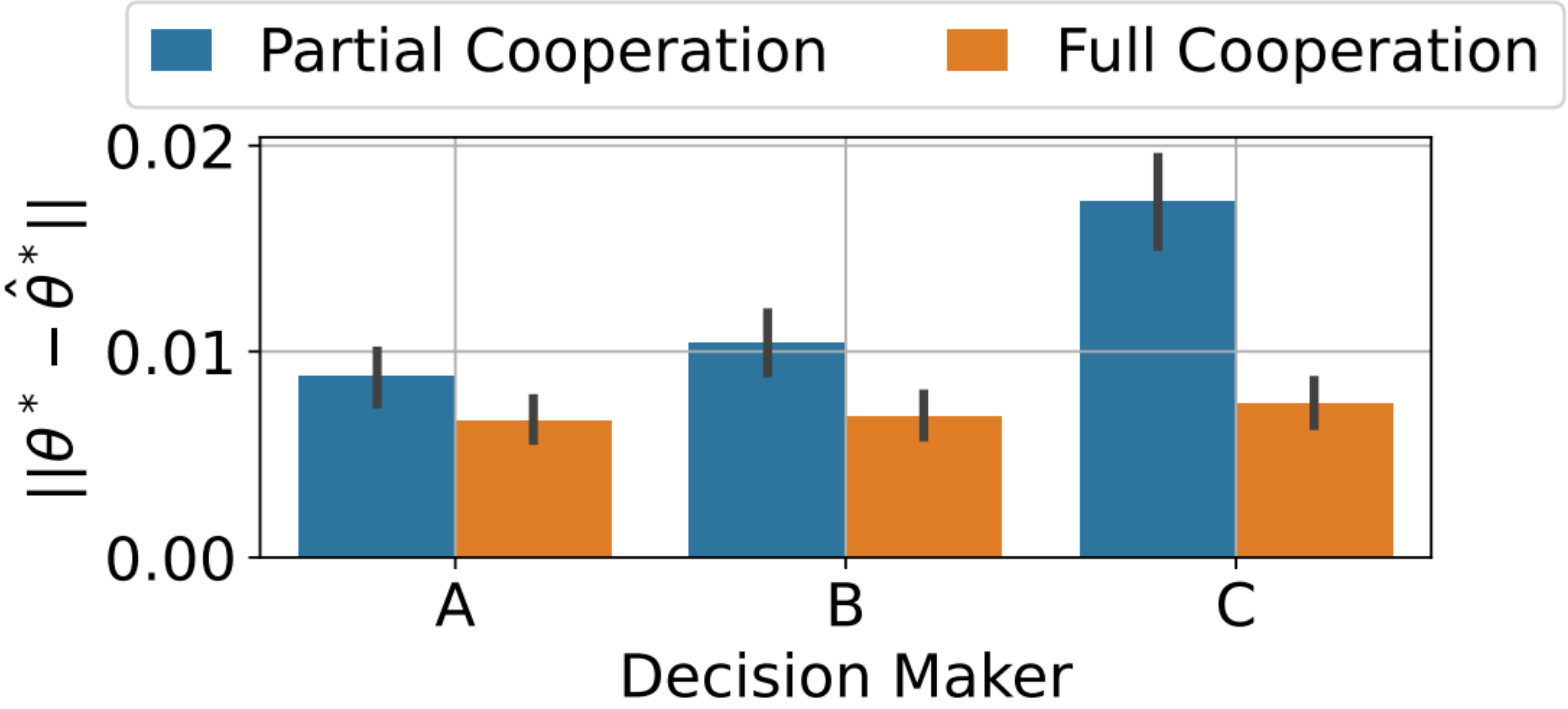
Cooperative Protocol

With ranking selection: $\delta_{\theta_{it}}(\mathbf{x}) = p\left(X_t^\top \theta_{it} \leq \mathbf{x}^\top \theta_{it}\right) = \mathrm{CDF}_{X_t^\top \theta_{it}}\left(\mathbf{x}^\top \theta_{it}\right)$, and when it holds that, for all decision makers i, $\exists \{t, t'\} : \theta_{it} = k_i \theta_{it'}$, then



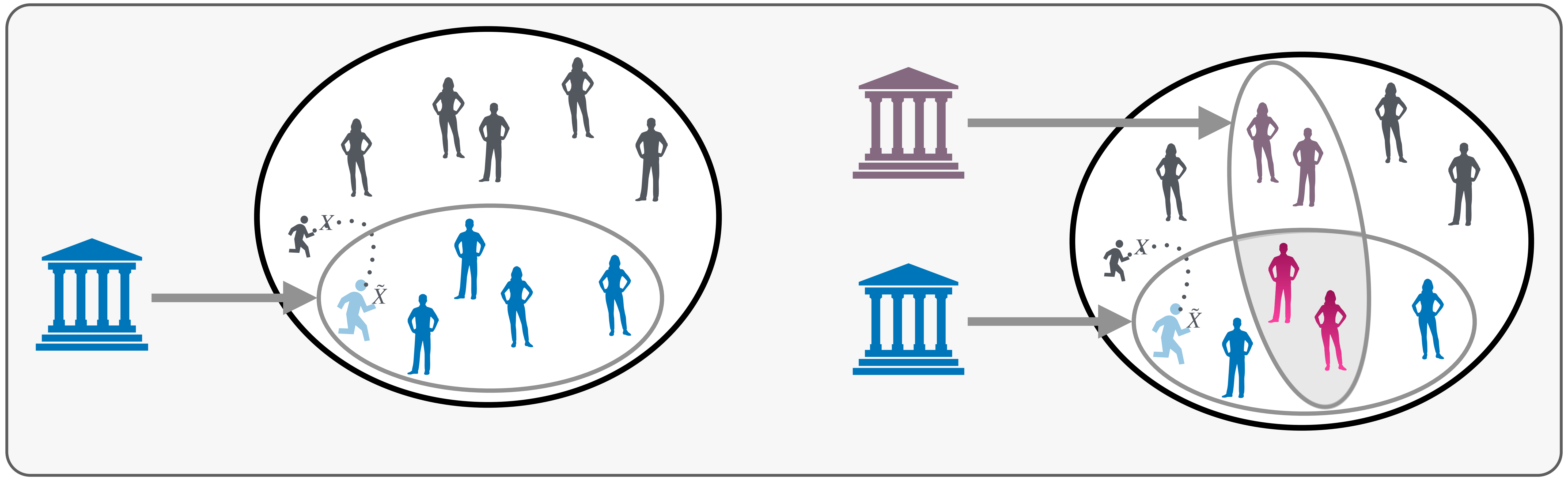
Identifying Causal Parameters

Three decision makers in two scenarios: {{A,B} & {C}} vs. {A,B,C}



Conclusion

When explanation = full-model disclosure



1. With agent selection:

- There is a trade-off between choosing the best candidates & incentivising them.
- There is selection bias in observational data.

2. With competitive selection

- Each agent's improvement is split in different directions.
- No DM can correct for selection bias alone.

3. Causal modelling is important for incentivisation, but we need a regulator

- to align the two objectives: choosing vs. incentivising,
- to coordinate DMs for (a) safeguarding agents' efforts and (b) parameter estimation.

Explanation Design in Strategic Learning: Sufficient Explanations that Induce Non-harmful Responses

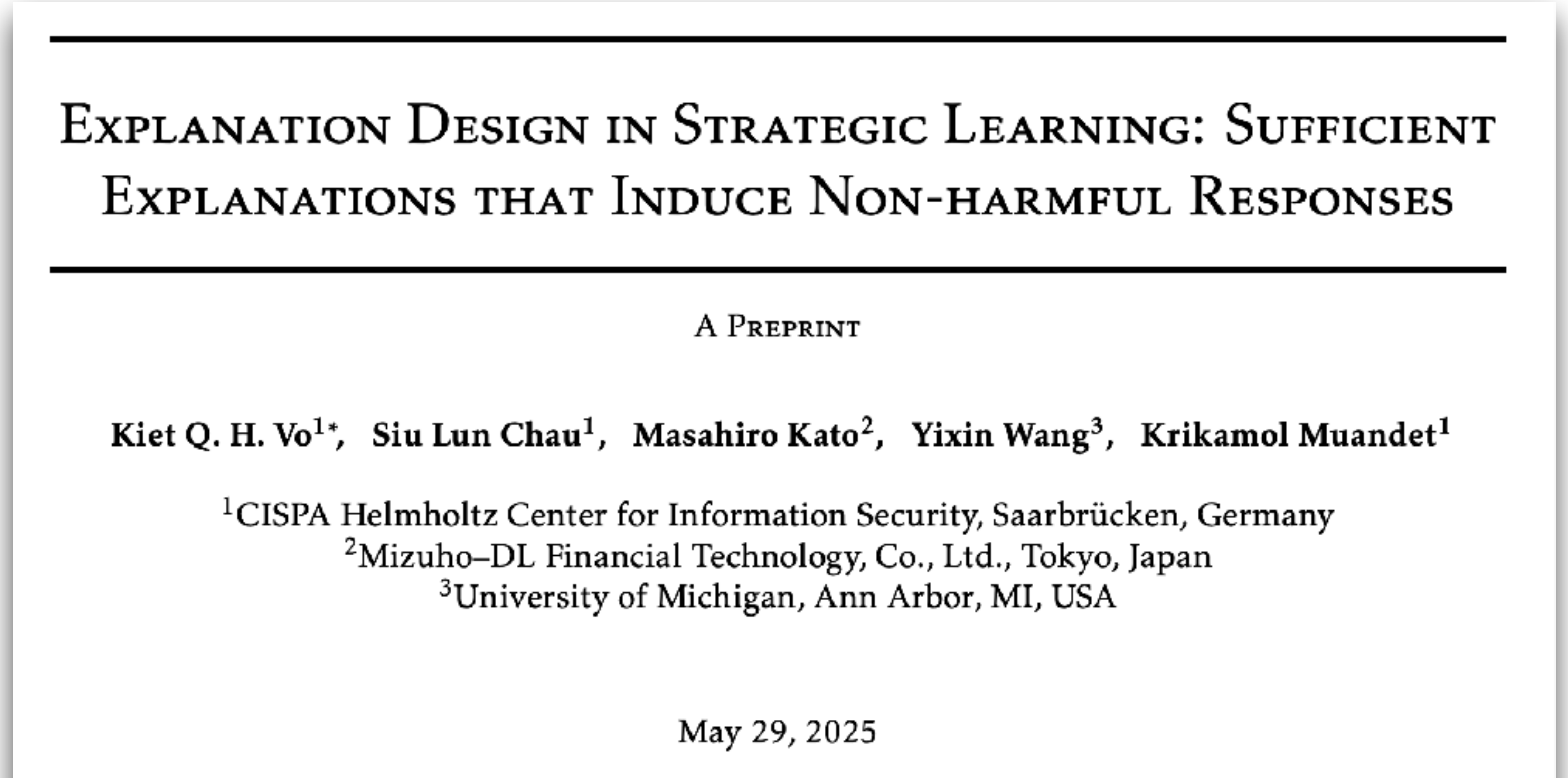
Kiet Vo
CISPA

Siu Lun Chau
CISPA

Masahiro Kato
Mizuho-DL

Yixin Wang
Michigan

Krikamol Muandet
CISPA



EXPLANATION DESIGN IN STRATEGIC LEARNING: SUFFICIENT EXPLANATIONS THAT INDUCE NON-HARMFUL RESPONSES

A PREPRINT

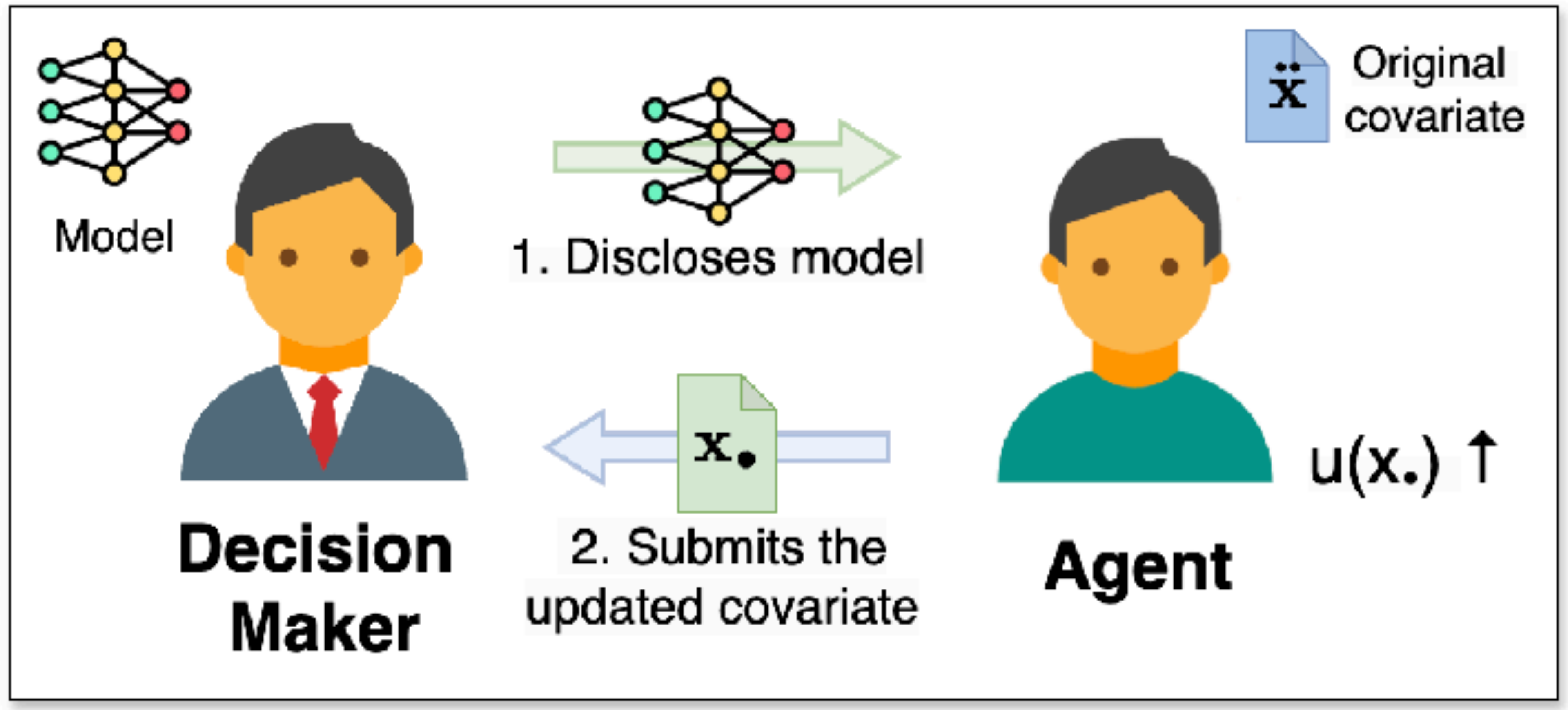
Kiet Q. H. Vo[1*], **Siu Lun Chau**[1], **Masahiro Kato**[2], **Yixin Wang**[3], **Krikamol Muandet**[1]

[1]CISPA Helmholtz Center for Information Security, Saarbrücken, Germany
[2]Mizuho–DL Financial Technology, Co., Ltd., Tokyo, Japan
[3]University of Michigan, Ann Arbor, MI, USA

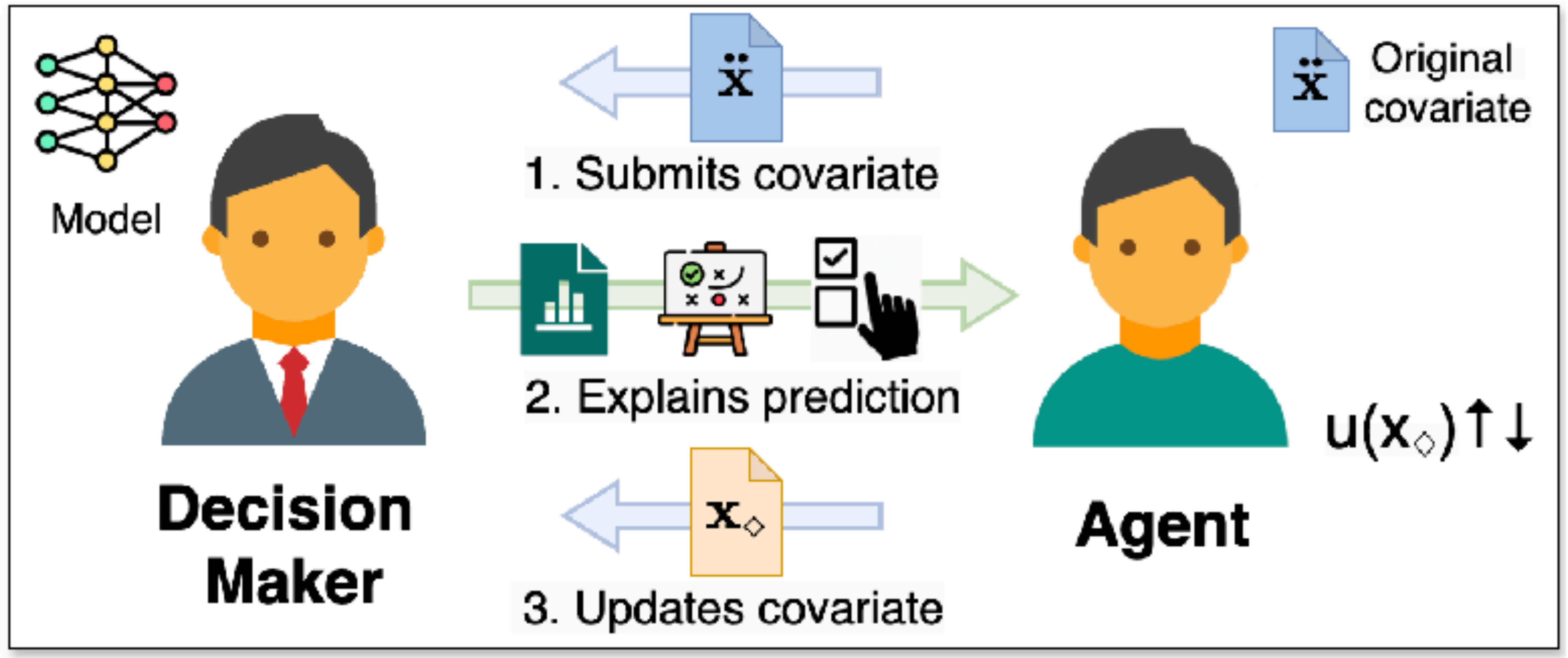
May 29, 2025

Partial Disclosure in Strategic Learning

Full Disclosure of Predictive Model



Partial Disclosure via Explanations (Ours)



Here, the agent can **correctly anticipate** how changing $\ddot{x}$ affects the prediction, then picks an update $x_\bullet$ that **improves** utility $u(x_\bullet)$.

With only an explanation (i.e., partial information), the agent's update $x_\diamond$ **might not improve** utility $u(x_\diamond)$.

Q1: Can we ensure **no reduction** in agents' utilities (do no harm)?
Q2: Is there a **sufficient** class of explanations that guarantee this?

Setup

- We assume that an **agent** is realised with $(\ddot{x}_t, c_t) \sim P_{\ddot{X},C}$, with a **cost function** $c_t : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$:

    1. **DM** predicts agent's risk with $g(\ddot{x}_t)$ where g is the DM's private information
    2. **DM** provides explanation $e_t := \sigma(g, \ddot{x}_t)$ to the **agent**.
    3. **Agent** modifies covariate to $x_t := \psi(\ddot{x}_t, e_t, c_t)$.
    4. **DM** updates the prediction from $g(\ddot{x}_t)$ to $g(x_t)$.

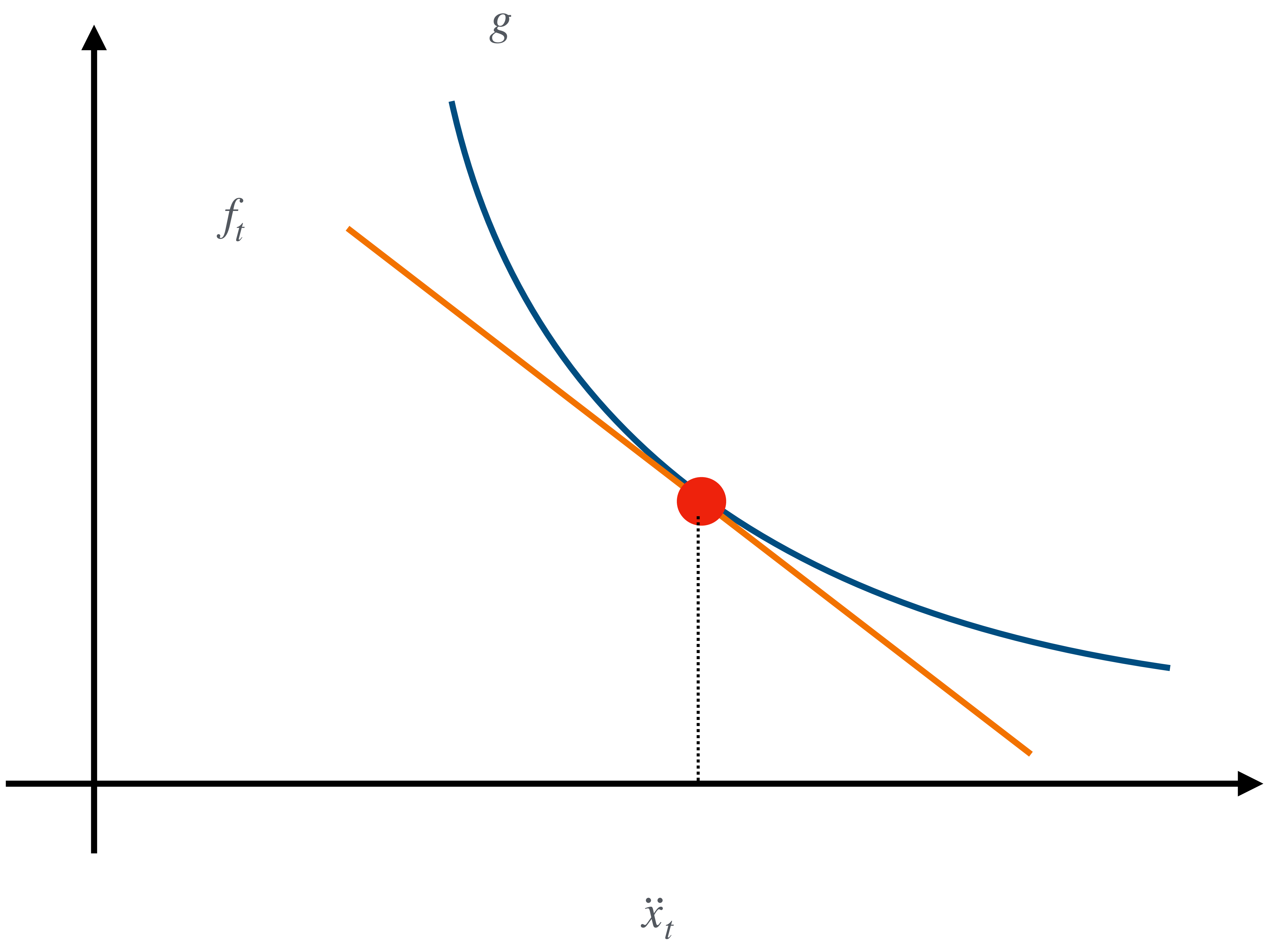
- For the **agent** t:
    - Their **true utility**: $u_t(g, x) := b(x) - c_t(\ddot{x}_t, x) = -g(x) - c_t(\ddot{x}_t, x)$.
    - Their non-harmful responses: $\nu_t := \{x \in \mathcal{X} : u_t(g, x) \geq u_t(g, \ddot{x}_t)\}$.

Surrogate Models as Explanations

- The explanation is a **surrogate model** $f_t : \mathcal{X} \to \mathcal{Y}$ to the predictive model $g : \mathcal{X} \to \mathcal{Y}$ (e.g., LIME and Taylor expansions)
- The **agent** is assumed to change from $\ddot{x}_t$ to x_t that maximises the **surrogate utility function**

$$u_t(f_t, x) := (-f_t(x)) - c_t(\ddot{x}_t, x).$$

- When f_t exaggerates gain at some region x', i.e., $f_t(\ddot{x}_t) - f_t(x') > g(\ddot{x}_t) - g(x')$, then x_t might be outside ν_t.

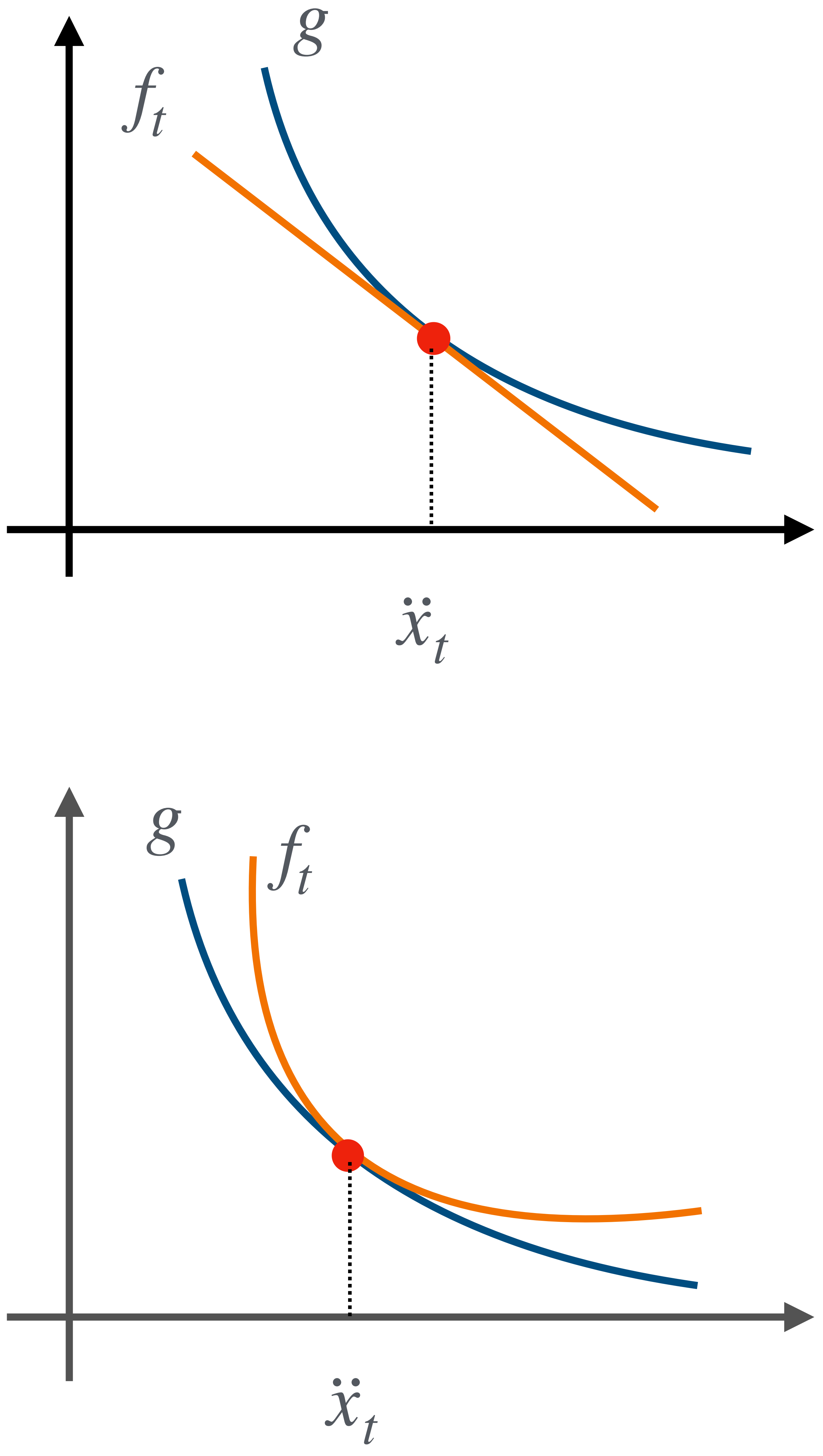


A Necessary Condition

- Consider a general form of cost function $c_t : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$, where $c_t(x, x) = 0$ and $c_t(x, x') > 0$ for all $x, x' \in \mathcal{X}$ and $x \neq x'$.
- If, for any cost function c_t, the induced response x_t is in ν_t, then

$$f_t(\ddot{x}_t) - f_t(x') \leq g(\ddot{x}_t) - g(x')$$

- This result extends to settings where agents construct surrogate models from partial information, e.g., Bayesian agents.



ARexes

Action recommendation-based explanations

- The explanation $(\vec{x}_t, \hat{\vec{y}}_t)$ contains
  - a recommended covariate update $\vec{x}_t$, and
  - the associated prediction score $\hat{\vec{y}}_t := g(\vec{x}_t)$.

- The agent is assumed to choose either $\ddot{x}_t$ or $\vec{x}_t$ as follows:

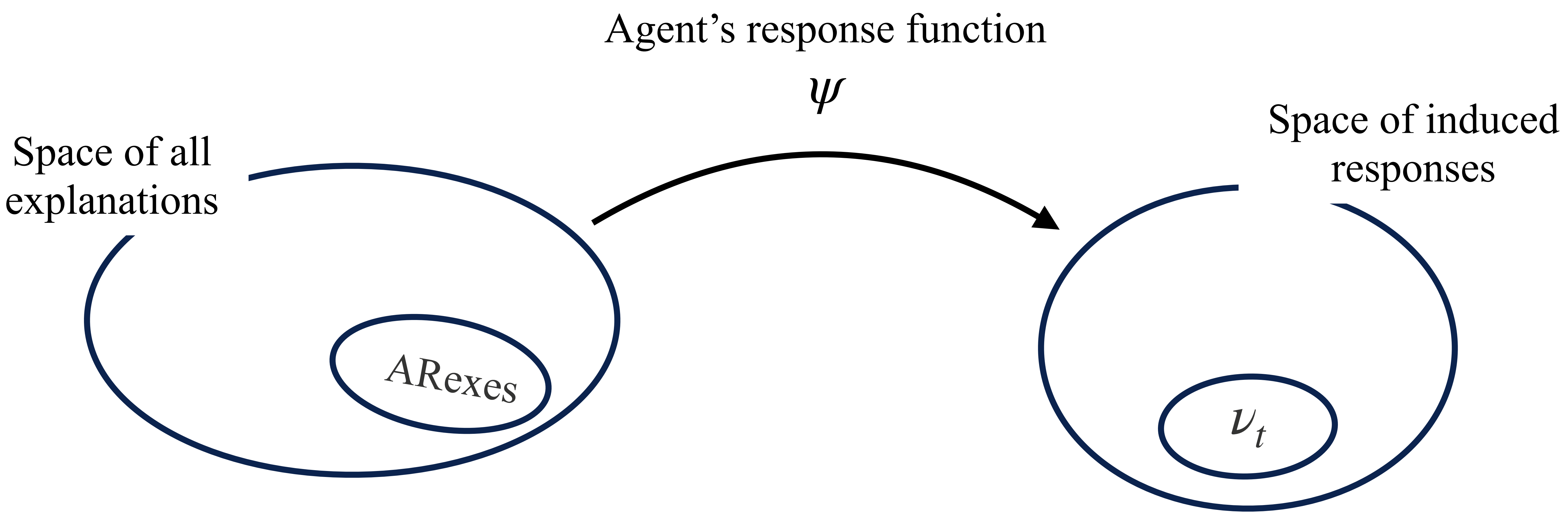
$$x_t := \arg \max_{x \in \{\ddot{x}_t, \vec{x}_t\}} \{-g(x) - c_t(\ddot{x}_t, x)\}$$

Distinction Between ARexes and Surrogates

- **ARexes** are transparent about the potential gains, i.e., via disclosing $\hat{\vec{y}}_t$.
  - ARexes guarantee non-harmful responses by construction.

- **ARexes** do not disclose information about other options x' where $x' \notin \{\ddot{x}_t, \vec{x}_t\}$.
  - Each ARex limits the set of agent's feasible responses to $\{\ddot{x}_t, \vec{x}_t\}$.
  - ARexes restrict the DM's uncertainty about how an agent responds.
  - ARexes are sufficient to induce all non-harmful responses.

Sufficiency of ARex-generating Methods



- If the DM has enough knowledge to partition a population of T agents into subgroups of homogeneous response behaviour, then

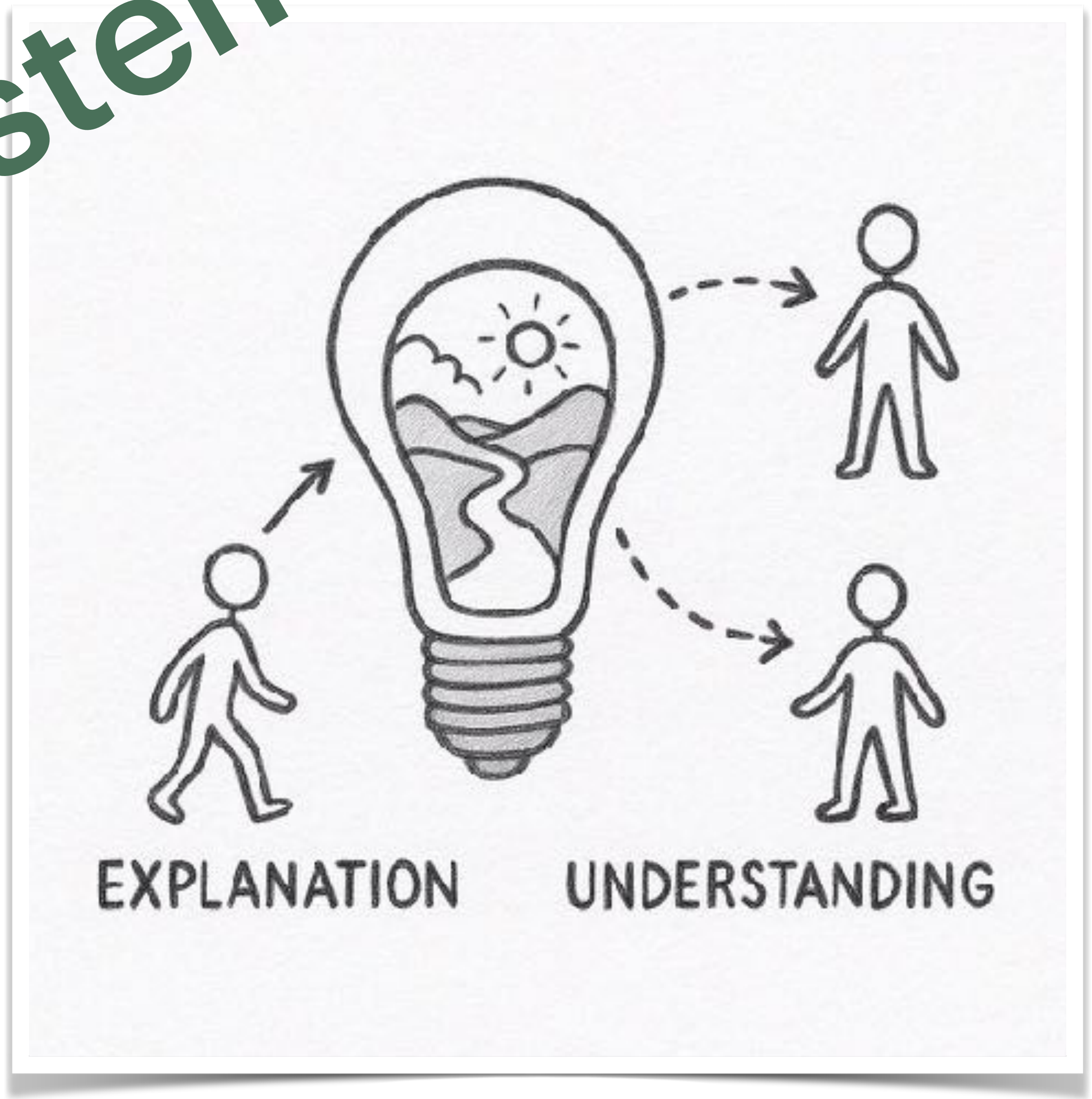
  For any arbitrary explanation method that induces non-harmful responses $(x_t^\bullet)_{t\in[T]}$, there exists an ARex generating function whose induced responses $(x_t^\diamond)_{t\in[T]}$ satisfy $(x_t^\diamond)_{t\in[T]} = (x_t^\bullet)_{t\in[T]}$.

Conclusion

- When explanations contain partial information, agents can overestimate their potential gains and unintentionally take utility-harming actions.
- We clarify the distinction between explanation methods, through the lens of strategic learning:
    - Surrogate-based methods must at least satisfy the **necessary condition** to guarantee the induced actions are non-harmful.
    - ARexes fix this issue by being transparent about the gains and withholding ambiguous information.
- ARex-generating methods are **sufficient**, when DM can ensure conditional homogeneity of agents' responses.

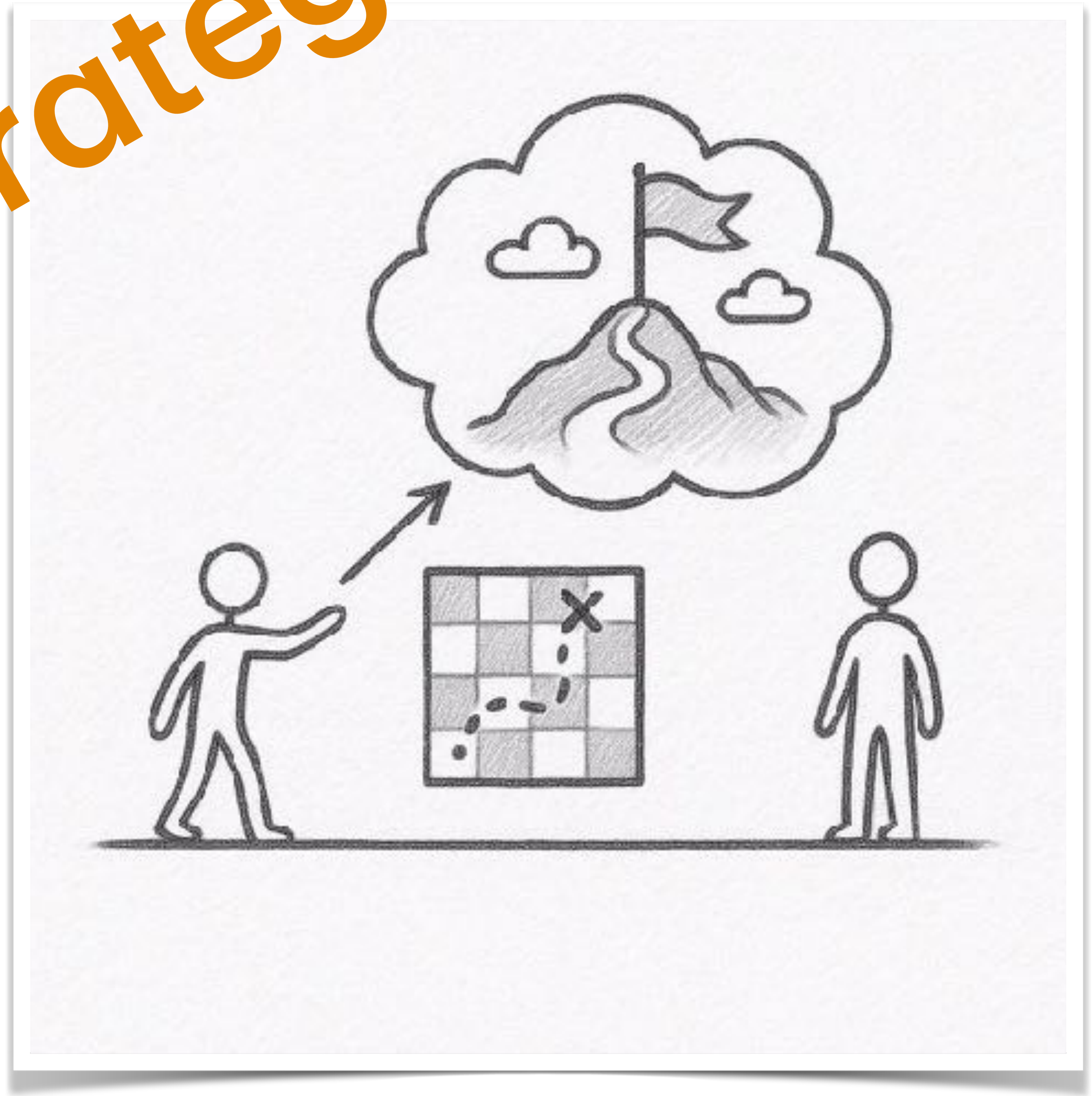
What Makes Great Explanation?

Epistemic



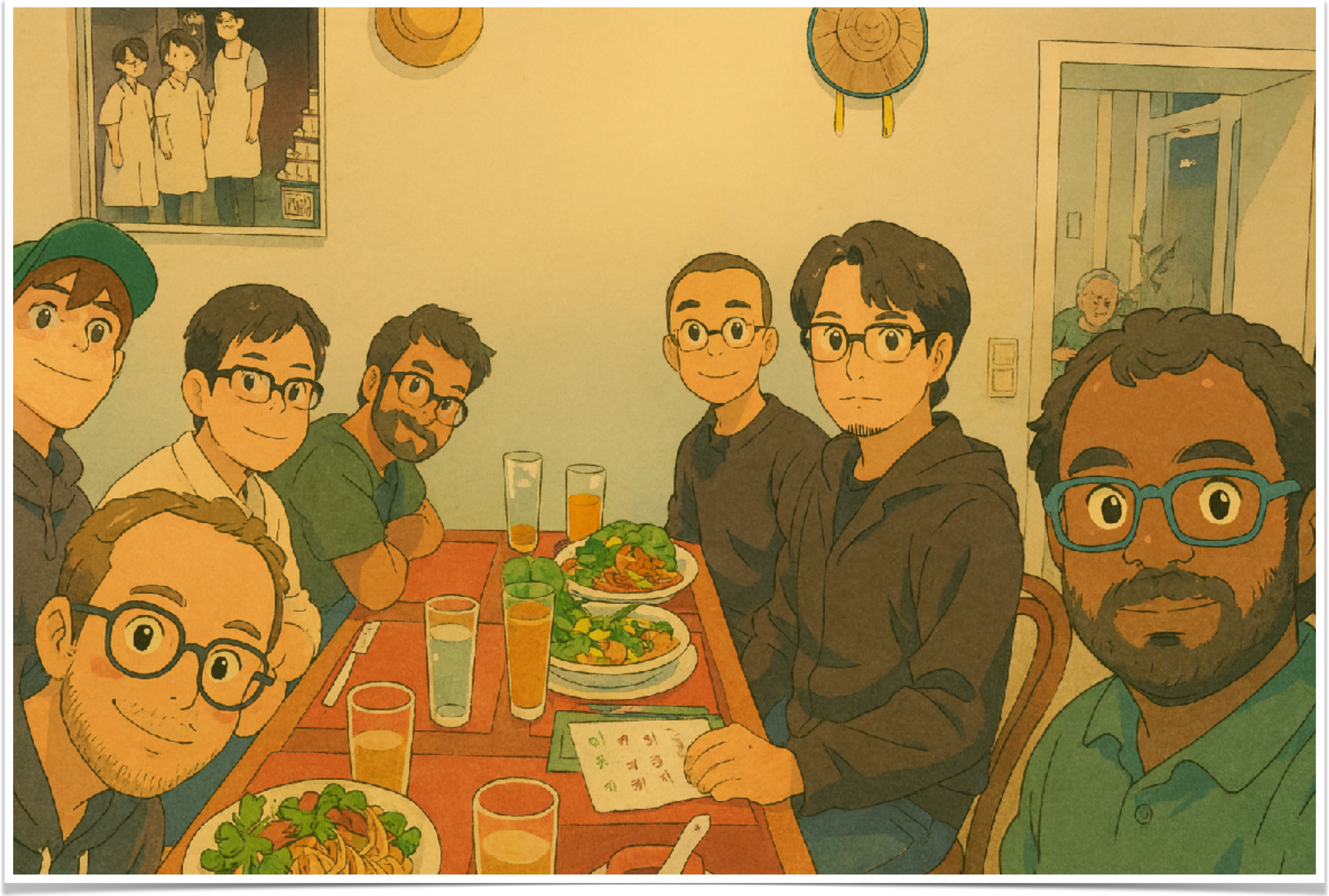
Explanation as understanding:
It turns information into understanding and knowledge.

Strategic



Explanation as influence:
It informs, convinces, and guides others toward desirable actions.

Thank you



Rational Intelligence Lab
https://ri-lab.org/

muandet@cispa.de