An Information-Flow Perspective on Explainability Requirements:

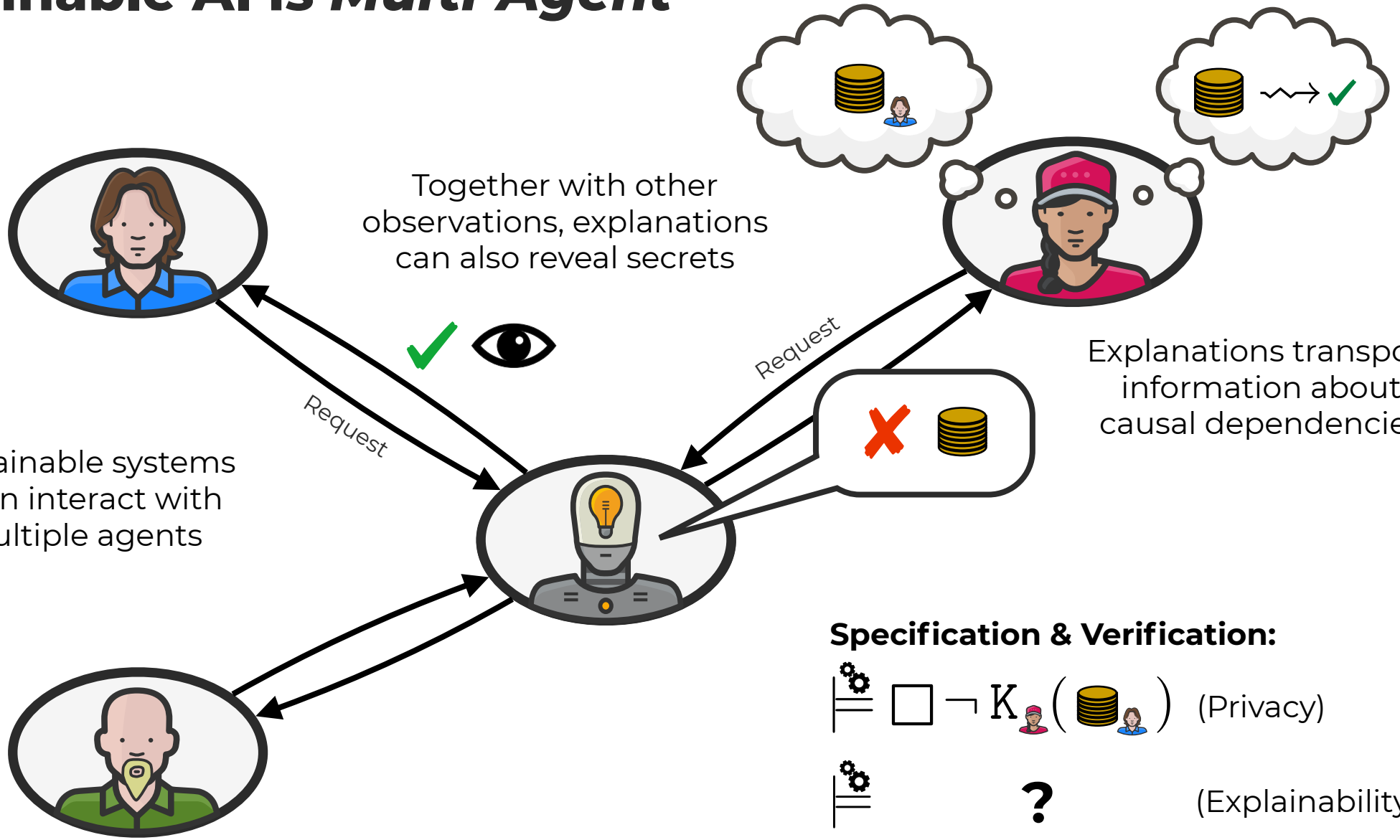
Specification and Verification

Bernd Finkbeiner[1], Hadar Frenkel[2], and Julian Siber[1]

[1] *CISPA Helmholtz Center for Information Security*
[2] *Bar-Ilan University*

KR 2025 — 16 November 2025

Explainable AI is *Multi-Agent*



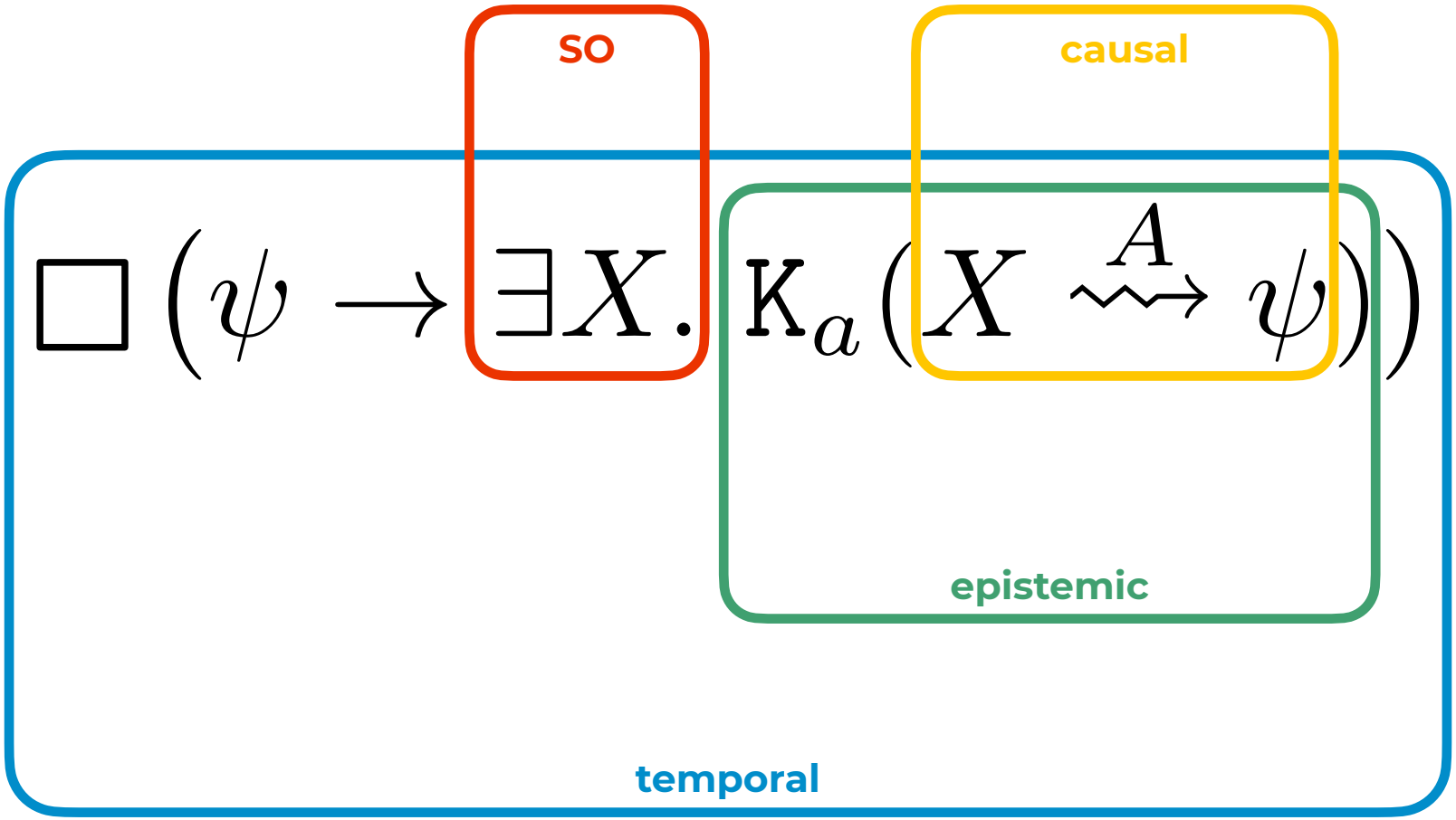
Specification & Verification:

$\models \Box \neg K_{\text{agent}}(\text{secret})$ (Privacy)

$\models$? (Explainability)

Key Takeaway

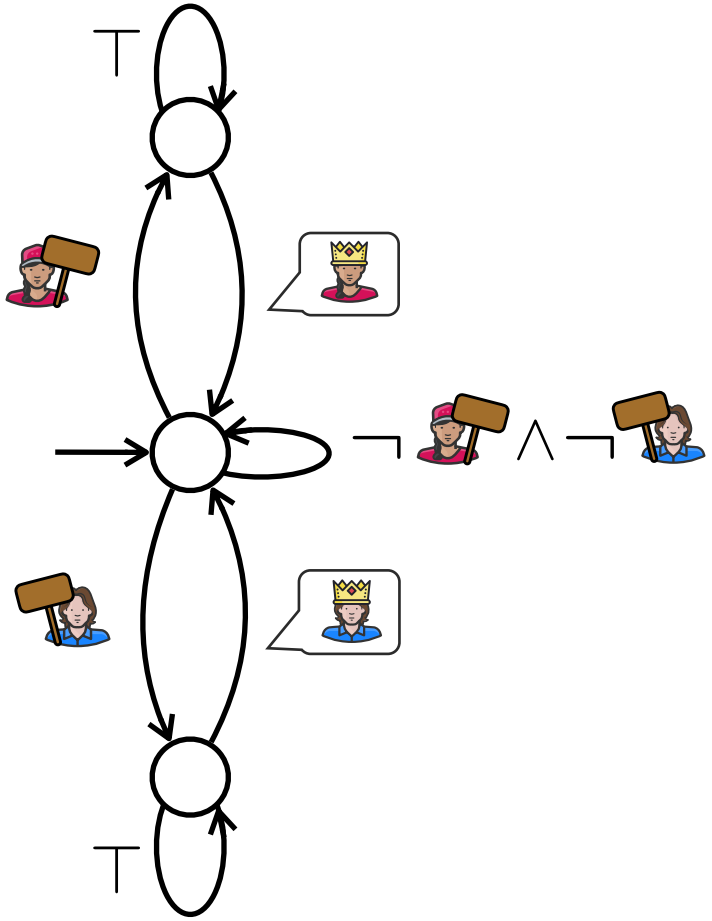
Specifying explainability requires a combination of causal, epistemic and temporal reasoning.

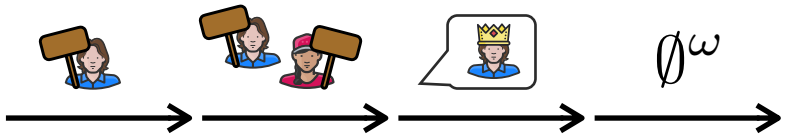


Running Example

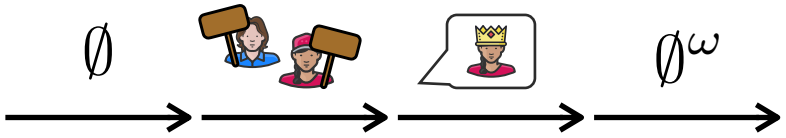
$$\square\left(\psi \to \exists X.\ \mathrm{K}_a(X \overset{A}{\rightsquigarrow} \psi)\right)$$

The state machine models a blind Dutch auction between  and  . The first bid wins.





Models generate traces.



Key Takeaway

Specifying explainability requires a combination of causal, epistemic and temporal reasoning.

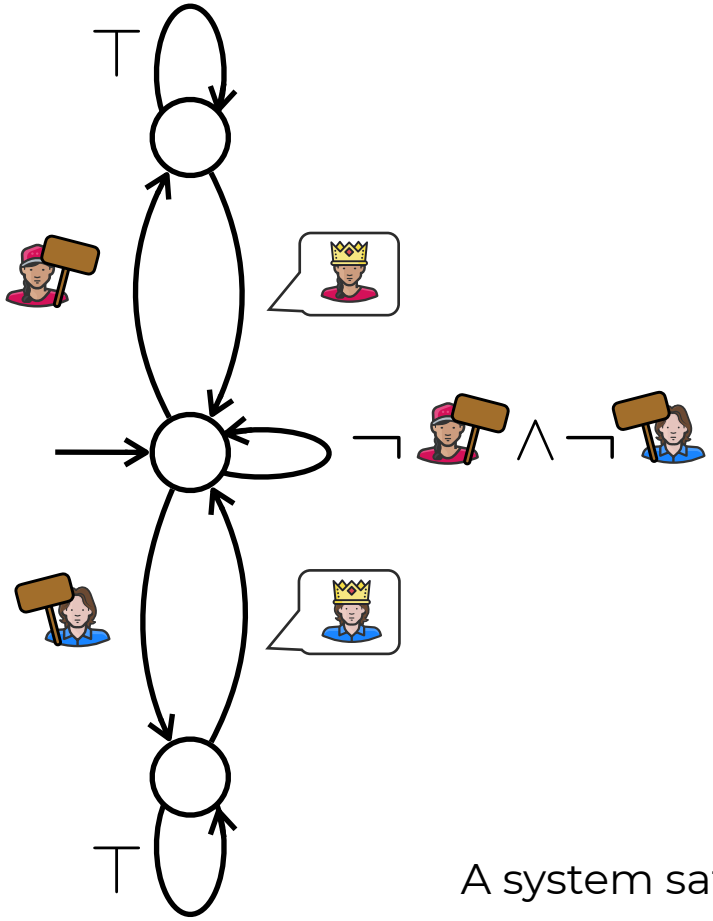
$$\Box\left(\psi \rightarrow \exists X.\ \mathtt{K}_a\left(X \overset{A}{\rightsquigarrow} \psi\right)\right)$$

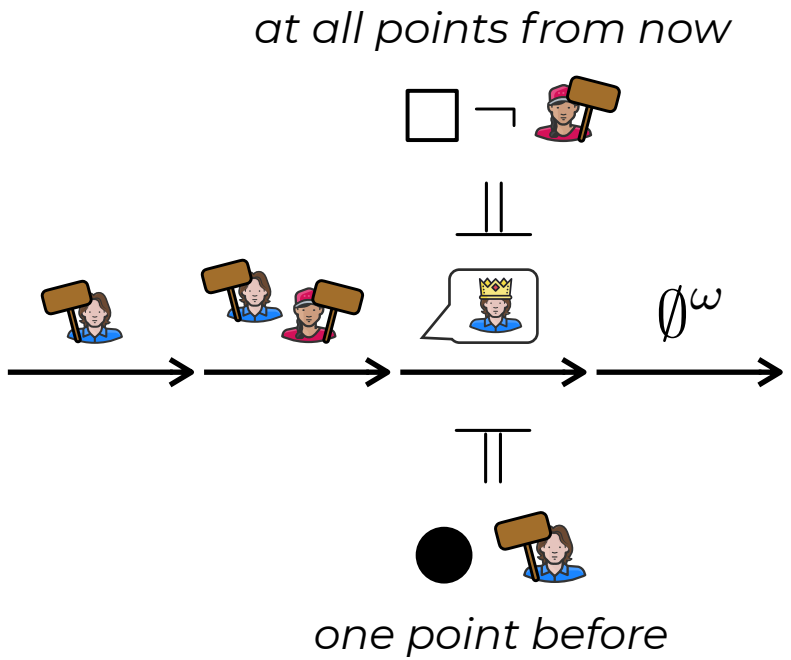
temporal

Temporal Operators [1]

$$\Box\left(\psi \rightarrow \exists X.\ \mathrm{K}_a(X \overset{A}{\rightsquigarrow} \psi)\right)$$

Formulas are evaluated with respect to a trace and time point (*anchor point*).



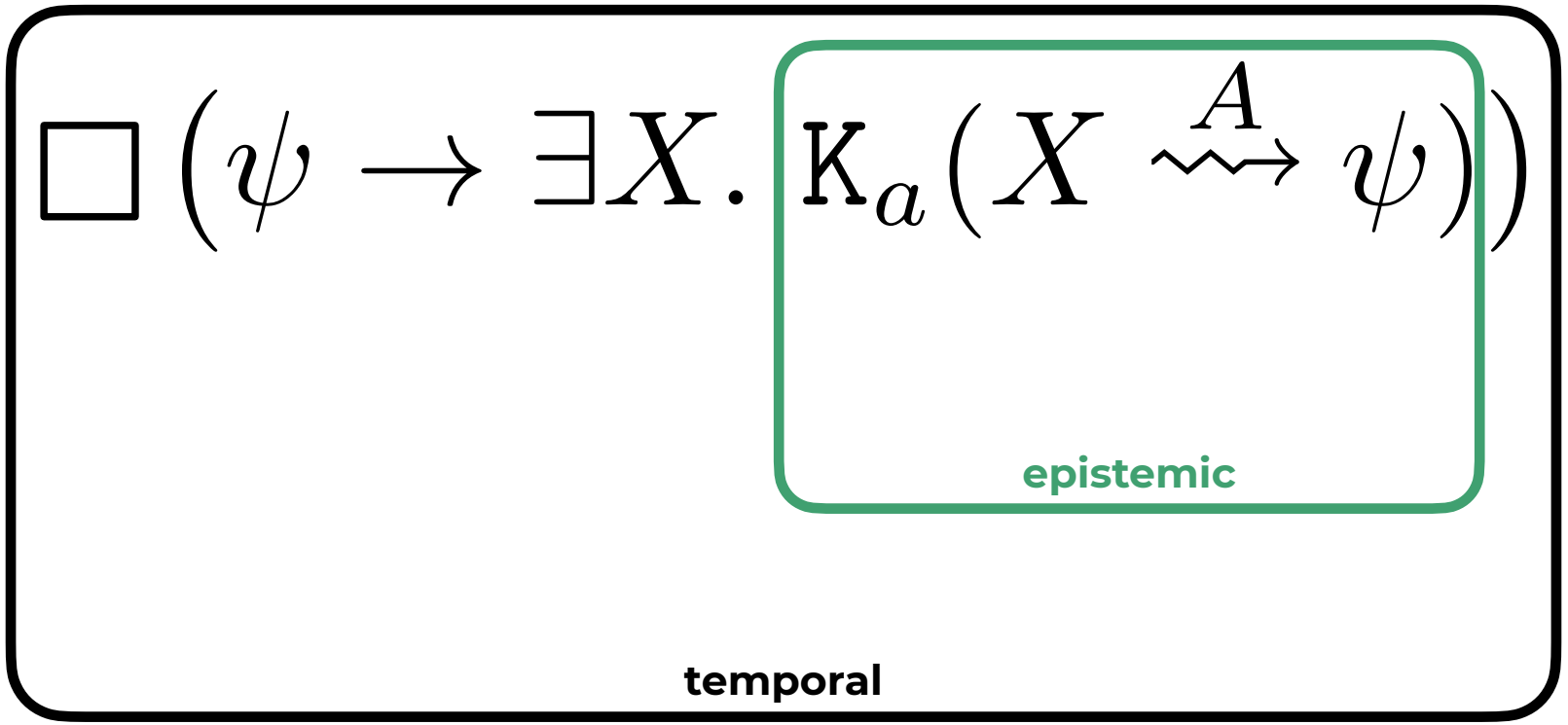


A system satisfies a formula if all initial trace satisfy the formula.

[1] Amir Pneuli. *The Temporal Logic of Programs.* FOCS 1977.

Key Takeaway

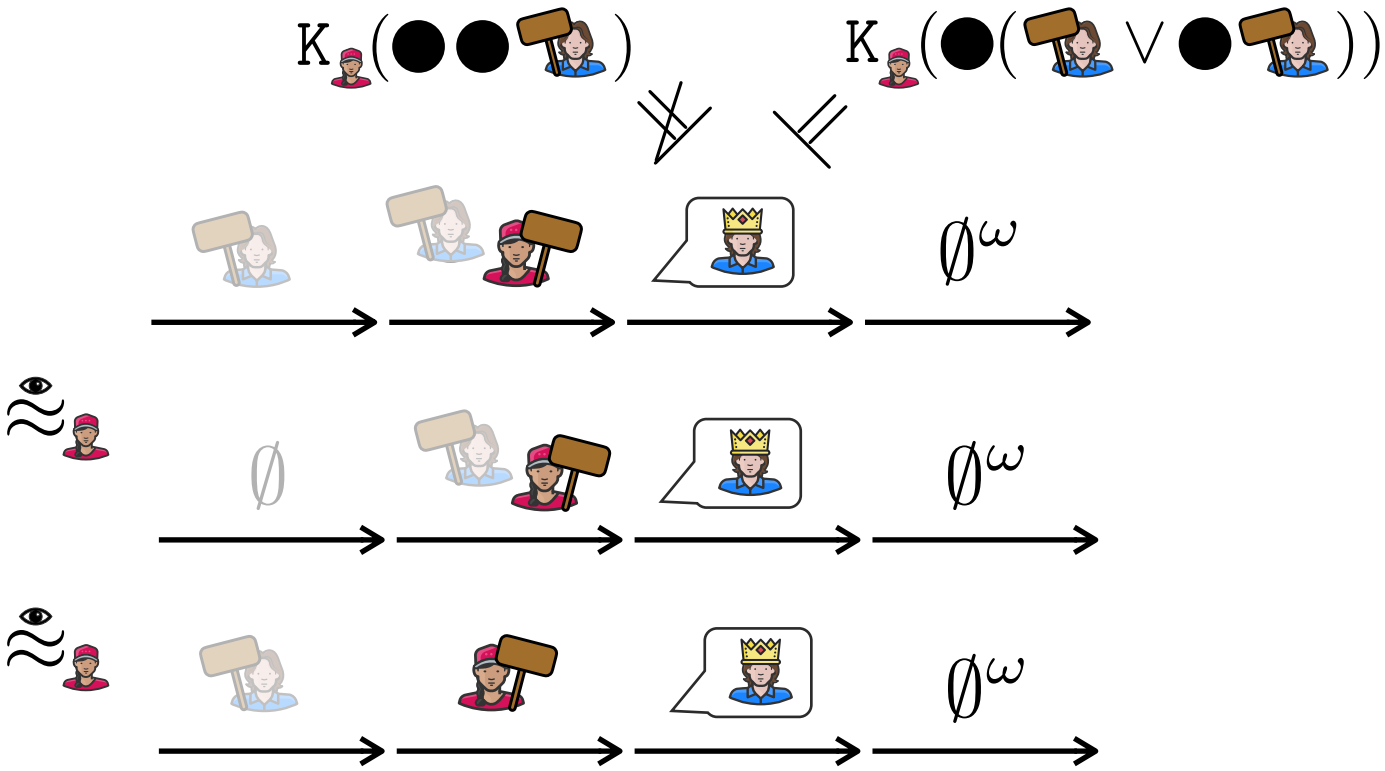
Specifying explainability requires a combination of causal, epistemic and temporal reasoning.



Epistemic Operators [2]

$$\square(\psi \to \exists X.\, \mathtt{K}_a(X \overset{A}{\rightsquigarrow} \psi))$$

Knowledge operators require their subformula to hold on all observation equivalent ($\approx_a$) traces.

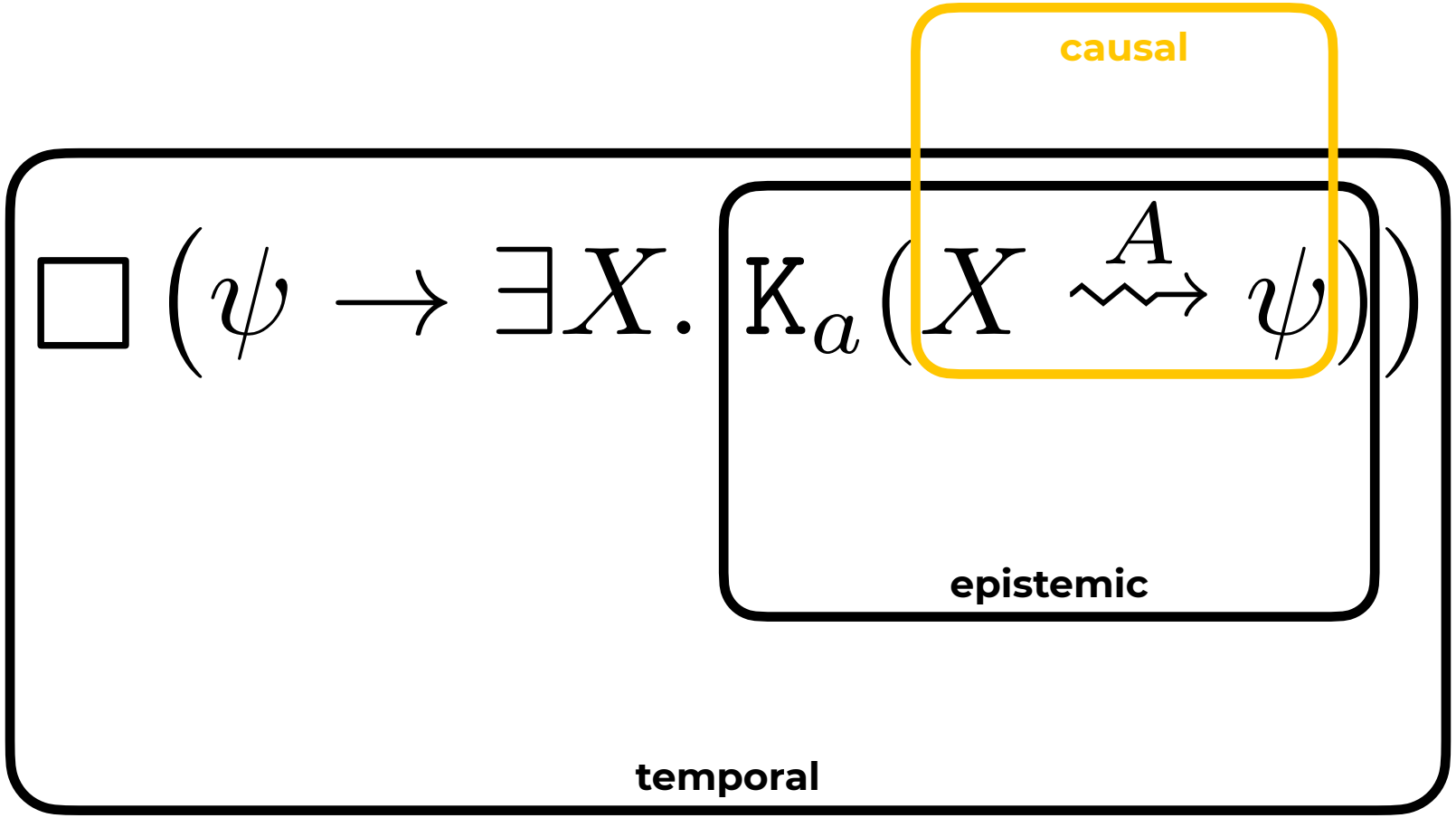


Synchronous perfect-recall semantics: Same observable propositions on full prefixes.

[2] Fagin, Halpern, Moses, and Vardi. *Reasoning About Knowledge*. MIT Press 1995.

Key Takeaway

Specifying explainability requires a combination of causal, epistemic and temporal reasoning.



Causal Dependencies

$$\square\big(\psi \to \exists X.\ \mathrm{K}_a(X \overset{A}{\rightsquigarrow} \psi)\big)$$

$X \overset{A}{\rightsquigarrow} \varphi$: Property X is a cause for property φ (when fixing all actions not in A).

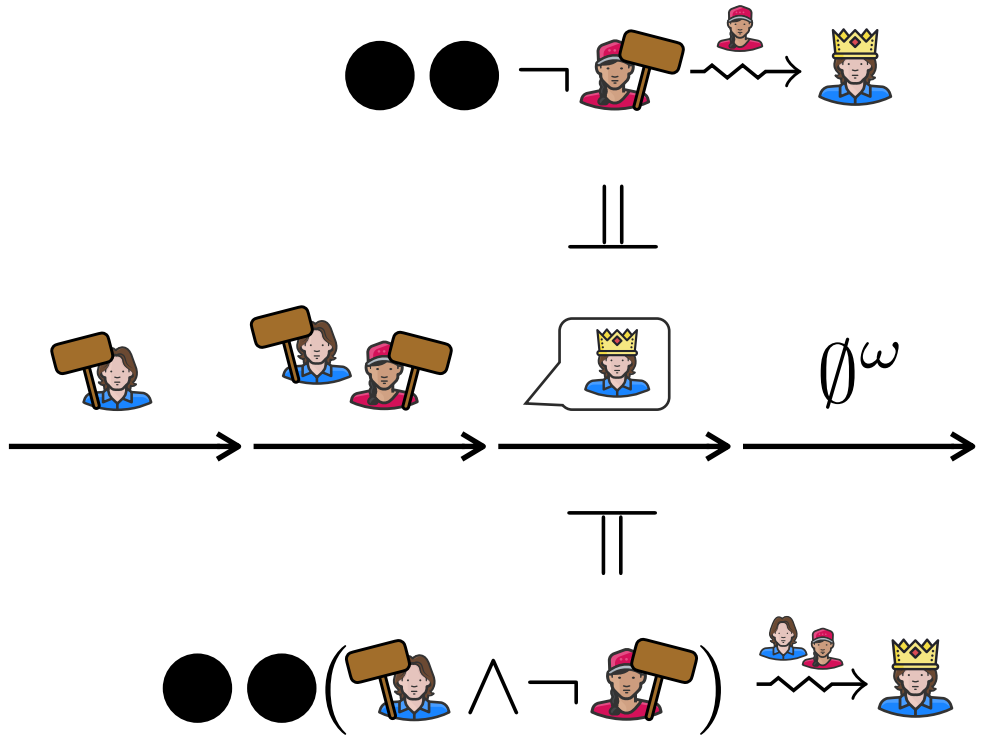
Temporal Causality [3]

SAT: Cause and effect are satisfied by the trace.

CF: Most similar traces that do not satisfy the cause also do not satisfy the effect.

MIN: Cause is (semantically) minimal.

Heavily inspired by actual causality [4].

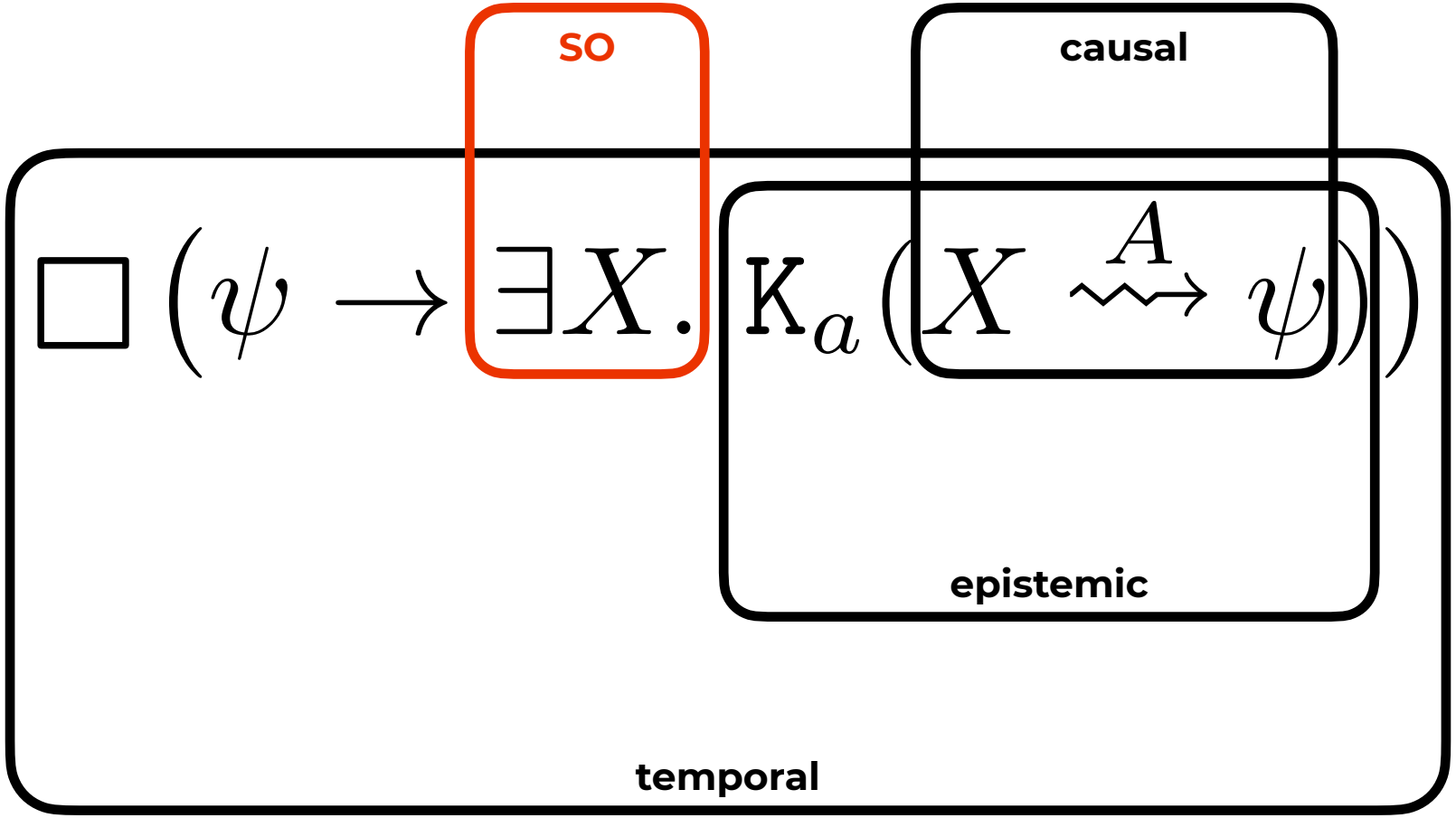


! There exists at most one unique causal property (which may symbolically describe multiple *events*).

[3] Finkbeiner, Frenkel, Metzger and Siber. Synthesis of Temporal Causality. CAV 2024.
[4] Halpern and Pearl. *Causes and Explanations: A Structural-Model Approach — Part 1: Causes.* British Journal for the Philosophy of Science 56 (4). 2005.

Key Takeaway

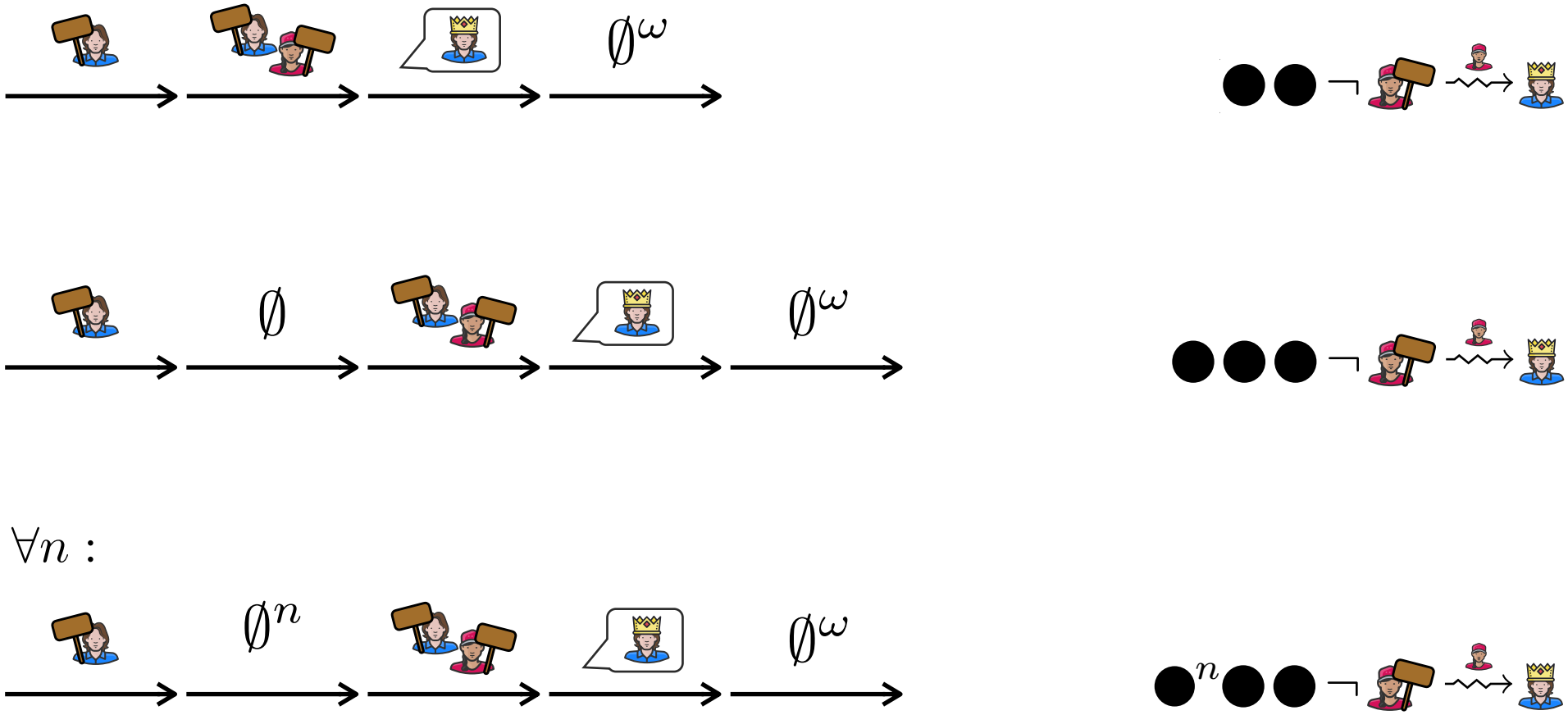
Specifying explainability requires a combination of causal, epistemic and temporal reasoning.



Why Second-Order Quantification?

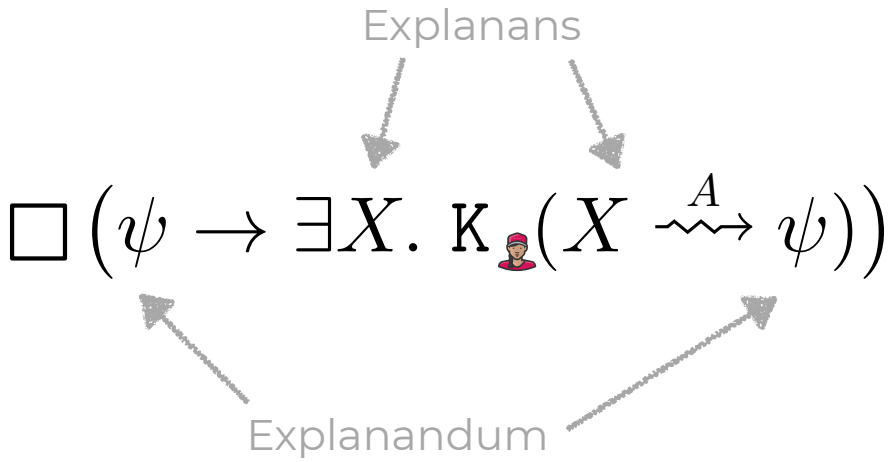
$$\Box\left(\psi \to \exists X.\ \mathrm{K}_a(X \overset{A}{\rightsquigarrow} \psi)\right)$$

A single explanandum can have infinitely many different explanantia:



Putting Everything Together

Explainability: Whenever a certain effect (*explanandum*) happens, an agent has knowledge about a cause (*explanans*) for this outcome.



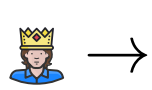
Internal Causal Explainability (ICE): $A = \{$ 👩 $\}$

External Causal Explainability (ECE): $A = Agents \setminus \{$ 👩 $\}$

Full Causal Explainability (FCE): $A = Agents$

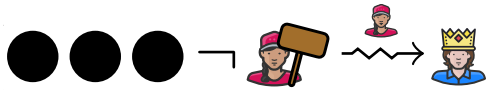
Information-Flow through Explanations

Semantics: Traces with different causes for the same explanandum must be distinguishable.

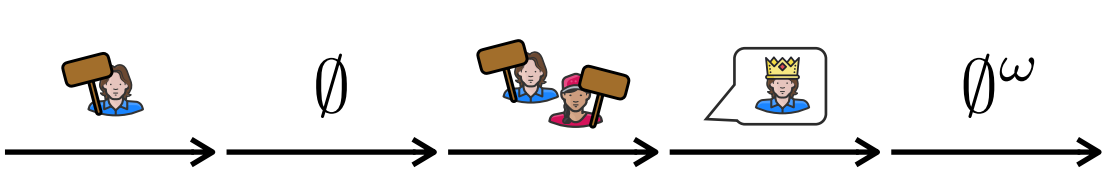
$\square\big($  $\rightarrow \exists X.\ \mathtt{K}_{}$  $(X \rightsquigarrow$  $))\big)$

$\not\Vdash$

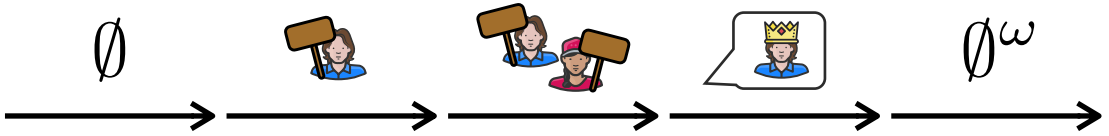
$\mathcal{T}_{blind}$



$\Vdash$



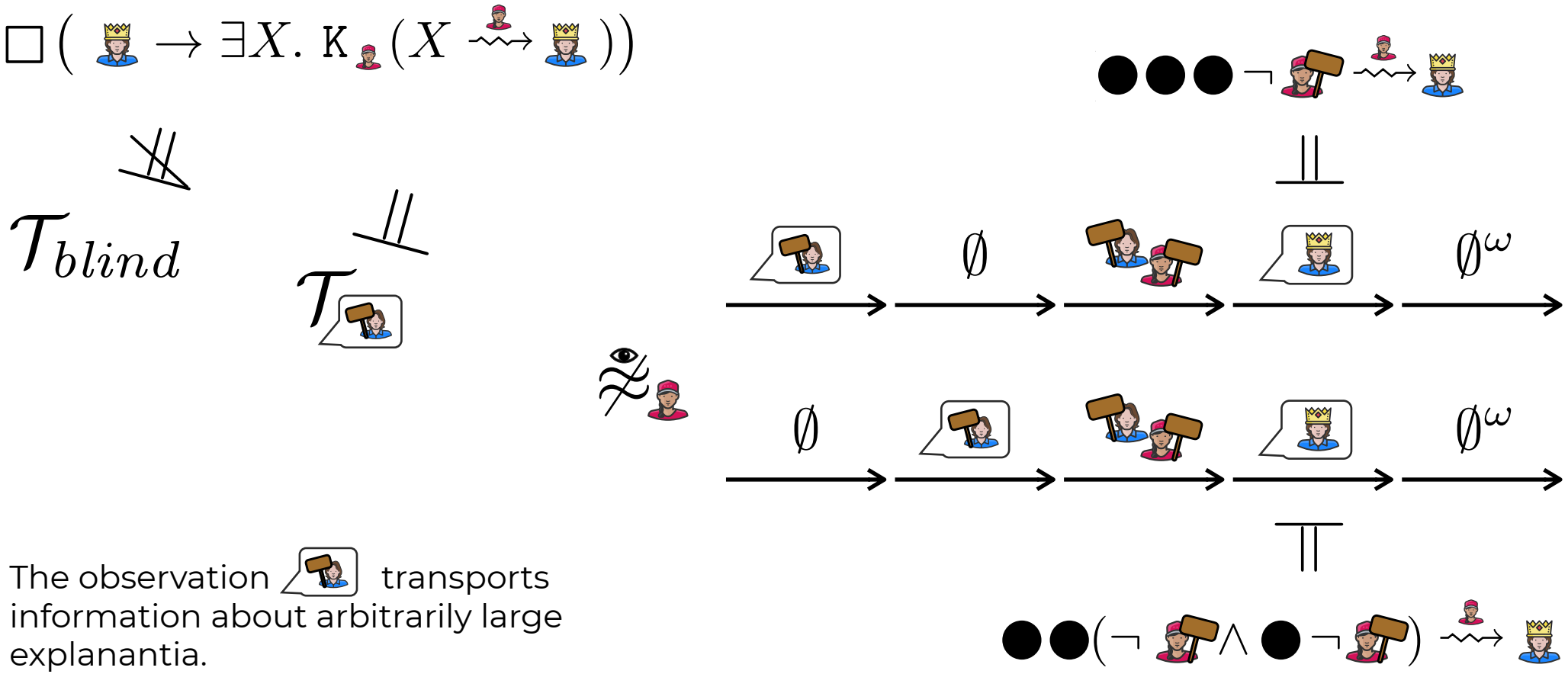
$\approx$ 



$\Vdash$

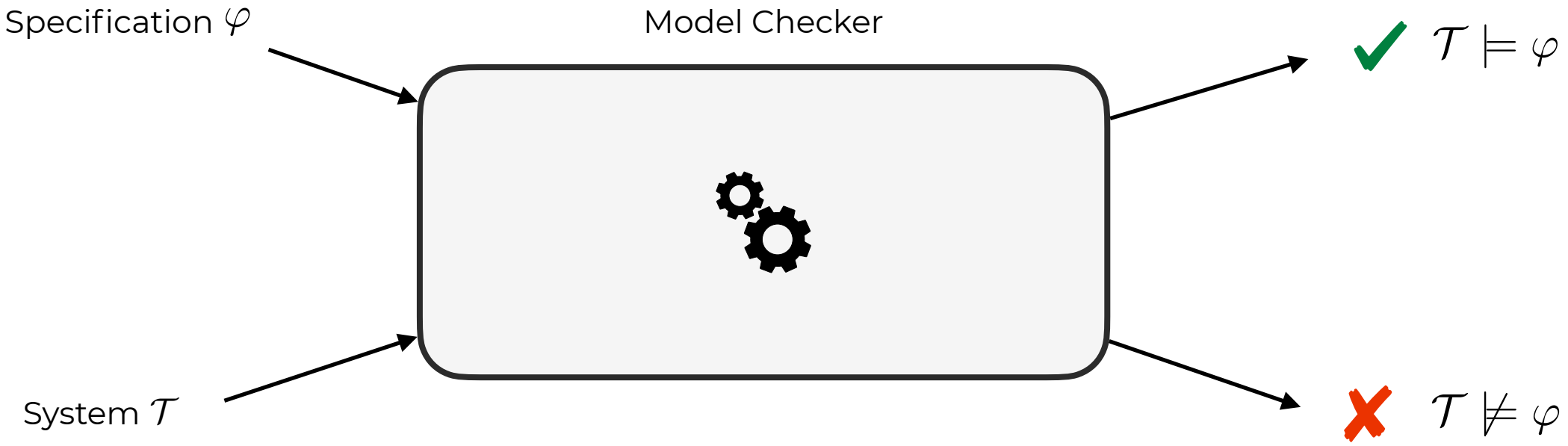
Information-Flow through Explanations

Semantics: Traces with different causes for the same explanandum must be distinguishable.

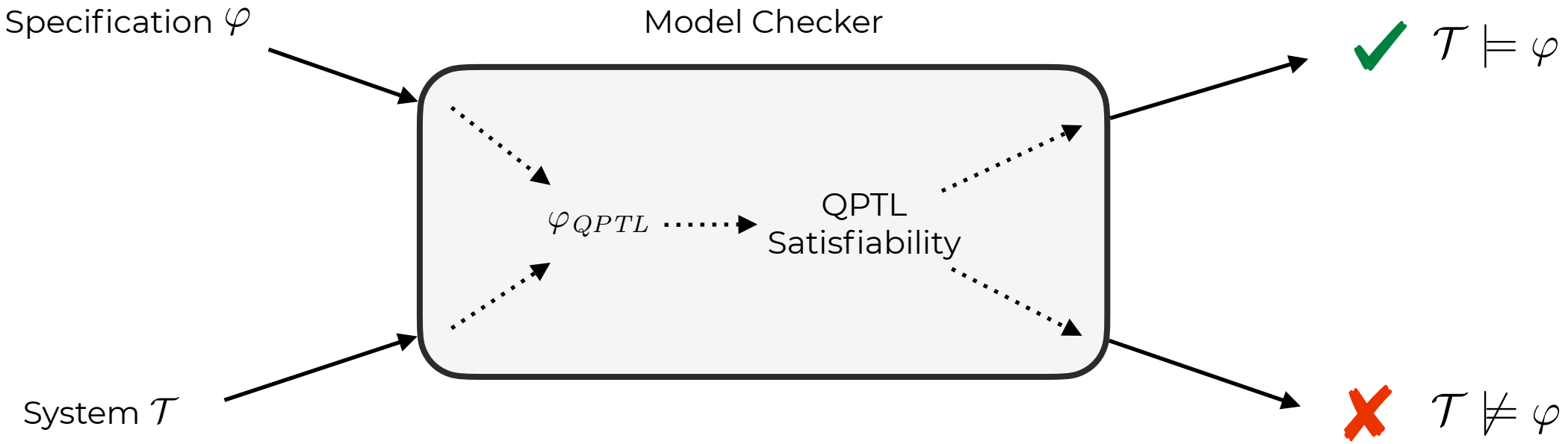


The observation transports information about arbitrarily large explanantia.

Automated Verification



High-Level Idea



QPTL: LTL + quantification over propositional sequences: $\xi ::= \forall q \in (2^{\{q\}})^{\omega}. \xi$

Simulate quantification over paths [5]:

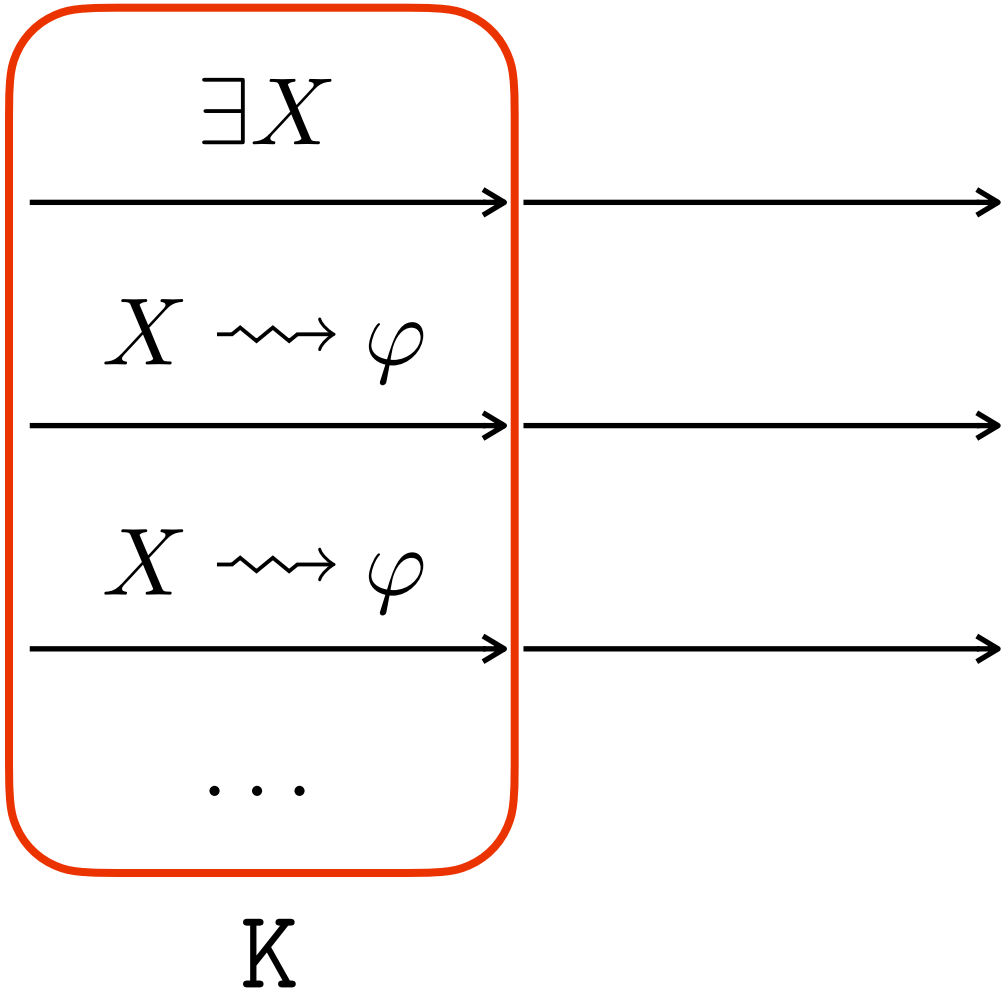
"behave like $\mathcal{T}$"

$$\forall a \in Actions, s \in States.\ \blacklozenge\left((\neg \bullet \top) \wedge \Box \widetilde{\mathcal{T}(a,s)}\right) \rightarrow \varphi_{LTL}(a,s)$$

[5] Bozzelli, Maubert, and Pinchinat. *Unifying Hyper and Epistemic Temporal Logics.* FoSSaCS 2015.

More Details

How to encode $\exists X.\ \mathtt{K}_a(X \rightsquigarrow \varphi)$?



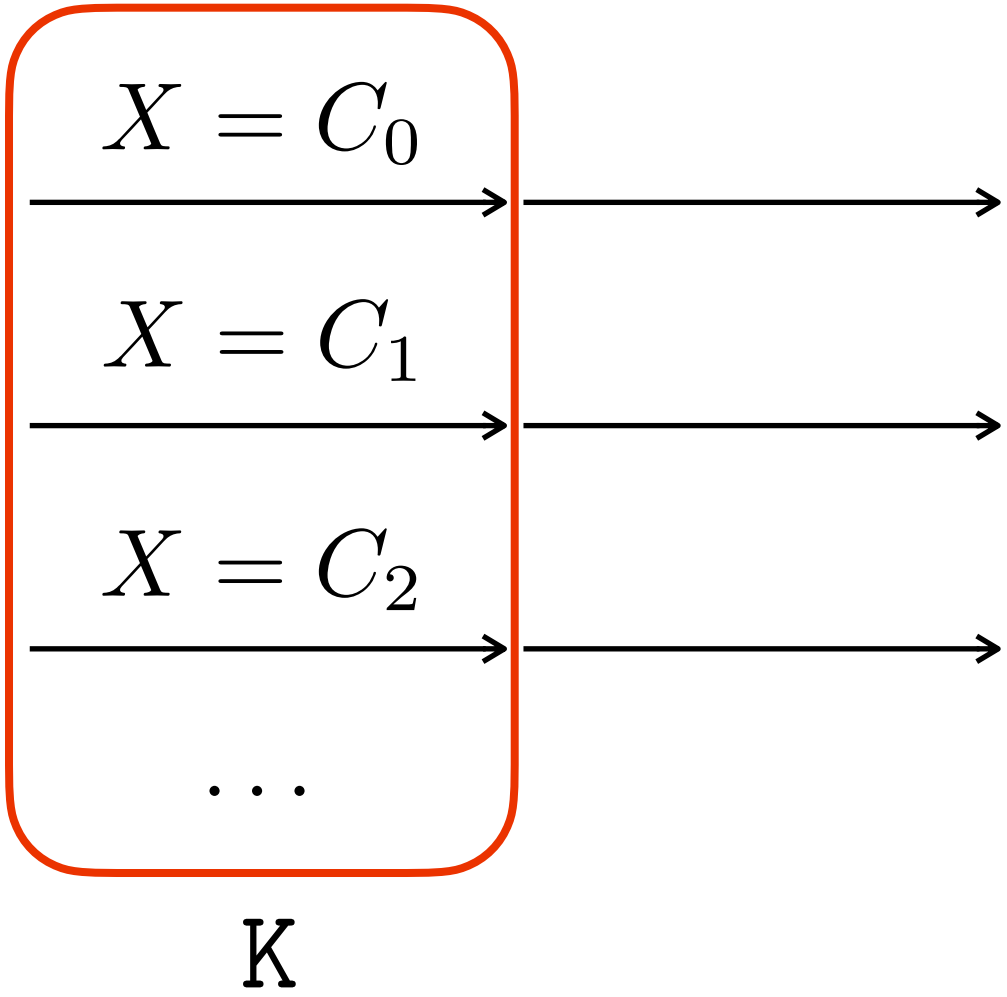
More Details

How to encode $\exists X.\ \mathtt{K}_a(X \rightsquigarrow \varphi)$?

There is a unique solution C_n for every $\rightsquigarrow \varphi$ and anchor point.

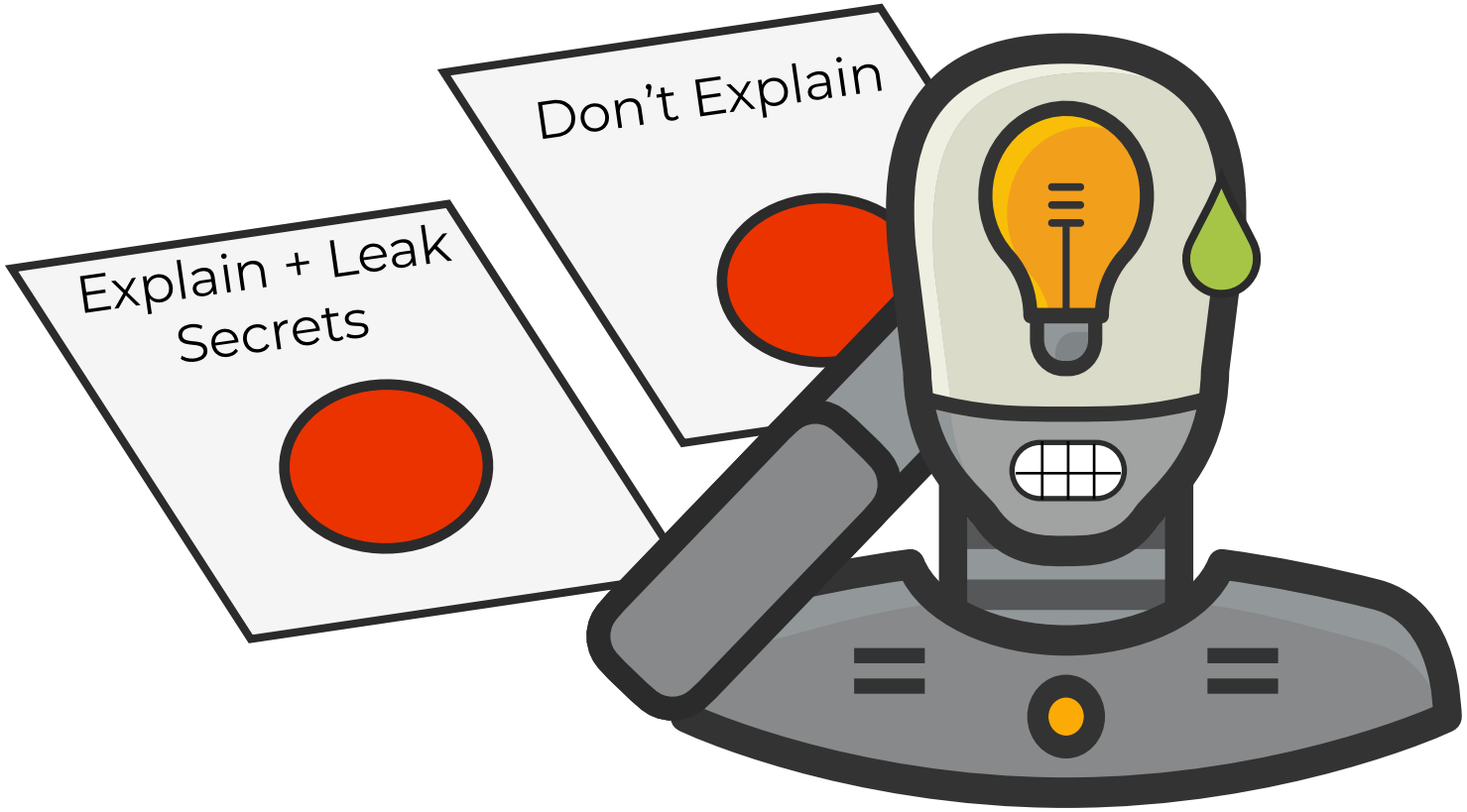
The formula requires that all the scoped solutions are equivalent:

$$C_0 = C_1 = C_2 = \dots$$

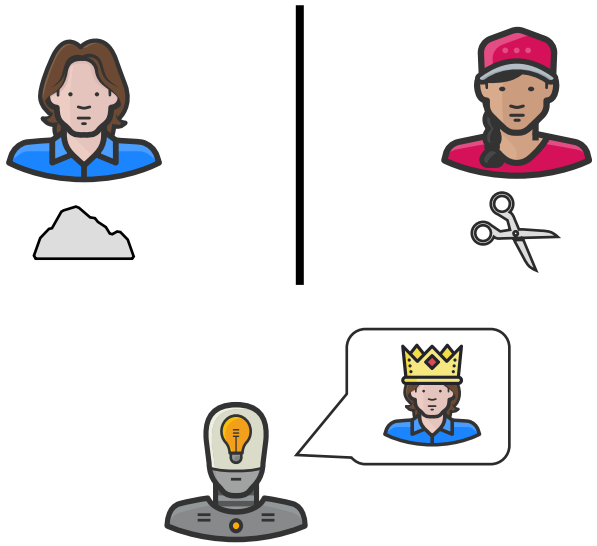


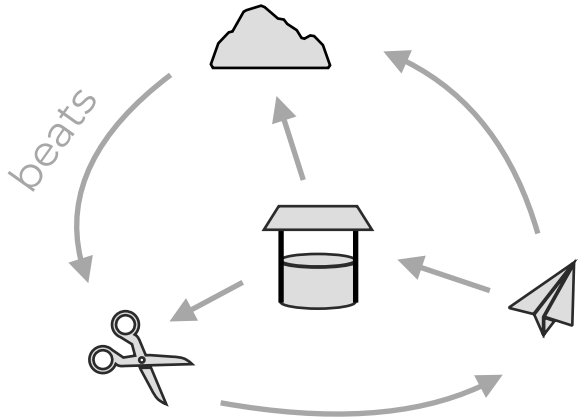
Use simulated quantification over paths to search for a trace that breaks the equality.

Experiments



Experiments: Blind RPS

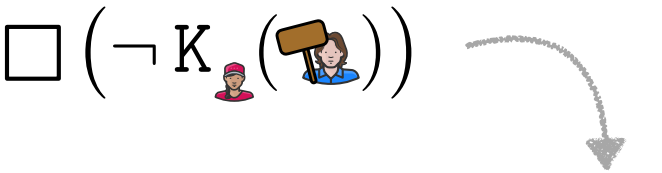


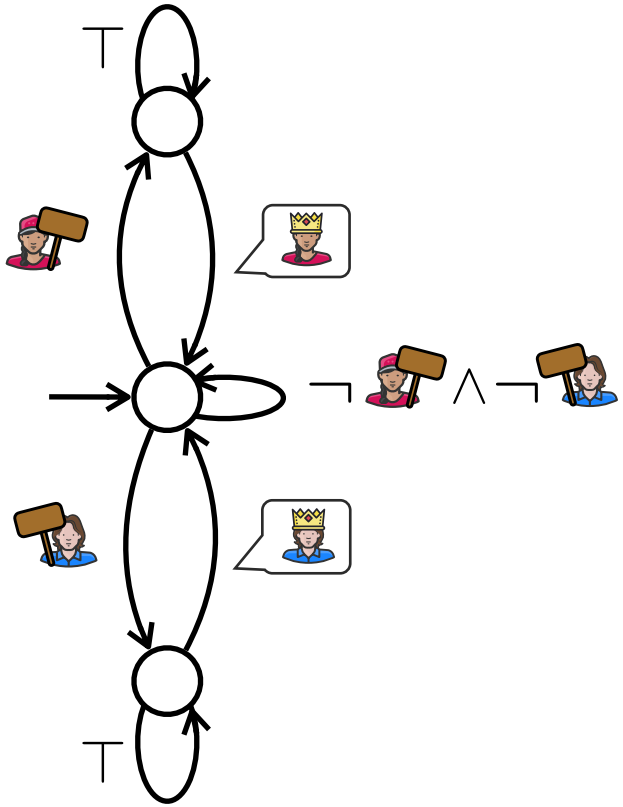


	ICE	ECE	FCE	Privacy
STANDARD	✓/0.49	✓/0.53	✓/0.55	✗/0.24

$$\Box\left(\neg draw \rightarrow \neg \mathtt{K}_{\text{👩}}(\text{☁}_{\text{👩‍🦰}})\right)$$

Experiments: Dutch Auction

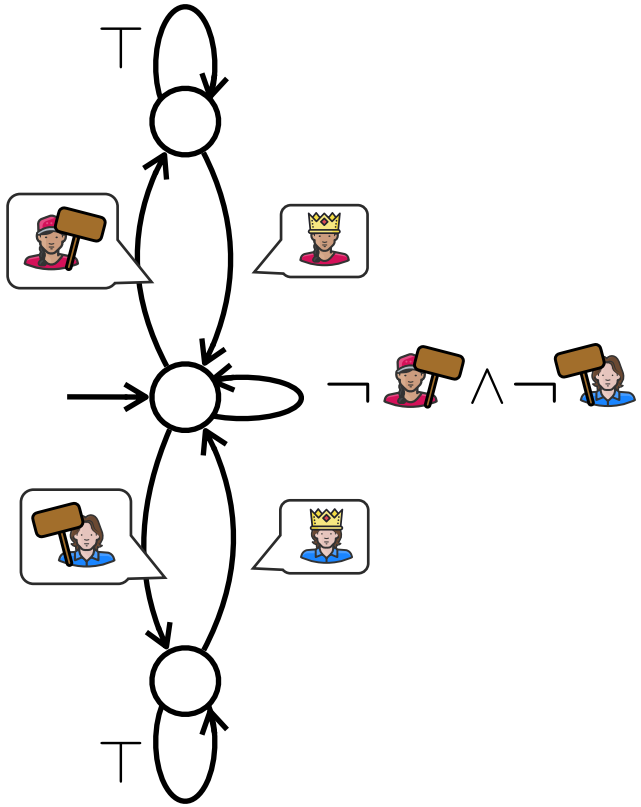




	$\lvert B \rvert$	ICE	ECE	FCE	Privacy
BLIND	2	✗/0.85	✗/0.88	✗/0.96	✓/0.25
	3	✗/1.57	✗/1.77	✗/1.73	✓/0.27
	4	✗/3.42	✗/3.68	✗/3.61	✓/0.27
	5	✗/7.85	✗/9.45	✗/10.3	✓/0.22

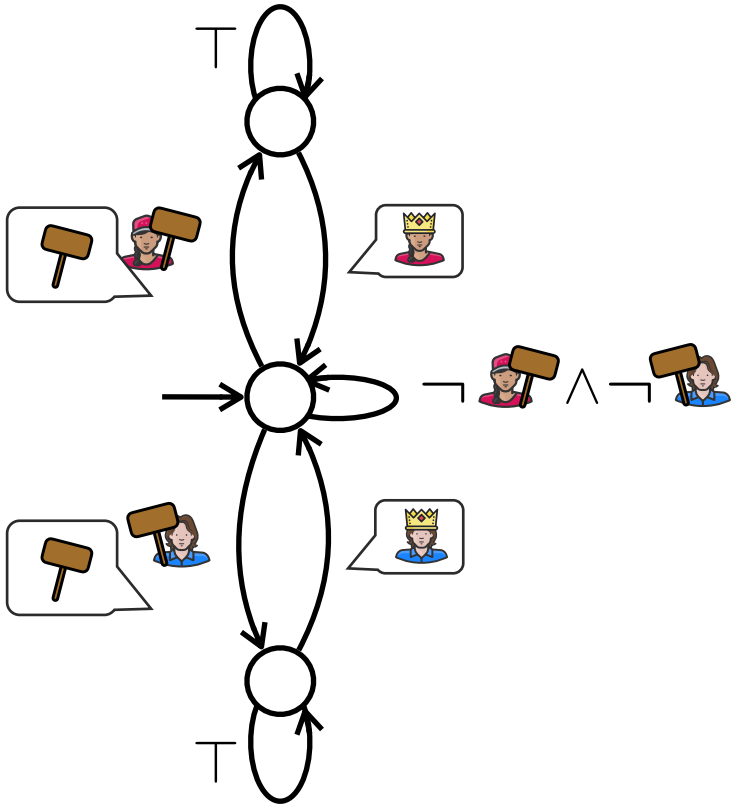
Experiments: Dutch Auction

$\Box\left(\neg \mathtt{K}_{\text{  }}\left(\text{  }\right)\right)$



	$\lvert B\rvert$	ICE	ECE	FCE	Privacy
BLIND	2	✗/0.85	✗/0.88	✗/0.96	✓/0.25
	3	✗/1.57	✗/1.77	✗/1.73	✓/0.27
	4	✗/3.42	✗/3.68	✗/3.61	✓/0.27
	5	✗/7.85	✗/9.45	✗/10.3	✓/0.22
PUBLIC	2	✓/0.78	✓/0.87	✓/0.81	✗/0.30
	3	✓/1.47	✓/1.80	✓/1.63	✗/0.30
	4	✓/3.19	✓/3.73	✓/3.48	✗/0.50
	5	✓/7.24	✓/ 9.52	✓/10.3	✗/1.21

Experiments: Dutch Auction



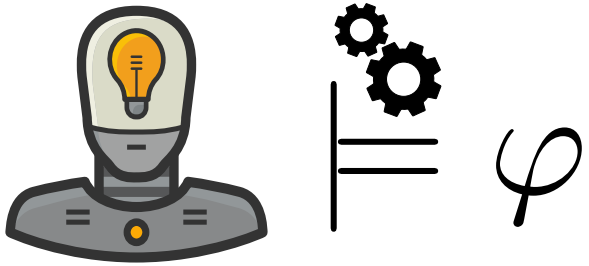
	$\lvert B\rvert$	ICE	ECE	FCE	Privacy
BLIND	2	✗/0.85	✗/0.88	✗/0.96	✓/0.25
	3	✗/1.57	✗/1.77	✗/1.73	✓/0.27
	4	✗/3.42	✗/3.68	✗/3.61	✓/0.27
	5	✗/7.85	✗/9.45	✗/10.3	✓/0.22
PUBLIC	2	✓/0.78	✓/0.87	✓/0.81	✗/0.30
	3	✓/1.47	✓/1.80	✓/1.63	✗/0.30
	4	✓/3.19	✓/3.73	✓/3.48	✗/0.50
	5	✓/7.24	✓/ 9.52	✓/10.3	✗/1.21
EXPLAIN	2	✓/0.86	✗/1.02	✗/1.02	✗/0.24
	3	✓/1.39	✗/1.89	✗/1.82	✓/0.25
	4	✓/3.40	✗/3.82	✗/4.03	✓/0.29
	5	✓/8.03	✗/ 9.92	✗/10.4	✓/0.29

Conclusion

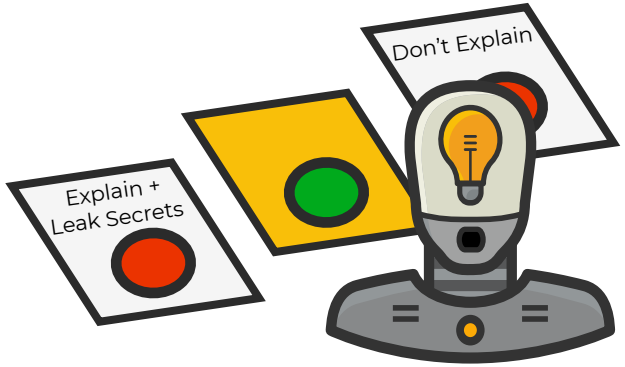
Explainability and privacy need to be balanced

$$\Box\left(\text{✗} \rightarrow \exists X.\ \mathtt{K}_{\text{👩}}\left(X \overset{\text{👩}}{\rightsquigarrow} \text{✗}\right)\right)$$

Explainability is an information-flow property



The information flow of explainability can be verified



Good explanations do not need to sacrifice privacy