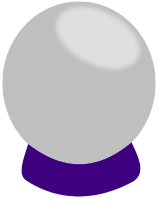
Formal Quality Measures for Predictors in Markov Decision Processes

Christel Baier, Sascha Klüppelholz, Jakob Piribauer, **Robin Ziemek**

Chair of Algebraic and Logical Foundations of Computer Science
Faculty of Computer Science, Technische Universität Dresden

AAAI Presentation, 1st March 2025



Motivation

- AI systems advance $\Rightarrow$ AI systems become more complicated

Motivation

- AI systems advance $\Rightarrow$ AI systems become more complicated

$\Rightarrow$ Modern AI systems can be very complex and incomprehensible

Motivation

- AI systems advance $\Rightarrow$ AI systems become more complicated

$\Rightarrow$ Modern AI systems can be very complex and incomprehensible

Motivation

- AI systems advance $\Rightarrow$ AI systems become more complicated

$\Rightarrow$ Modern AI systems can be very complex and incomprehensible

- There are many ways to model systems compactly

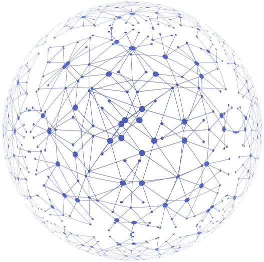
Motivation

- AI systems advance $\Rightarrow$ AI systems become more complicated

$\Rightarrow$ Modern AI systems can be very complex and incomprehensible

- There are many ways to model systems compactly

simplification $\longrightarrow$



Towards Perspicuity

Formal Verification

- Given a compact system model M

Towards Perspicuity

Formal Verification

- Given a compact system model M and a property specification φ

Towards Perspicuity

Formal Verification

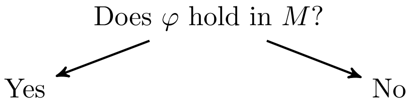
- Given a compact system model M and a property specification φ

$$\text{Does } \varphi \text{ hold in } M?$$

Towards Perspicuity

Formal Verification

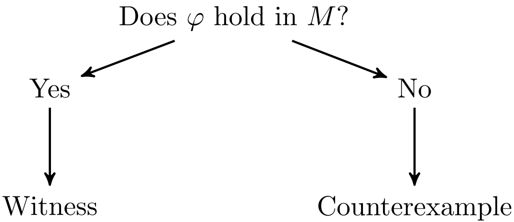
- Given a compact system model M and a property specification φ



Towards Perspicuity

Formal Verification

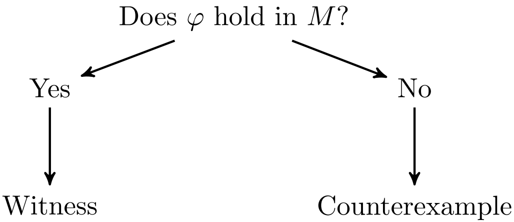
- Given a compact system model M and a property specification φ



Towards Perspicuity

Formal Verification

- Given a compact system model M and a property specification φ

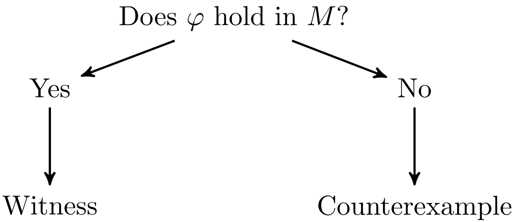


- Behavior can be ensured

Towards Perspicuity

Formal Verification

- Given a compact system model M and a property specification φ



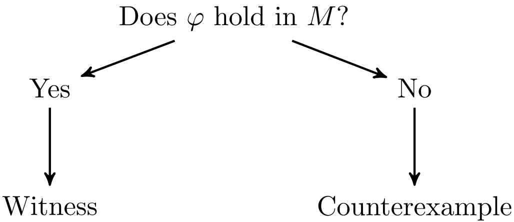
- Behavior can be ensured

Perspicuity

Towards Perspicuity

Formal Verification

- Given a compact system model M and a property specification φ



- Behavior can be ensured

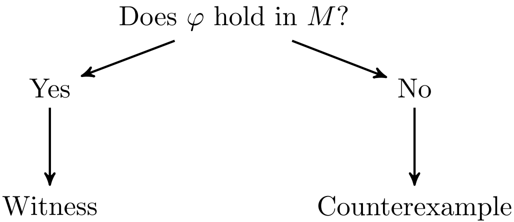
Perspicuity

- Understanding the system

Towards Perspicuity

Formal Verification

- Given a compact system model M and a property specification φ



- Behavior can be ensured

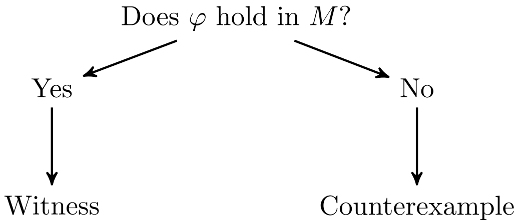
Perspicuity

- Understanding the system
- Compact and comprehensive

Towards Perspicuity

Formal Verification

- Given a compact system model M and a property specification φ



- Behavior can be ensured

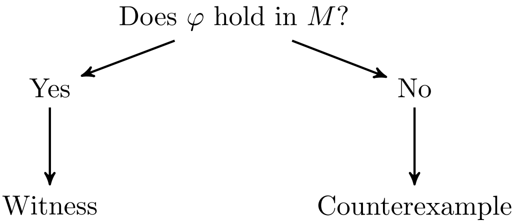
Perspicuity

- Understanding the system
- Compact and comprehensive
- Inherently difficult to find especially for larger systems

Towards Perspicuity

Formal Verification

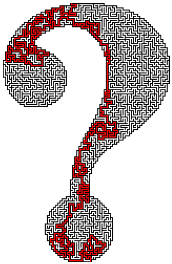
- Given a compact system model M and a property specification φ



- Behavior can be ensured

Perspicuity

- Understanding the system
- Compact and comprehensive
- Inherently difficult to find especially for larger systems



Systems with Probabilistic Behavior and Uncertainty

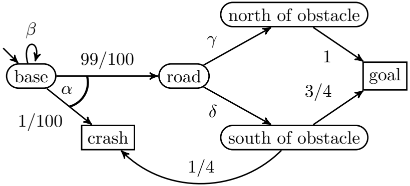
- Markov decision processes (MDPs) model systems with probabilism and uncertainty

Systems with Probabilistic Behavior and Uncertainty

- Markov decision processes (MDPs) model systems with probabilism and uncertainty

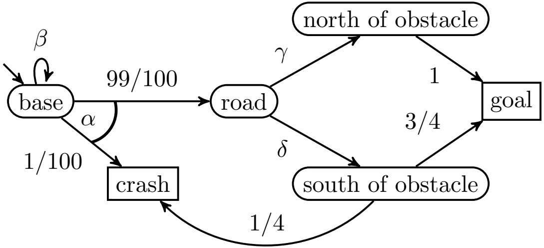
Systems with Probabilistic Behavior and Uncertainty

- Markov decision processes (MDPs) model systems with probabilism and uncertainty



Markov Decision Process

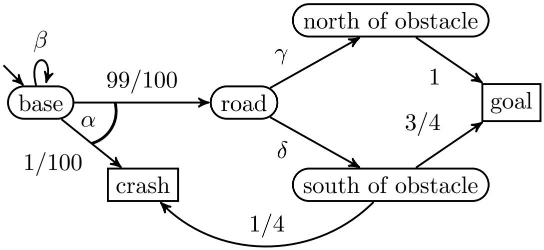
A *Markov decision process* (MDP) $\mathcal{M} = (\mathcal{S}, Act, \mathbf{P}, s_0)$ consists of



Markov Decision Process

A *Markov decision process* (MDP) $\mathcal{M} = (S, Act, \mathbf{P}, s_0)$ consists of

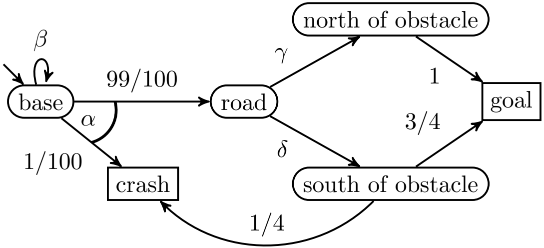
- a finite set of states S,



Markov Decision Process

A *Markov decision process* (MDP) $\mathcal{M} = (S, Act, \mathbf{P}, s_0)$ consists of

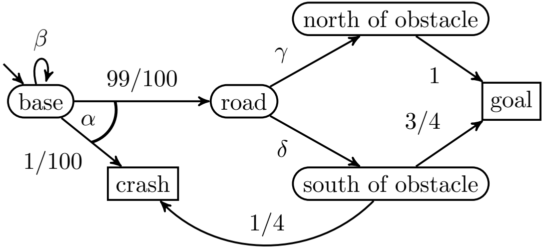
- a finite set of states S,
- a finite set of actions Act,



Markov Decision Process

A *Markov decision process* (MDP) $\mathcal{M} = (S, Act, \mathbf{P}, s_0)$ consists of

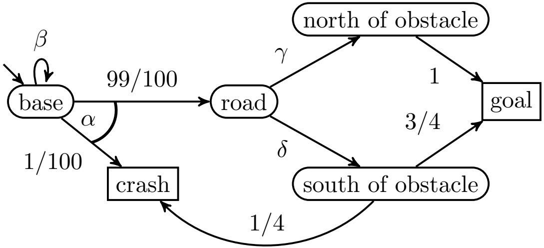
- a finite set of states S,
- a finite set of actions Act,
- an initial state s_0,



Markov Decision Process

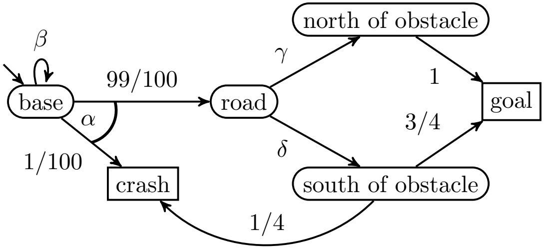
A *Markov decision process* (MDP) $\mathcal{M} = (S, Act, \mathbf{P}, s_0)$ consists of

- a finite set of states S,
- a finite set of actions Act,
- an initial state s_0,
- and a probabilistic transition function $\mathbf{P}\colon S \times Act \times S \to [0, 1]$.



Markov Decision Process – Policy

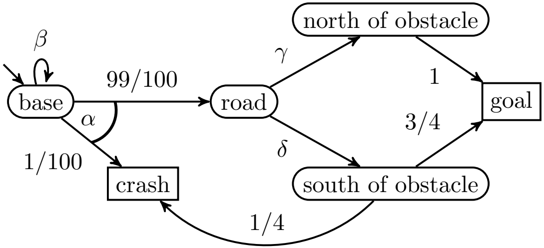
A *policy* of an MDP $\mathcal{M}$ is a function $\mathfrak{S}$ with



Markov Decision Process – Policy

A *policy* of an MDP $\mathcal{M}$ is a function $\mathfrak{S}$ with

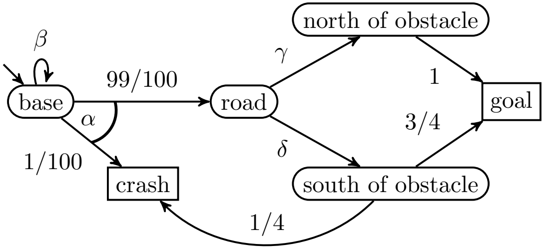
- Input: finite path in $\mathcal{M}$



Markov Decision Process – Policy

A *policy* of an MDP $\mathcal{M}$ is a function $\mathfrak{S}$ with

- Input: finite path in $\mathcal{M}$
- Output: distribution over enabled actions

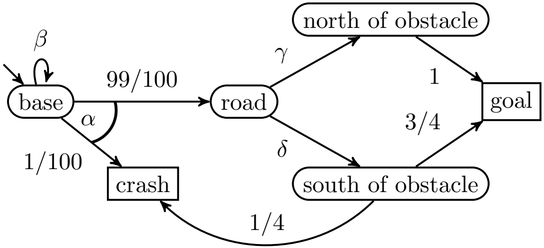


Markov Decision Process – Policy

A *policy* of an MDP $\mathcal{M}$ is a function $\mathfrak{S}$ with

- Input: finite path in $\mathcal{M}$
- Output: distribution over enabled actions

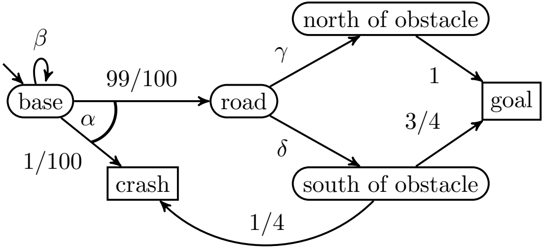
The purely probabilistic model is a *discrete time Markov chain* (DTMC)



Towards Perspicuity in MDPs

We consider general predictors in MDPs:

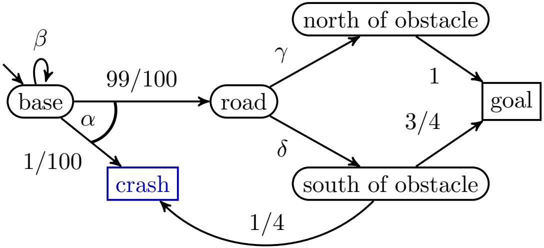
Given an MDP $\mathcal{M}$,



Towards Perspicuity in MDPs

We consider general predictors in MDPs:

Given an MDP $\mathcal{M}$, and a reachability event of interest $\Diamond E$,

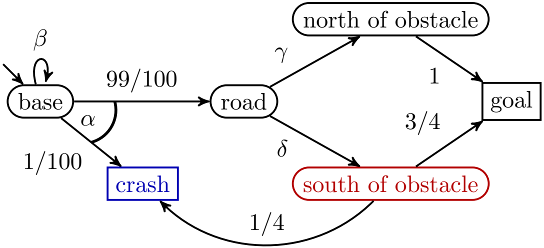


Towards Perspicuity in MDPs

We consider general predictors in MDPs:

Given an MDP $\mathcal{M}$, and a reachability event of interest $\Diamond E$,

how *good* does another reachability event $\Diamond C$ predict $\Diamond E$?



Confusion Matrix

Consider the confusion matrix for an event E and a predictor C

Confusion Matrix

Consider the confusion matrix for an event E and a predictor C

$\wedge$	$\Diamond E$	$\neg\Diamond E$
$\Diamond C$	$\Diamond C \wedge \Diamond E$ = tp true positive	$\Diamond C \wedge \neg\Diamond E$ = fp false positive
$\neg\Diamond C$	$\neg\Diamond C \wedge \Diamond E$ = fn false negative	$\neg\Diamond C \wedge \neg\Diamond E$ = tn true negative

Confusion Matrix

Consider the confusion matrix for an event E and a predictor C together with a policy $\mathfrak{S}$

$\wedge$	$\Diamond E$	$\neg\Diamond E$
$\Diamond C$	$\Pr^{\mathfrak{S}}(\Diamond C \wedge \Diamond E) = \mathrm{tp}^{\mathfrak{S}}$ true positive	$\Pr^{\mathfrak{S}}(\Diamond C \wedge \neg\Diamond E) = \mathrm{fp}^{\mathfrak{S}}$ false positive
$\neg\Diamond C$	$\Pr^{\mathfrak{S}}(\neg\Diamond C \wedge \Diamond E) = \mathrm{fn}^{\mathfrak{S}}$ false negative	$\Pr^{\mathfrak{S}}(\neg\Diamond C \wedge \neg\Diamond E) = \mathrm{tn}^{\mathfrak{S}}$ true negative

Confusion Matrix

Consider the confusion matrix for an event E and a predictor C together with a policy $\mathfrak{S}$

$\wedge$	$\Diamond E$	$\neg\Diamond E$
$\Diamond C$	$\Pr^{\mathfrak{S}}(\Diamond C \wedge \Diamond E) = \mathsf{tp}^{\mathfrak{S}}$ true positive	$\Pr^{\mathfrak{S}}(\Diamond C \wedge \neg\Diamond E) = \mathsf{fp}^{\mathfrak{S}}$ false positive
$\neg\Diamond C$	$\Pr^{\mathfrak{S}}(\neg\Diamond C \wedge \Diamond E) = \mathsf{fn}^{\mathfrak{S}}$ false negative	$\Pr^{\mathfrak{S}}(\neg\Diamond C \wedge \neg\Diamond E) = \mathsf{tn}^{\mathfrak{S}}$ true negative

- Measures from statistical analysis can rate the quality of a predictor

Quality Measures

Given MDP $\mathcal{M}$, event $\lozenge E$, predictor $\lozenge C$ and policy $\mathfrak{S}$,

Quality Measures

Given MDP $\mathcal{M}$, event $\Diamond E$, predictor $\Diamond C$ and policy $\mathfrak{S}$, define e.g.

$$\mathrm{recall}^{\mathfrak{S}}(C) = \frac{\mathrm{tp}^{\mathfrak{S}}}{\mathrm{tp}^{\mathfrak{S}}+\mathrm{fn}^{\mathfrak{S}}} \qquad \mathrm{fscore}^{\mathfrak{S}}(C) = \frac{2\cdot\mathrm{tp}^{\mathfrak{S}}}{2\cdot\mathrm{tp}^{\mathfrak{S}}+\mathrm{fp}^{\mathfrak{S}}+\mathrm{fn}^{\mathfrak{S}}}$$

$$\mathrm{mcc}^{\mathfrak{S}}(C) = \frac{\mathrm{tp}^{\mathfrak{S}}\cdot\mathrm{tn}^{\mathfrak{S}}-\mathrm{fp}^{\mathfrak{S}}\cdot\mathrm{fn}^{\mathfrak{S}}}{\sqrt{(\mathrm{tp}^{\mathfrak{S}}+\mathrm{fp}^{\mathfrak{S}})\cdot(\mathrm{tp}^{\mathfrak{S}}+\mathrm{fn}^{\mathfrak{S}})\cdot(\mathrm{tn}^{\mathfrak{S}}+\mathrm{fp}^{\mathfrak{S}})\cdot(\mathrm{tn}^{\mathfrak{S}}+\mathrm{fn}^{\mathfrak{S}})}}$$

Quality Measures

Given MDP $\mathcal{M}$, event $\Diamond E$, predictor $\Diamond C$ and policy $\mathfrak{S}$, define e.g.

$$\mathsf{recall}^{\mathfrak{S}}(C) = \frac{\mathsf{tp}^{\mathfrak{S}}}{\mathsf{tp}^{\mathfrak{S}}+\mathsf{fn}^{\mathfrak{S}}} \qquad \mathsf{fscore}^{\mathfrak{S}}(C) = \frac{2\cdot\mathsf{tp}^{\mathfrak{S}}}{2\cdot\mathsf{tp}^{\mathfrak{S}}+\mathsf{fp}^{\mathfrak{S}}+\mathsf{fn}^{\mathfrak{S}}}$$

$$\mathsf{mcc}^{\mathfrak{S}}(C) = \frac{\mathsf{tp}^{\mathfrak{S}}\cdot\mathsf{tn}^{\mathfrak{S}}-\mathsf{fp}^{\mathfrak{S}}\cdot\mathsf{fn}^{\mathfrak{S}}}{\sqrt{(\mathsf{tp}^{\mathfrak{S}}+\mathsf{fp}^{\mathfrak{S}})\cdot(\mathsf{tp}^{\mathfrak{S}}+\mathsf{fn}^{\mathfrak{S}})\cdot(\mathsf{tn}^{\mathfrak{S}}+\mathsf{fp}^{\mathfrak{S}})\cdot(\mathsf{tn}^{\mathfrak{S}}+\mathsf{fn}^{\mathfrak{S}})}}$$

- Best/Worst case policies computable

Quality Measures

Given MDP $\mathcal{M}$, event $\Diamond E$, predictor $\Diamond C$ and policy $\mathfrak{S}$, define e.g.

$$\mathrm{recall}^{\mathfrak{S}}(C) = \frac{\mathrm{tp}^{\mathfrak{S}}}{\mathrm{tp}^{\mathfrak{S}}+\mathrm{fn}^{\mathfrak{S}}} \qquad \mathrm{fscore}^{\mathfrak{S}}(C) = \frac{2\cdot\mathrm{tp}^{\mathfrak{S}}}{2\cdot\mathrm{tp}^{\mathfrak{S}}+\mathrm{fp}^{\mathfrak{S}}+\mathrm{fn}^{\mathfrak{S}}}$$

$$\mathrm{mcc}^{\mathfrak{S}}(C) = \frac{\mathrm{tp}^{\mathfrak{S}}\cdot\mathrm{tn}^{\mathfrak{S}}-\mathrm{fp}^{\mathfrak{S}}\cdot\mathrm{fn}^{\mathfrak{S}}}{\sqrt{(\mathrm{tp}^{\mathfrak{S}}+\mathrm{fp}^{\mathfrak{S}})\cdot(\mathrm{tp}^{\mathfrak{S}}+\mathrm{fn}^{\mathfrak{S}})\cdot(\mathrm{tn}^{\mathfrak{S}}+\mathrm{fp}^{\mathfrak{S}})\cdot(\mathrm{tn}^{\mathfrak{S}}+\mathrm{fn}^{\mathfrak{S}})}}$$

- Best/Worst case policies computable *but* are corner cases

Quality Measures

Given MDP $\mathcal{M}$, event $\lozenge E$, predictor $\lozenge C$ and policy $\mathfrak{S}$, define e.g.

$$\mathrm{recall}^{\mathfrak{S}}(C) = \frac{\mathrm{tp}^{\mathfrak{S}}}{\mathrm{tp}^{\mathfrak{S}} + \mathrm{fn}^{\mathfrak{S}}} \qquad \mathrm{fscore}^{\mathfrak{S}}(C) = \frac{2 \cdot \mathrm{tp}^{\mathfrak{S}}}{2 \cdot \mathrm{tp}^{\mathfrak{S}} + \mathrm{fp}^{\mathfrak{S}} + \mathrm{fn}^{\mathfrak{S}}}$$

$$\mathrm{mcc}^{\mathfrak{S}}(C) = \frac{\mathrm{tp}^{\mathfrak{S}} \cdot \mathrm{tn}^{\mathfrak{S}} - \mathrm{fp}^{\mathfrak{S}} \cdot \mathrm{fn}^{\mathfrak{S}}}{\sqrt{(\mathrm{tp}^{\mathfrak{S}} + \mathrm{fp}^{\mathfrak{S}}) \cdot (\mathrm{tp}^{\mathfrak{S}} + \mathrm{fn}^{\mathfrak{S}}) \cdot (\mathrm{tn}^{\mathfrak{S}} + \mathrm{fp}^{\mathfrak{S}}) \cdot (\mathrm{tn}^{\mathfrak{S}} + \mathrm{fn}^{\mathfrak{S}})}}$$

- Best/Worst case policies computable *but* are corner cases
- Average quality measures for all possible (memoryless) policies e.g.

$$\int_{\mathfrak{P}} \mathrm{fscore}^{x}(C)\, \mathrm{d}x \;/\; \int_{\mathfrak{P}} 1\, \mathrm{d}x,$$

Quality Measures

Given MDP $\mathcal{M}$, event $\lozenge E$, predictor $\lozenge C$ and policy $\mathfrak{S}$, define e.g.

$$\text{recall}^{\mathfrak{S}}(C) = \frac{\text{tp}^{\mathfrak{S}}}{\text{tp}^{\mathfrak{S}} + \text{fn}^{\mathfrak{S}}} \qquad \text{fscore}^{\mathfrak{S}}(C) = \frac{2 \cdot \text{tp}^{\mathfrak{S}}}{2 \cdot \text{tp}^{\mathfrak{S}} + \text{fp}^{\mathfrak{S}} + \text{fn}^{\mathfrak{S}}}$$

$$\text{mcc}^{\mathfrak{S}}(C) = \frac{\text{tp}^{\mathfrak{S}} \cdot \text{tn}^{\mathfrak{S}} - \text{fp}^{\mathfrak{S}} \cdot \text{fn}^{\mathfrak{S}}}{\sqrt{(\text{tp}^{\mathfrak{S}} + \text{fp}^{\mathfrak{S}}) \cdot (\text{tp}^{\mathfrak{S}} + \text{fn}^{\mathfrak{S}}) \cdot (\text{tn}^{\mathfrak{S}} + \text{fp}^{\mathfrak{S}}) \cdot (\text{tn}^{\mathfrak{S}} + \text{fn}^{\mathfrak{S}})}}$$

- Best/Worst case policies computable *but* are corner cases
- Average quality measures for all possible (memoryless) policies e.g.

$$\int_{\mathfrak{P}} \text{fscore}^{x}(C)\, dx \;/\; \int_{\mathfrak{P}} 1\, dx,$$

where $\mathfrak{P} = \{x \in [0,1]^{StAct} \mid \sum_{\alpha \in Act(s)} x_{s,\alpha} = 1 \text{ for all } s \in S \setminus T\}$

Quality Measures

Given MDP $\mathcal{M}$, event $\Diamond E$, predictor $\Diamond C$ and policy $\mathfrak{S}$, define e.g.

$$\mathrm{recall}^{\mathfrak{S}}(C) = \frac{\mathrm{tp}^{\mathfrak{S}}}{\mathrm{tp}^{\mathfrak{S}}+\mathrm{fn}^{\mathfrak{S}}} \qquad \mathrm{fscore}^{\mathfrak{S}}(C) = \frac{2\cdot\mathrm{tp}^{\mathfrak{S}}}{2\cdot\mathrm{tp}^{\mathfrak{S}}+\mathrm{fp}^{\mathfrak{S}}+\mathrm{fn}^{\mathfrak{S}}}$$

$$\mathrm{mcc}^{\mathfrak{S}}(C) = \frac{\mathrm{tp}^{\mathfrak{S}}\cdot\mathrm{tn}^{\mathfrak{S}}-\mathrm{fp}^{\mathfrak{S}}\cdot\mathrm{fn}^{\mathfrak{S}}}{\sqrt{(\mathrm{tp}^{\mathfrak{S}}+\mathrm{fp}^{\mathfrak{S}})\cdot(\mathrm{tp}^{\mathfrak{S}}+\mathrm{fn}^{\mathfrak{S}})\cdot(\mathrm{tn}^{\mathfrak{S}}+\mathrm{fp}^{\mathfrak{S}})\cdot(\mathrm{tn}^{\mathfrak{S}}+\mathrm{fn}^{\mathfrak{S}})}}$$

- Best/Worst case policies computable *but* are corner cases
- Average quality measures for all possible (memoryless) policies e.g.

$$\int_{\mathfrak{P}} \mathrm{fscore}^{x}(C)\,dx \;/\; \int_{\mathfrak{P}} 1\,dx,$$

where $\mathfrak{P} = \{x \in [0,1]^{StAct} \mid \sum_{\alpha\in Act(s)} x_{s,\alpha} = 1 \text{ for all } s \in S \setminus T\}$

- Exact computation very hard polytope volume #P-hard

Quality Measures

Given MDP $\mathcal{M}$, event $\lozenge E$, predictor $\lozenge C$ and policy $\mathfrak{S}$, define e.g.

$$\mathrm{recall}^{\mathfrak{S}}(C) = \frac{\mathrm{tp}^{\mathfrak{S}}}{\mathrm{tp}^{\mathfrak{S}}+\mathrm{fn}^{\mathfrak{S}}} \qquad \mathrm{fscore}^{\mathfrak{S}}(C) = \frac{2\cdot\mathrm{tp}^{\mathfrak{S}}}{2\cdot\mathrm{tp}^{\mathfrak{S}}+\mathrm{fp}^{\mathfrak{S}}+\mathrm{fn}^{\mathfrak{S}}}$$

$$\mathrm{mcc}^{\mathfrak{S}}(C) = \frac{\mathrm{tp}^{\mathfrak{S}}\cdot\mathrm{tn}^{\mathfrak{S}}-\mathrm{fp}^{\mathfrak{S}}\cdot\mathrm{fn}^{\mathfrak{S}}}{\sqrt{(\mathrm{tp}^{\mathfrak{S}}+\mathrm{fp}^{\mathfrak{S}})\cdot(\mathrm{tp}^{\mathfrak{S}}+\mathrm{fn}^{\mathfrak{S}})\cdot(\mathrm{tn}^{\mathfrak{S}}+\mathrm{fp}^{\mathfrak{S}})\cdot(\mathrm{tn}^{\mathfrak{S}}+\mathrm{fn}^{\mathfrak{S}})}}$$

- Best/Worst case policies computable *but* are corner cases
- Average quality measures for all possible (memoryless) policies e.g.

$$\int_{\mathfrak{P}} \mathrm{fscore}^{x}(C)\,\mathrm{d}x \;/\; \int_{\mathfrak{P}} 1\,\mathrm{d}x,$$

where $\mathfrak{P} = \{x \in [0,1]^{StAct} \mid \sum_{\alpha\in Act(s)} x_{s,\alpha} = 1 \text{ for all } s \in S \setminus T\}$

- Exact computation very hard polytope volume #P-hard
- Simple constraints/functions favor Monte Carlo approx.

Probability-Raising Policies and Causes

Causes naturally qualify as predictors

Probability-Raising Policies and Causes

Causes naturally qualify as predictors as they *raise* the probability of the considered effect

Probability-Raising Policies and Causes

Causes naturally qualify as predictors as they *raise* the probability of the considered effect

$$\mathbf{P}(\mathit{Effect} \mid \mathit{Cause}) > \mathbf{P}(\mathit{Effect})$$

Probability-Raising Policies and Causes

Causes naturally qualify as predictors as they *raise* the probability of the considered effect

$$\mathrm{Pr}^{\mathfrak{S}}(\Diamond \mathit{Effect} \mid \Diamond \mathit{Cause}) > \mathrm{Pr}^{\mathfrak{S}}(\Diamond \mathit{Effect})^{1}$$

[1] Christel Baier, Jakob Piribauer, and Robin Ziemek. "Foundations of probability-raising causality in Markov decision processes". In: *Logical Methods in Computer Science* Volume 20, Issue 1 (Jan. 2024). DOI: 10.46298/lmcs-20(1:4)2024

Probability-Raising Policies and Causes

Causes naturally qualify as predictors as they *raise* the probability of the considered effect

$$\Pr^{\mathfrak{S}}(\Diamond \mathit{Effect} \mid \Diamond \mathit{Cause}) > \Pr^{\mathfrak{S}}(\Diamond \mathit{Effect})^{1}$$

A policy $\mathfrak{S}$ is *(globally) probability-raising* if the above holds

[1] Christel Baier, Jakob Piribauer, and Robin Ziemek. "Foundations of probability-raising causality in Markov decision processes". In: *Logical Methods in Computer Science* Volume 20, Issue 1 (Jan. 2024). DOI: 10.46298/lmcs-20(1:4)2024

Causal Volumes

Causes naturally qualify as predictors as they *raise* the probability of the considered effect

$$\mathsf{Pr}^{\mathfrak{S}}(\Diamond \mathit{Effect} \mid \Diamond \mathit{Cause}) > \mathsf{Pr}^{\mathfrak{S}}(\Diamond \mathit{Effect})$$

A policy $\mathfrak{S}$ is *(globally) probability-raising* if the above holds

Causal Volumes

Causes naturally qualify as predictors as they *raise* the probability of the considered effect

$$\Pr^{\mathfrak{S}}(\Diamond \mathit{Effect} \mid \Diamond \mathit{Cause}) > \Pr^{\mathfrak{S}}(\Diamond \mathit{Effect})$$

A policy $\mathfrak{S}$ is *(globally) probability-raising* if the above holds

Idea: Measure the set

$$\mathfrak{P}_{PR} = \{x \in \mathfrak{P} \mid x \text{ constitutes a probability-raising policy}\}$$

Causal Volumes

Causes naturally qualify as predictors as they *raise* the probability of the considered effect

$$\mathsf{Pr}^{\mathfrak{S}}(\Diamond \mathit{Effect} \mid \Diamond \mathit{Cause}) > \mathsf{Pr}^{\mathfrak{S}}(\Diamond \mathit{Effect})$$

A policy $\mathfrak{S}$ is *(globally) probability-raising* if the above holds

Idea: Measure the set

$$\mathfrak{P}_{\mathrm{PR}} = \{x \in \mathfrak{P} \mid x \text{ constitutes a probability-raising policy}\}$$

The *causal volume* of C is defined as

$$\mathrm{V}_{\mathrm{PR}}(C) = \int_{\mathfrak{P}_{\mathrm{PR}}} 1\,\mathrm{d}x \;/\; \int_{\mathfrak{P}} 1\,\mathrm{d}x$$

Causal Volumes

Causes naturally qualify as predictors as they *raise* the probability of the considered effect

$$\Pr^{\mathfrak{S}}(\Diamond \mathit{Effect} \mid \Diamond \mathit{Cause}) > \Pr^{\mathfrak{S}}(\Diamond \mathit{Effect})$$

A policy $\mathfrak{S}$ is *(globally) probability-raising* if the above holds

Idea: Measure the set

$$\mathfrak{P}_{\mathrm{PR}} = \{x \in \mathfrak{P} \mid x \text{ constitutes a probability-raising policy}\}$$

The *causal volume* of C is defined as

$$\mathrm{V}_{\mathrm{PR}}(C) = \int_{\mathfrak{P}_{\mathrm{PR}}} 1 \,\mathrm{d}x \;/\; \int_{\mathfrak{P}} 1 \,\mathrm{d}x$$

- Exact computation very hard polytope volume #P-hard
- Simple constraints/functions favor Monte Carlo approx.

Conclusion

- Predicting events in AI systems helps *perspicuity*

Conclusion

- Predicting events in AI systems helps *perspicuity*

In order to leverage this our paper

- uses tools from formal verification to consider *predictors* in MDPs

Conclusion

- Predicting events in AI systems helps *perspicuity*

In order to leverage this our paper

- uses tools from formal verification to consider *predictors* in MDPs
- defines ways of rating the *quality* of such predictors

Conclusion

- Predicting events in AI systems helps *perspicuity*

In order to leverage this our paper

- uses tools from formal verification to consider *predictors* in MDPs
- defines ways of rating the *quality* of such predictors
- contributes to a growing toolbox for enabling *perspicuous systems*

Conclusion

- Predicting events in AI systems helps *perspicuity*

In order to leverage this our paper

- uses tools from formal verification to consider *predictors* in MDPs
- defines ways of rating the *quality* of such predictors
- contributes to a growing toolbox for enabling *perspicuous systems*

Thank You!