Counterfactual Off-Policy Evaluation with Gumbel-Max Structural Causal Models

Michael Oberst
MIT

David Sontag
MIT

@MichaelOberst

CSAIL

imes
INSTITUTE FOR MEDICAL ENGINEERING & SCIENCE

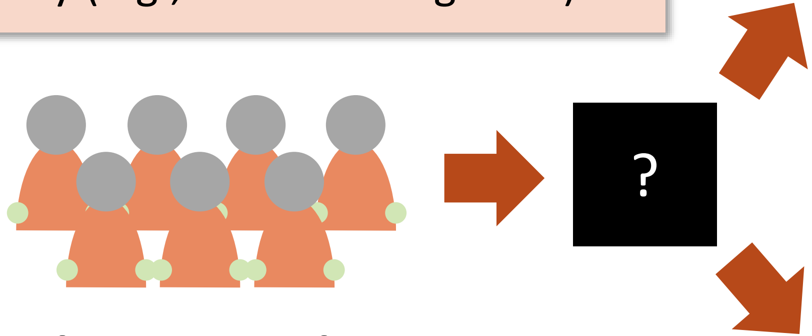
HST
HEALTH SCIENCES & TECHNOLOGY

Motivation: Building trust in RL policies

- **Goal**: Apply reinforcement learning in high risk settings (e.g., healthcare)
- **Problem**: How to safely evaluate a policy? No simulator, and off-policy evaluation can fail due to
  - Confounding
  - Small sample sizes
  - Poorly specified rewards
- Could try to interpret the policy directly, but if not possible, what can we do?

Motivation: Building trust in RL policies

Suppose we are given:
- Markov Decision Process (MDP)
- Policy (e.g., learned using MDP)



Observational Data

Markov Decision Process (MDP)

$P(S', R \mid S, A)$

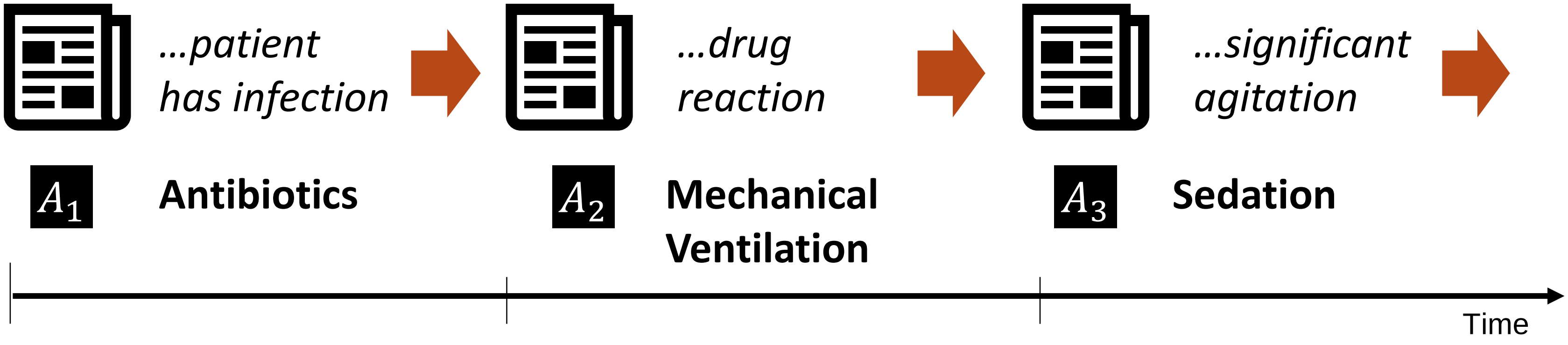
S: Current State
A: Action
R: Reward
S': Next State

Policy

$\pi(A \mid S)$

Using counterfactuals to "sanity check"

S: State
A: Action



Using counterfactuals to “sanity check”

*If the new policy **had been** applied to this patient…*

S: State
A: Action

…patient has infection

…drug reaction

…significant agitation

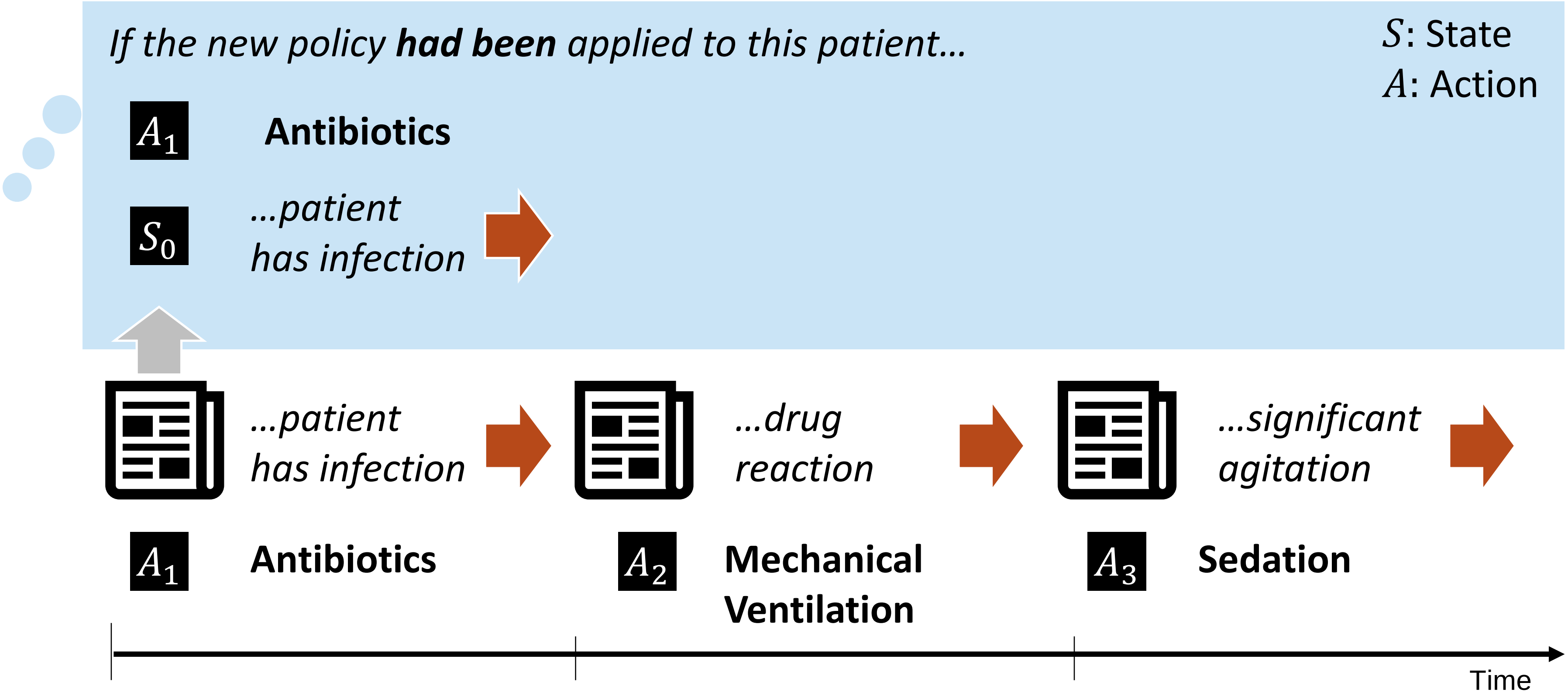
A_1 **Antibiotics**

A_2 **Mechanical Ventilation**

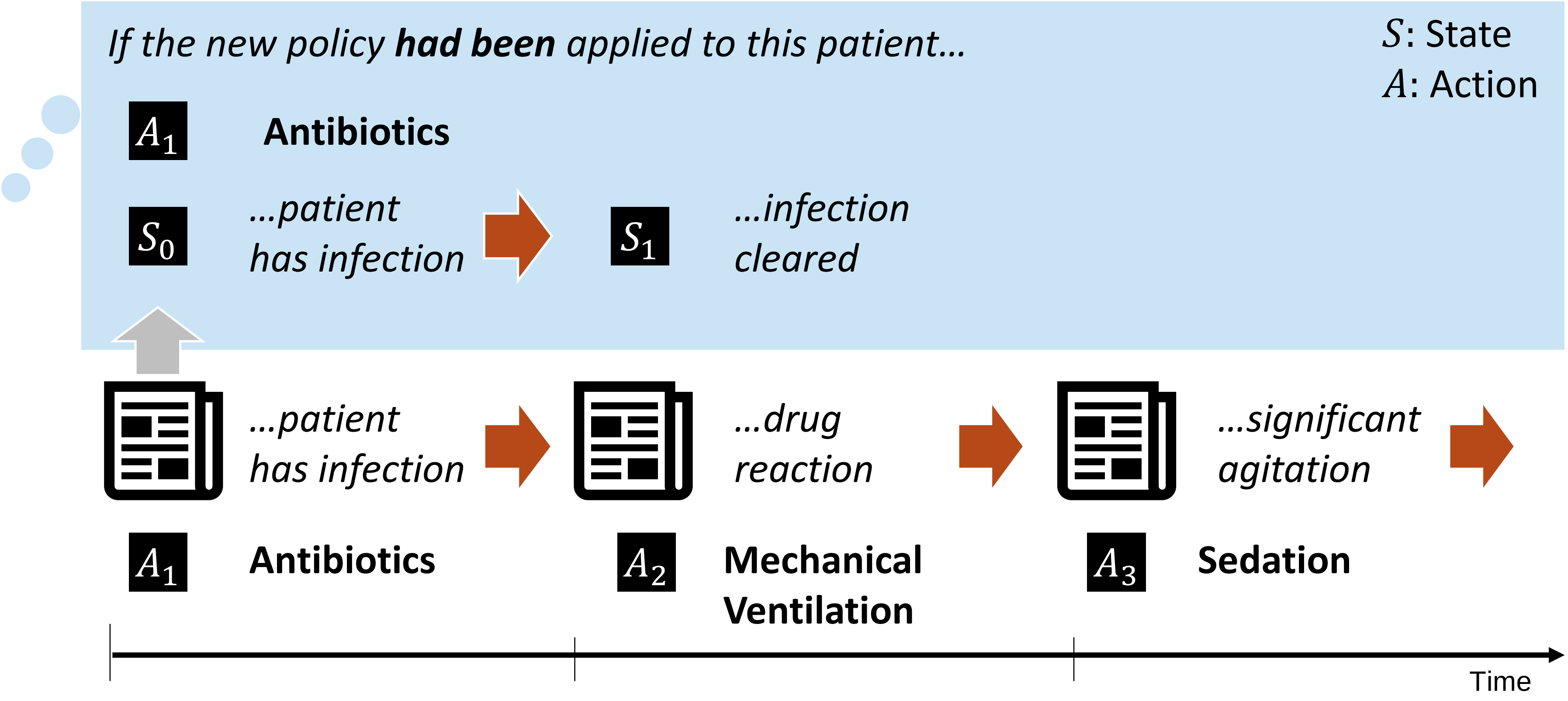
A_3 **Sedation**

Time

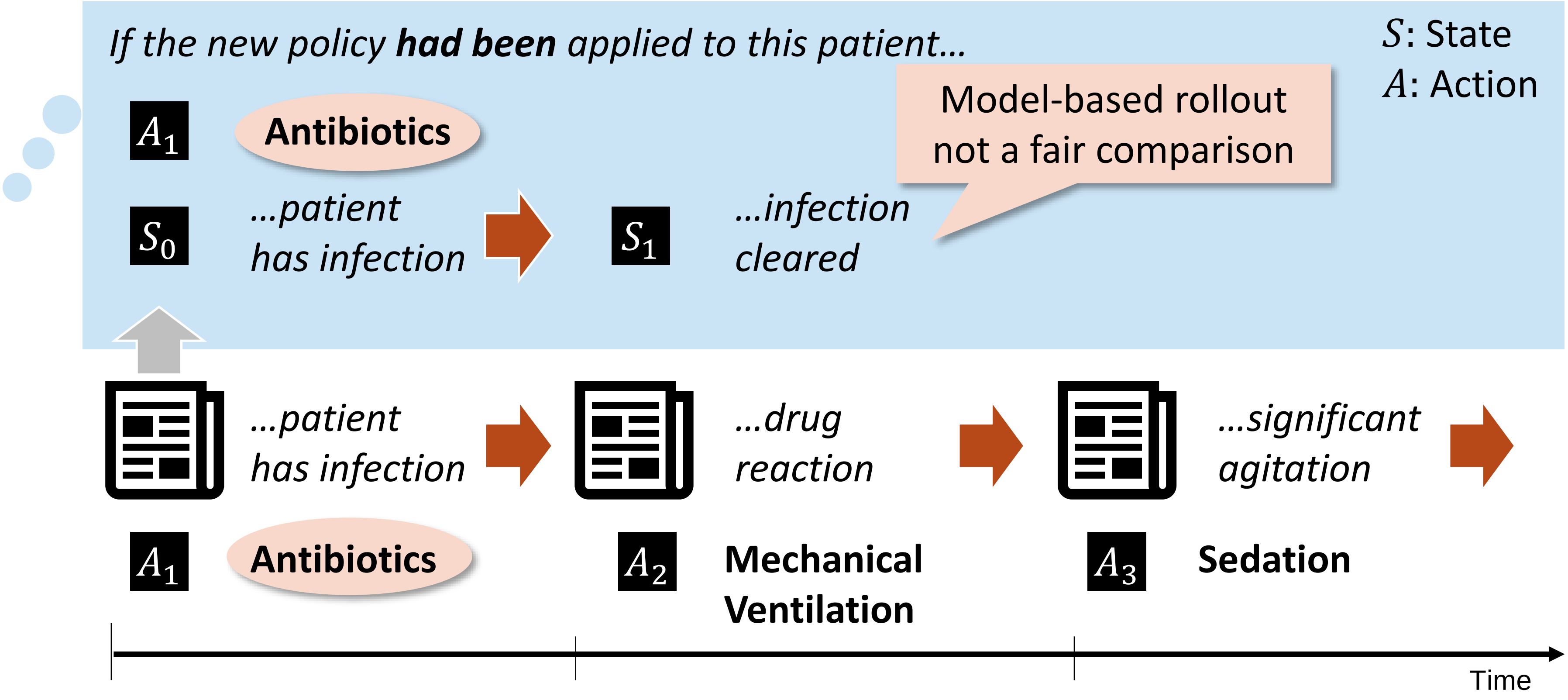
Using counterfactuals to “sanity check”



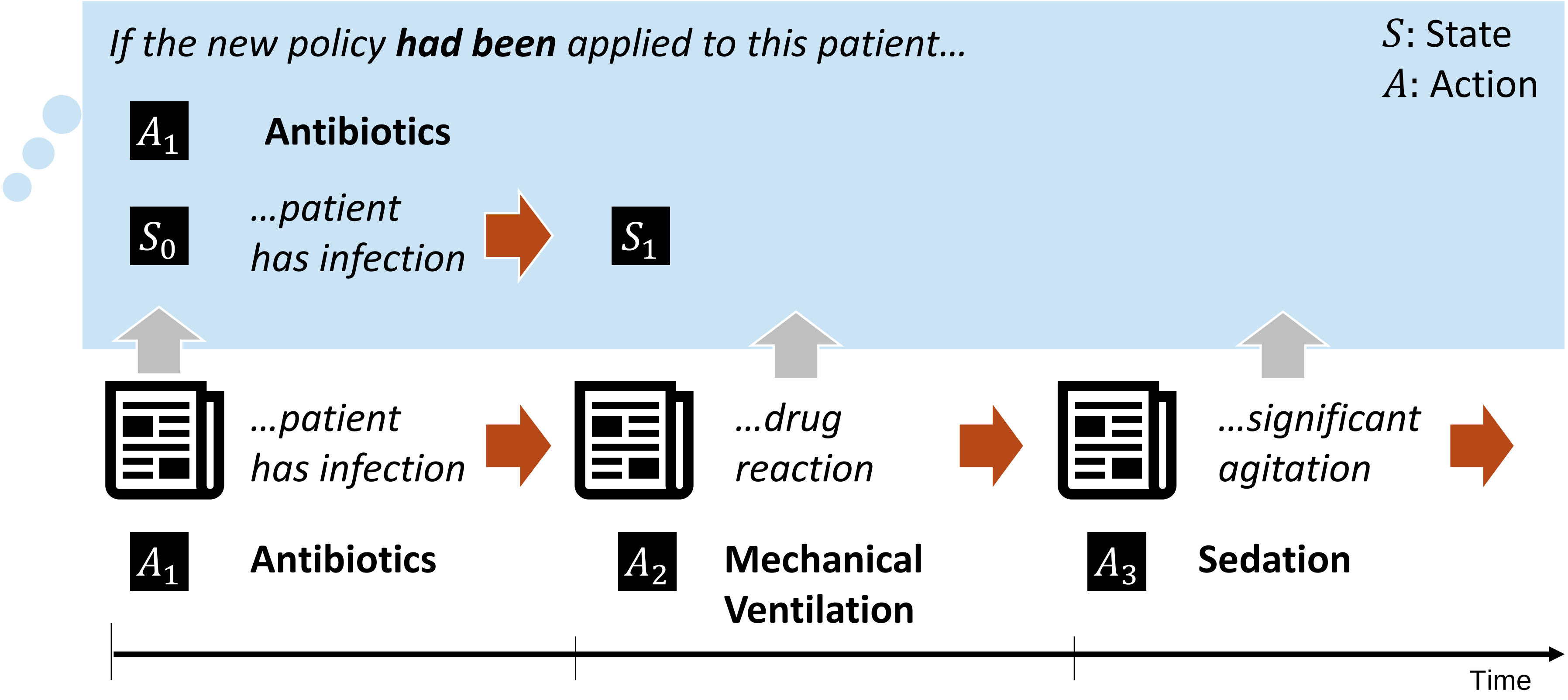
Using counterfactuals to "sanity check"



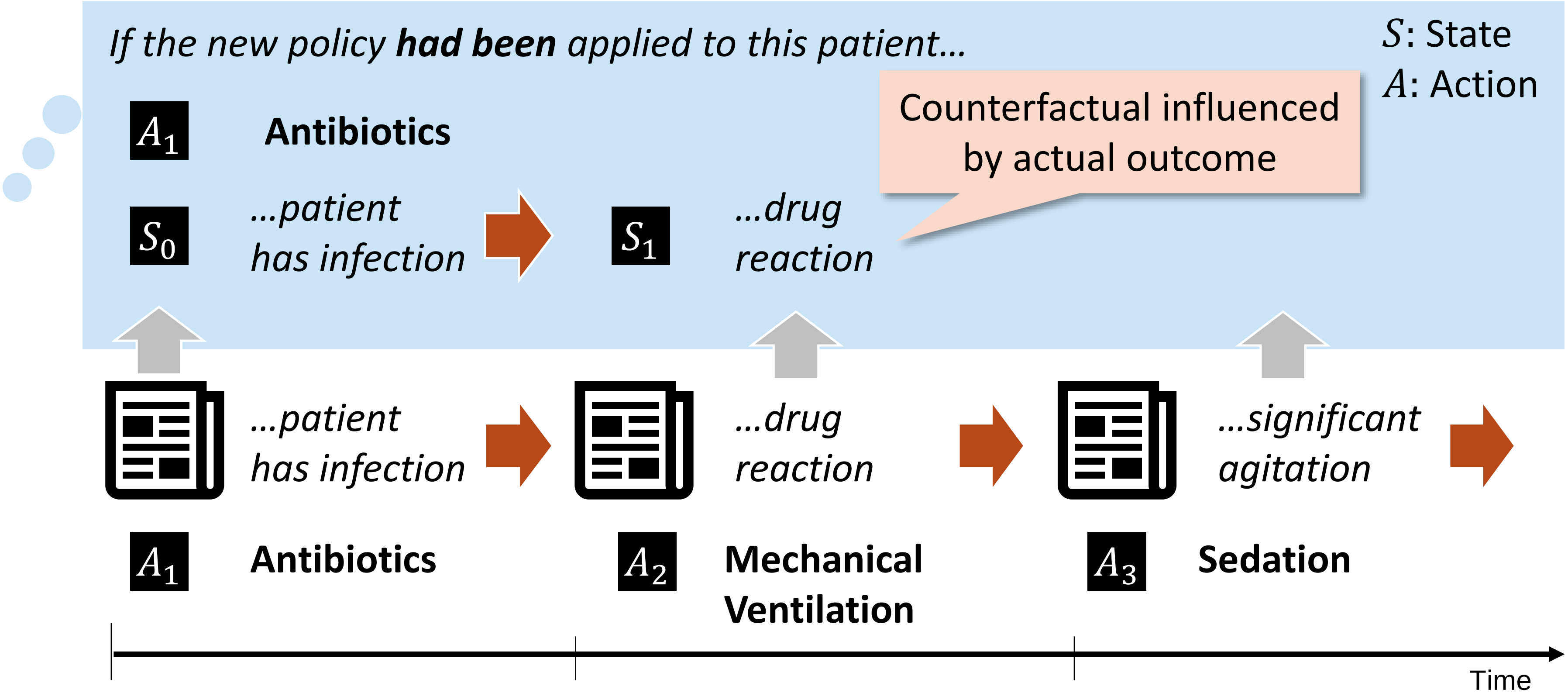
Using counterfactuals to "sanity check"



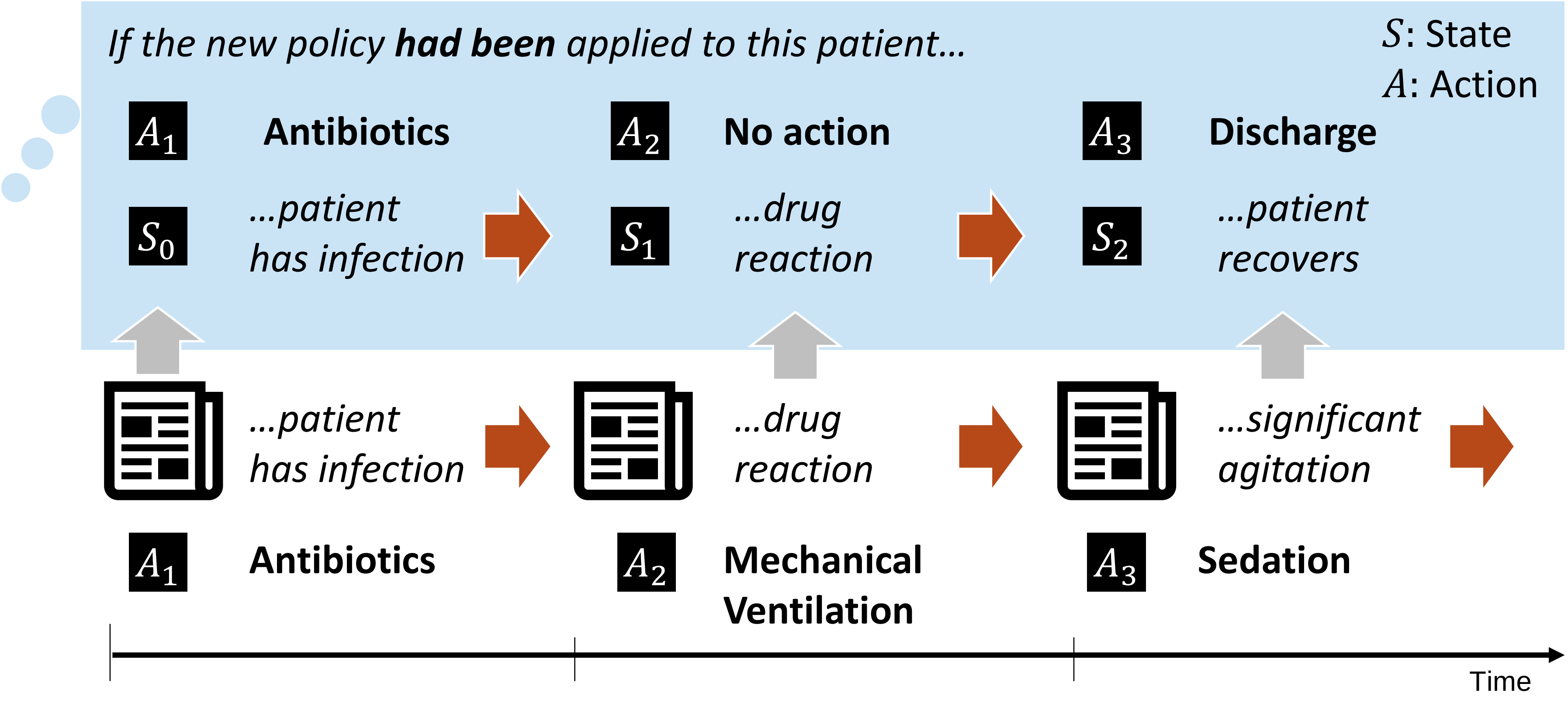
Using counterfactuals to “sanity check”



Using counterfactuals to “sanity check”



Using counterfactuals to “sanity check”



Using counterfactuals to “sanity check”

*If the new policy **had been** applied to this patient…*

S: State
A: Action

A_1	**Antibiotics**	A_2	**No action**	A_3	**Discharge**
S_0	*…patient has infection*	S_1	*…drug reaction*	S_2	*…patient recovers*

…patient has infection → *…drug reaction* → *…significant agitation* →

A_1	**Antibiotics**	A_2	**Mechanical Ventilation**	A_3	**Sedation**

Idea: If the counterfactual trajectory is unreasonable given full context of patient, the model / policy may be flawed

Using counterfactuals to "sanity check"

Approach

1. **Decomposition of reward** over real episodes, to identify interesting cases

See paper / poster for synthetic case study motivated by sepsis management

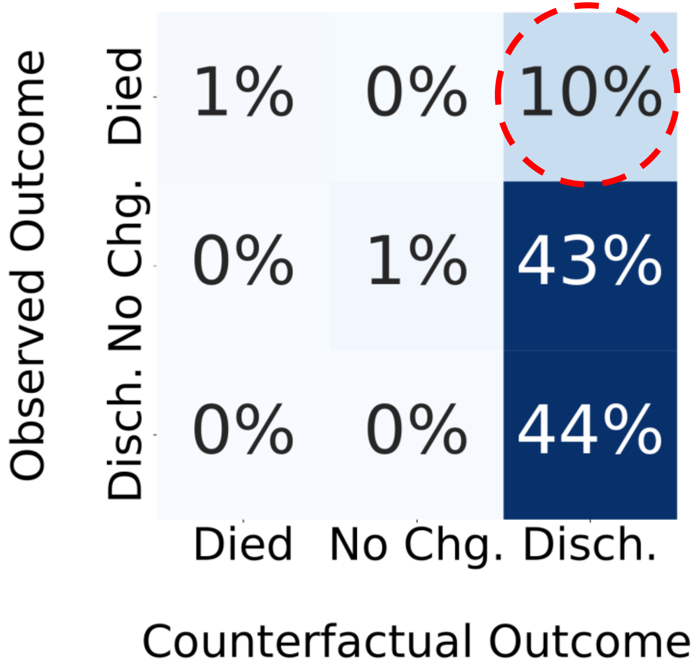
Using counterfactuals to "sanity check"

Approach

1. **Decomposition of reward** over real episodes, to identify interesting cases

See paper / poster for synthetic case study motivated by sepsis management

Example



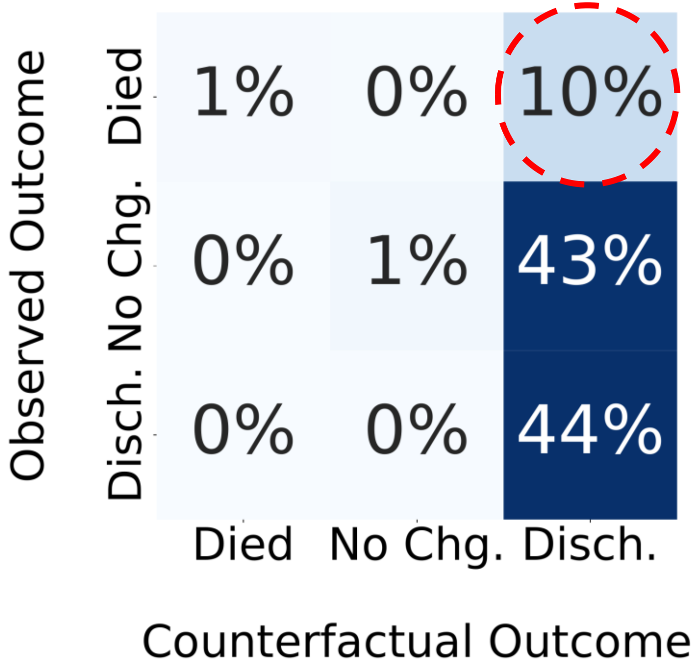
Using counterfactuals to “sanity check”

Approach

1. **Decomposition of reward** over real episodes, to identify interesting cases
2. **Examine counterfactual trajectories** under new policy
3. **Validate and/or criticize** conclusions, using full patient information (e.g., chart review)

See paper / poster for synthetic case study motivated by sepsis management

Example



Simulating counterfactual trajectories

What we need

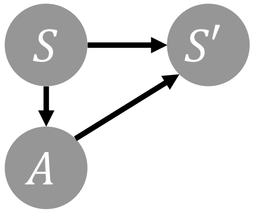
1 Observed trajectories

2 Policy to evaluate

$$\pi(A \mid S)$$

3 Model of discrete dynamics, e.g., Markov Decision Process

S: Current State
A: Action
S': Next State



Simulating counterfactual trajectories

What we need

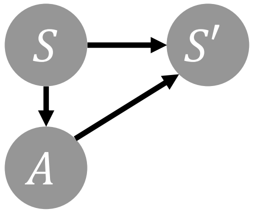
1. Observed trajectories

2. Policy to evaluate

$$\pi(A \mid S)$$

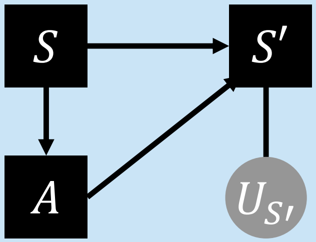
3. Model of discrete dynamics, e.g., Markov Decision Process

S: Current State
A: Action
S': Next State



\+

Structural Causal Model (SCM)



$$S' = f(S, A, U_{S'})$$
$$U_{S'} \sim P(U_{S'})$$

Simulating counterfactual trajectories

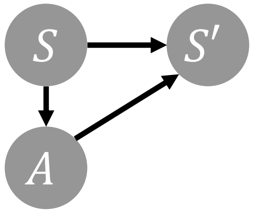
What we need

1. Observed trajectories
2. Policy to evaluate

   $\pi(A \mid S)$

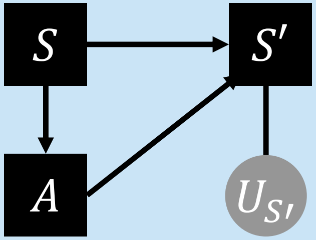
3. Model of discrete dynamics, e.g., Markov Decision Process

S: Current State
A: Action
S': Next State



\+

Structural Causal Model (SCM)



$$S' = f(S, A, U_{s'})$$
$$U_{s'} \sim P(U_{s'})$$

Problem: Choice of SCM is not identifiable from data!

So, what should we use for the structural causal model (SCM)?

Key challenge: Non-identifiability

There are multiple SCMs consistent with $P(S' \mid S, A)$ but with different *counterfactual* distributions

For **binary variables**, assuming the property of **monotonicity** (Pearl, 2000) is sufficient to identify the counterfactual distribution

But most real-world MDPs have non-binary states!

So, what should we use for the structural causal model (SCM)?

Key challenge: Non-identifiability

There are multiple SCMs consistent with $P(S' \mid S, A)$ but with different *counterfactual* distributions

For **binary variables**, assuming the property of **monotonicity** (Pearl, 2000) is sufficient to identify the counterfactual distribution

But most real-world MDPs have non-binary states!

Theorem 1 (informal): (Newly defined) property of **counterfactual stability** generalizes monotonicity to categorical variables

So, what should we use for the structural causal model (SCM)?

Key challenge: Non-identifiability

There are multiple SCMs consistent with $P(S' \mid S, A)$ but with different *counterfactual* distributions

For **binary variables**, assuming the property of **monotonicity** (Pearl, 2000) is sufficient to identify the counterfactual distribution

But most real-world MDPs have non-binary states!

Theorem 1 (informal): (Newly defined) property of **counterfactual stability** generalizes monotonicity to categorical variables

Gumbel-Max SCM

Use the *Gumbel-Max trick* to sample from a categorical distribution with k categories:

$$g_j \sim Gumbel$$
$$S' = argmax_j \{ \log P(S' = j \mid S, A) + g_j \}$$

Theorem 2: Gumbel-Max SCM satisfies the counterfactual stability condition

Thank you!

Come to our poster for more details: **Pacific Ballroom #72**