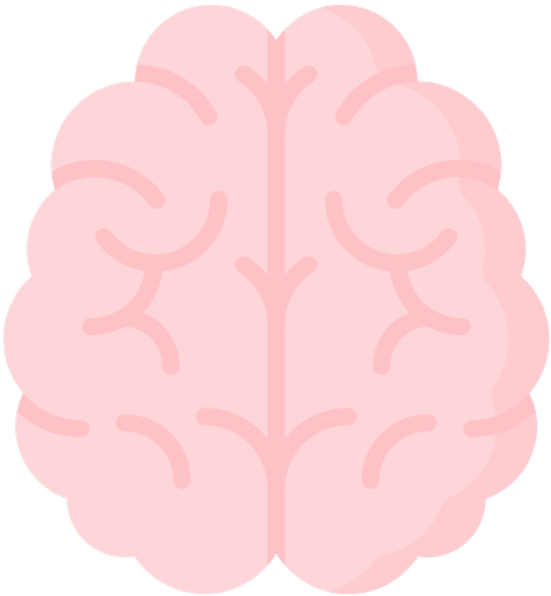
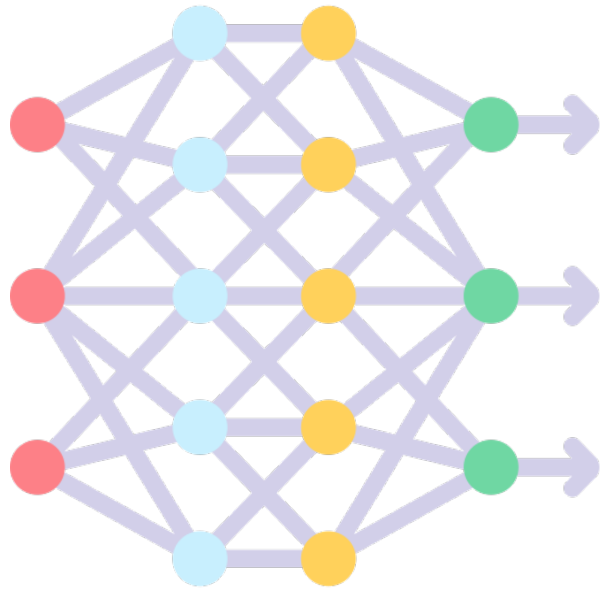
Outline



Cognitive science



Machine learning

Large language models

Causal machine learning

The amount of work at the interface of causality and machine learning, often referred to as **causal machine learning**, has been increasing very rapidly.

Kaddour et al. *"Causal machine learning: A survey and open problems."* arXiv preprint, 2022.
Peters et al. *"Elements of causal inference: foundations and learning algorithms."* The MIT Press, 2017.

Causal machine learning

The amount of work at the interface of causality and machine learning, often referred to as **causal machine learning**, has been increasing very rapidly.

Causal machine learning operationalizes causal (counterfactual) reasoning about

the **outputs** of machine learning models,
the **data** used by these models, and
the **users** of these models

using the theoretical framework of **structural causal models (SCMs).**

Kaddour et al. *"Causal machine learning: A survey and open problems."* arXiv preprint, 2022.
Peters et al. *"Elements of causal inference: foundations and learning algorithms."* The MIT Press, 2017.

Structural Causal Models (SCMs)

Given a set of random variables $\mathbf{X} = \{X_1, \ldots, X_n\}$, a SCM defines a **complete data-generating process** via a collection of assignments

$$X_i := f_i(\mathbf{PA}_i, U_i),$$

where $\mathbf{PA}_i \subseteq \mathbf{X} \backslash X_i$ are the direct causes of X_i,

$\mathbf{U} = \{U_1, \ldots, U_n\}$ are jointly independent noise variables

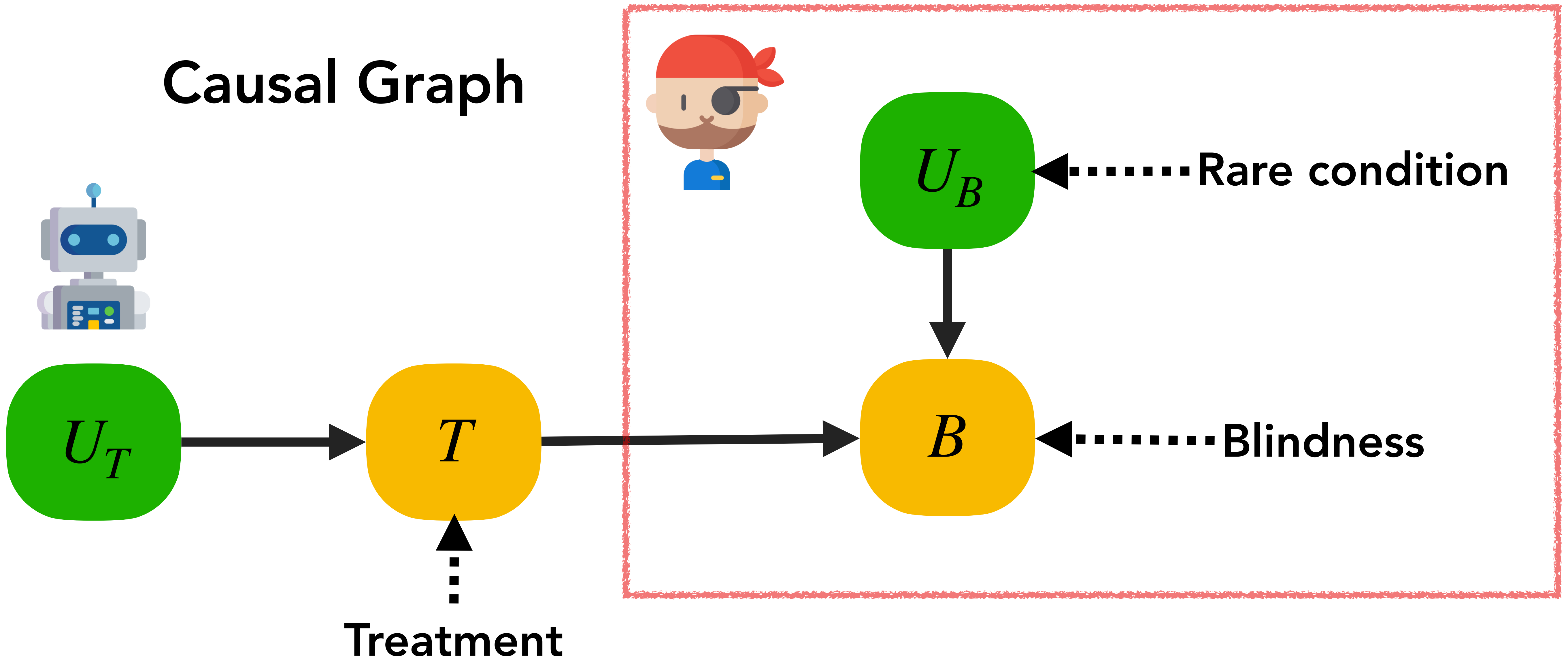
$\mathbf{F} = \{f_1, \ldots, f_n\}$ are deterministic causal mechanisms, and

$P(\mathbf{U})$ denotes the (prior) distribution of the noise variables.

Pearl. "*Causality.*" Cambridge university press, 2009.
Peters et al. *"Elements of causal inference: foundations and learning algorithms."* The MIT Press, 2017.

What kind of (causal) questions can we answer with SCMs?

(1) Observational, (2) Interventional and (3) Counterfactual Queries



Structural Causal Model $\mathscr{M}$

$$T := U_T$$

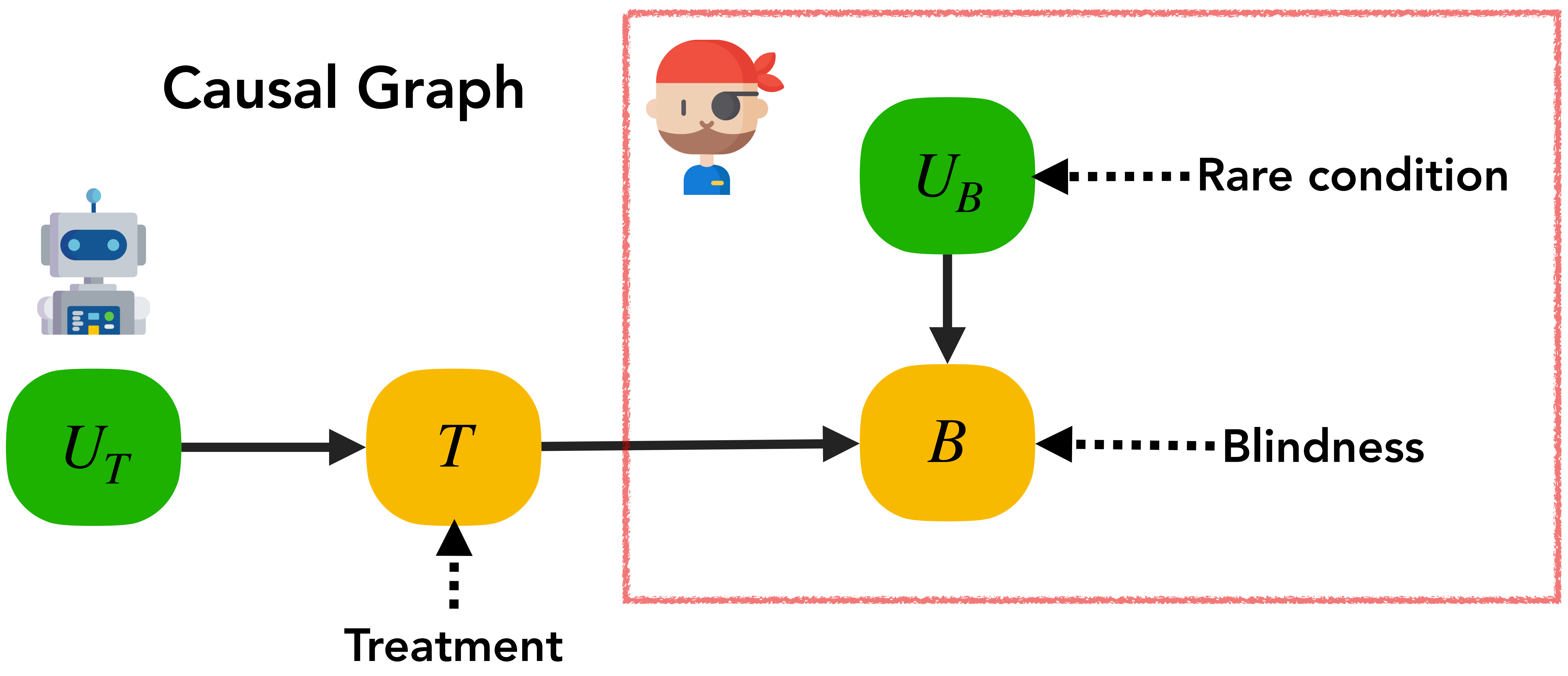
$$B := T \cdot U_B + (1 - T) \cdot (1 - U_B)$$

$$U_B \sim Ber(0.01), \quad U_T \sim Ber(0.5)$$

Example adapted from *Elements of causal inference, MIT Press, 2017*

What kind of (causal) questions can we answer with SCMs?

(1) **Observational**, (2) Interventional and (3) Counterfactual Queries



Structural Causal Model $\mathscr{M}$

$$T := U_T$$

$$B := T \cdot U_B + (1 - T) \cdot (1 - U_B)$$

$$U_B \sim Ber(0.01), \quad U_T \sim Ber(0.5)$$

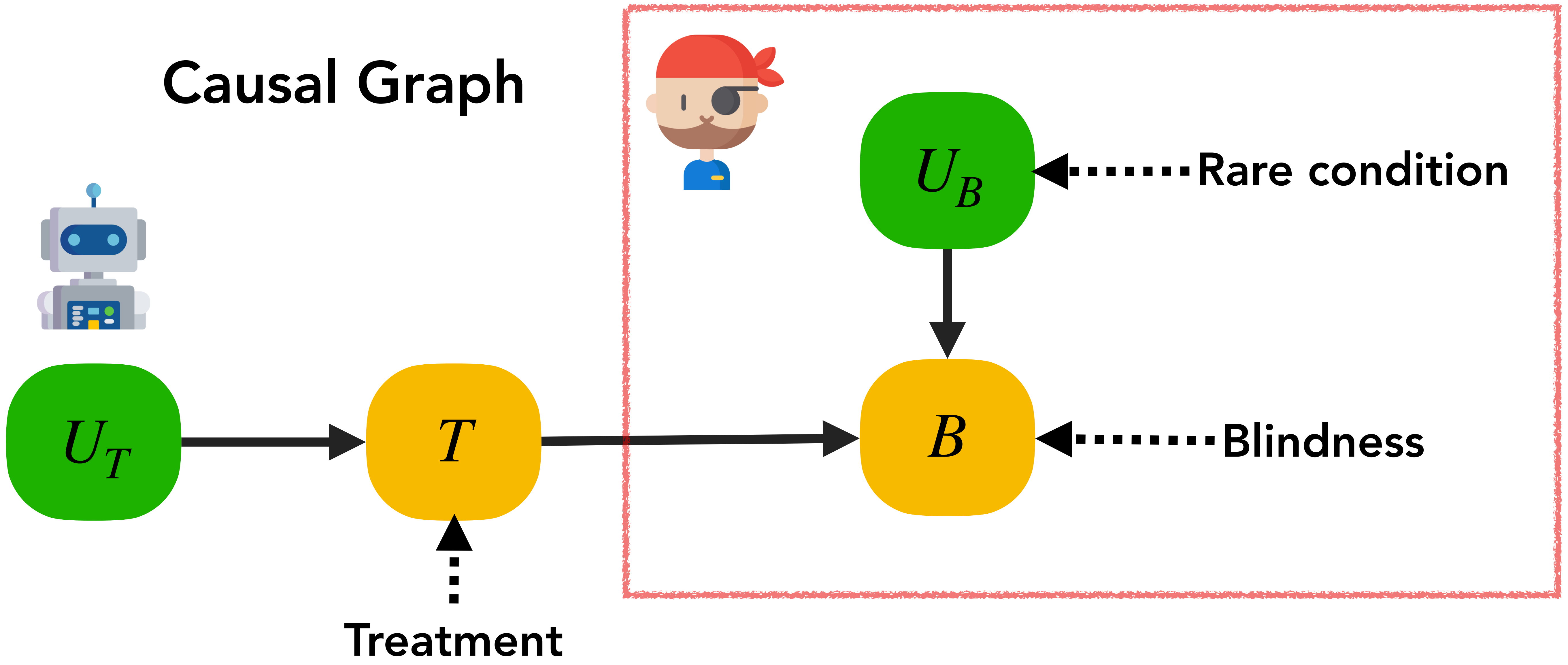
Observational question

What will happen to the patient?

Example adapted from *Elements of causal inference, MIT Press, 2017*

What kind of (causal) questions can we answer with SCMs?

(1) **Observational**, (2) Interventional and (3) Counterfactual Queries



Structural Causal Model $\mathscr{M}$

$$T := U_T$$

$$B := T \cdot U_B + (1 - T) \cdot (1 - U_B) \xrightarrow{\text{"observe"}}$$

$$U_B \sim Ber(0.01), \quad U_T \sim Ber(0.5)$$

Observational question

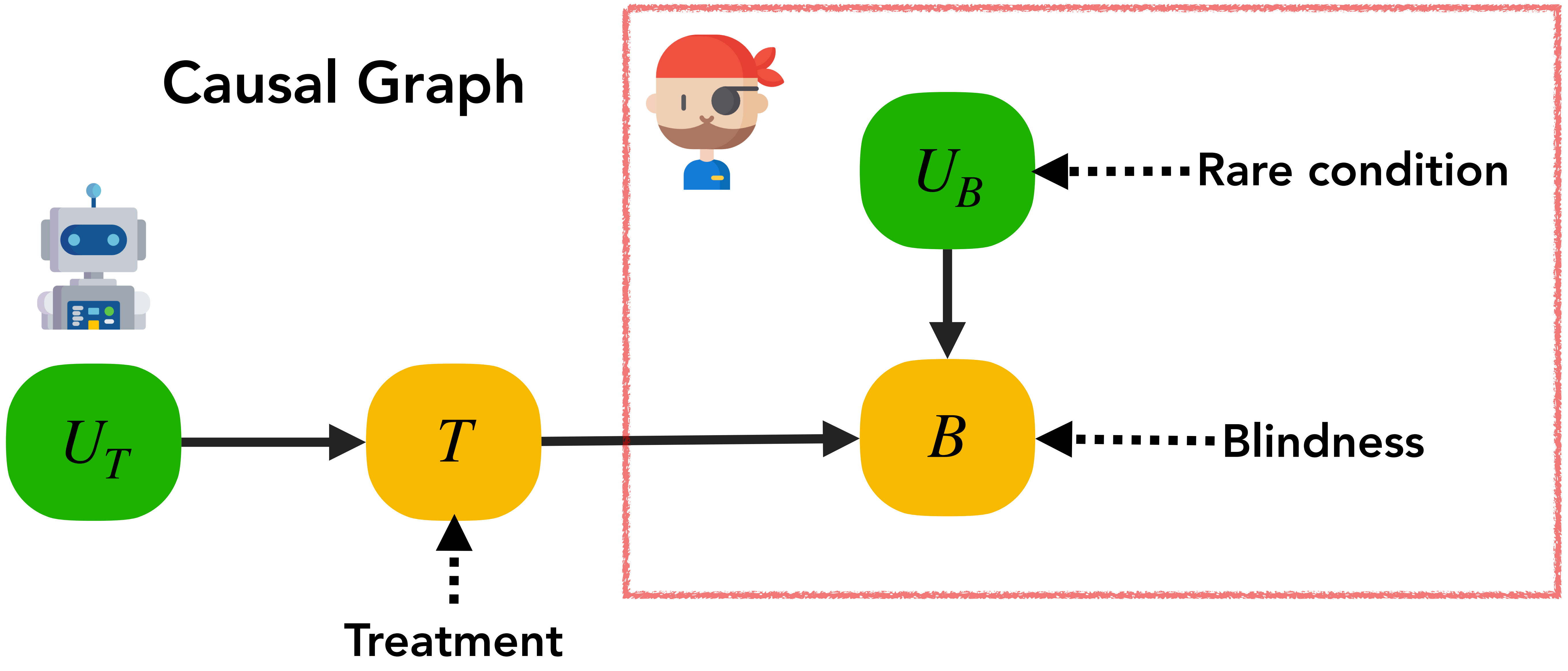
What will happen to the patient?

The patient will get blind ($B = 1$) with prob. 0.5

Example adapted from *Elements of causal inference, MIT Press, 2017*

What kind of (causal) questions can we answer with SCMs?

(1) **Observational**, (2) Interventional and (3) Counterfactual Queries



Structural Causal Model $\mathscr{M}$

$T := U_T$

$B := T \cdot U_B + (1 - T) \cdot (1 - U_B)$ $\xrightarrow{\text{"observe"}}$

$U_B \sim Ber(0.01), \quad U_T \sim Ber(0.5)$

Observational question

What will happen to the patient?

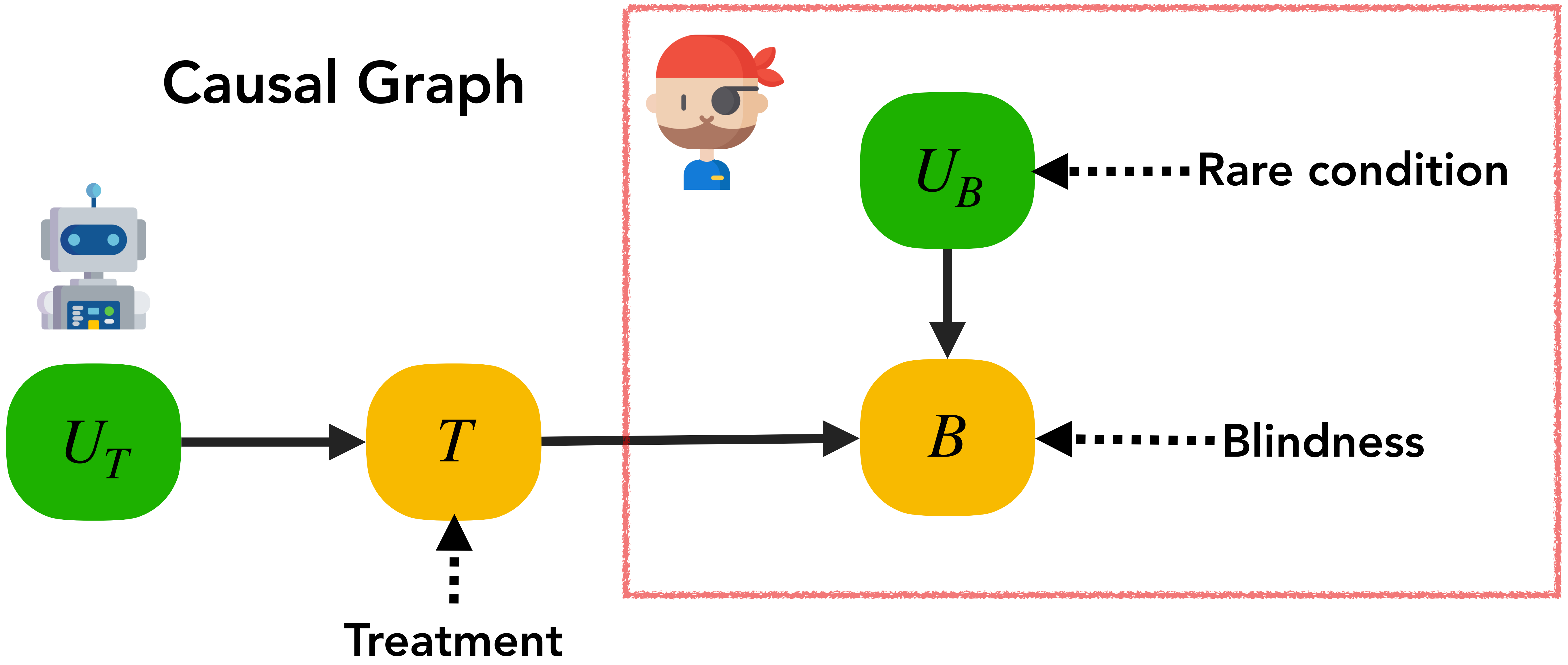
The patient will get blind ($B = 1$) with prob. 0.5

Formally, $P^{\mathscr{M}}(B = 1) = 0.5$

Example adapted from *Elements of causal inference, MIT Press, 2017*

What kind of (causal) questions can we answer with SCMs?

(1) Observational, (2) **Interventional** and (3) Counterfactual Queries



Structural Causal Model $\mathscr{M}$

$$T := U_T$$

$$B := T \cdot U_B + (1 - T) \cdot (1 - U_B)$$

$$U_B \sim Ber(0.01), \quad U_T \sim Ber(0.5)$$

Interventional question

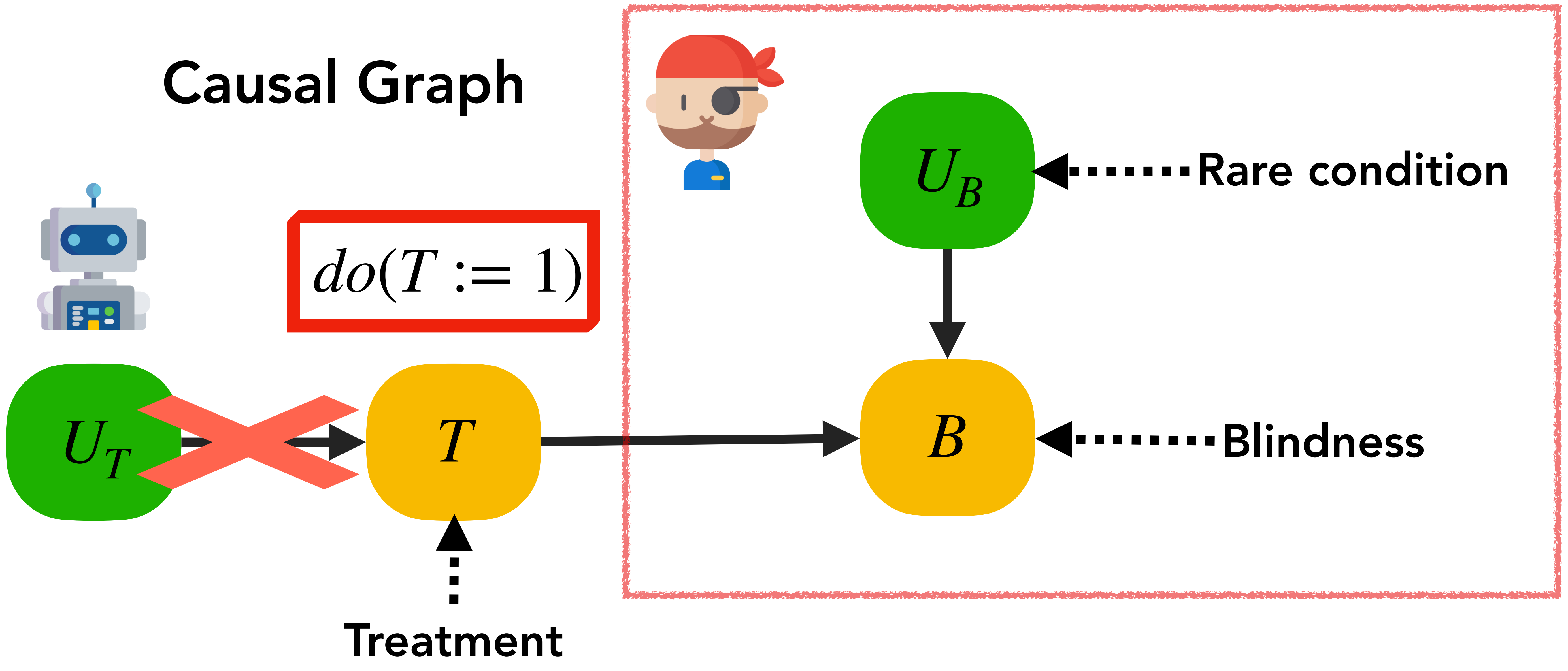
What will happen to the patient if a doctor breaks the robot and always administers the treatment?

Example adapted from *Elements of causal inference, MIT Press, 2017*

What kind of (causal) questions can we answer with SCMs?

(1) Observational, (2) **Interventional** and (3) Counterfactual Queries

Causal Graph



Structural Causal Model $\mathscr{M}$

$$\cancel{T := U_T} \quad T := 1$$

$$B := T \cdot U_B + (1 - T) \cdot (1 - U_B)$$

$$U_B \sim Ber(0.01), \quad \cancel{U_T \sim Ber(0.5)}$$

“do”

Interventional question

What will happen to the patient if a doctor breaks the robot and always administers the treatment?

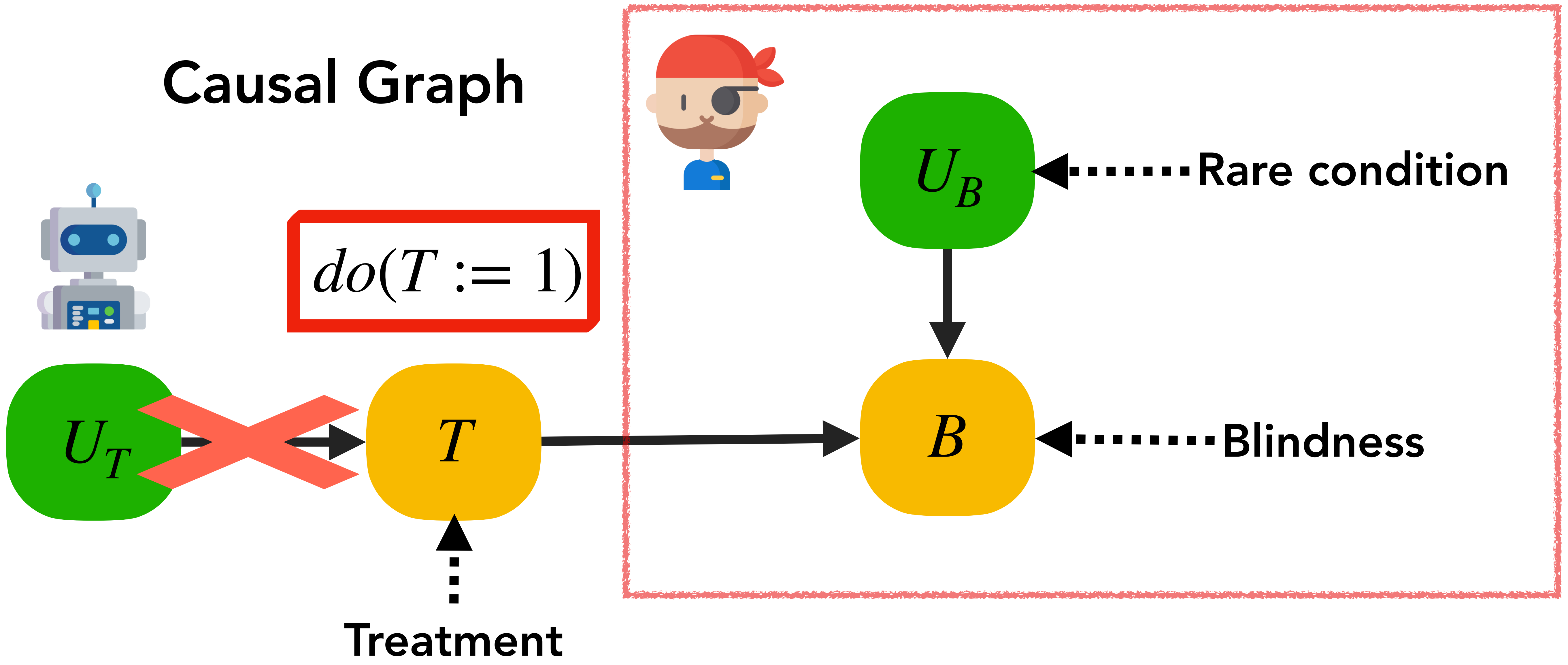
The patient will get blind ($B = 1$) with prob. 0.01

Example adapted from *Elements of causal inference, MIT Press, 2017*

What kind of (causal) questions can we answer with SCMs?

(1) Observational, (2) **Interventional** and (3) Counterfactual Queries

Causal Graph



Structural Causal Model $\mathscr{M}$

$\cancel{T := U_T}$ $T := 1$

$B := T \cdot U_B + (1 - T) \cdot (1 - U_B)$

$U_B \sim Ber(0.01)$, $\cancel{U_T \sim Ber(0.5)}$

“do”

Interventional question

What will happen to the patient if a doctor breaks the robot and always administers the treatment?

The patient will get blind ($B = 1$) with prob. 0.01

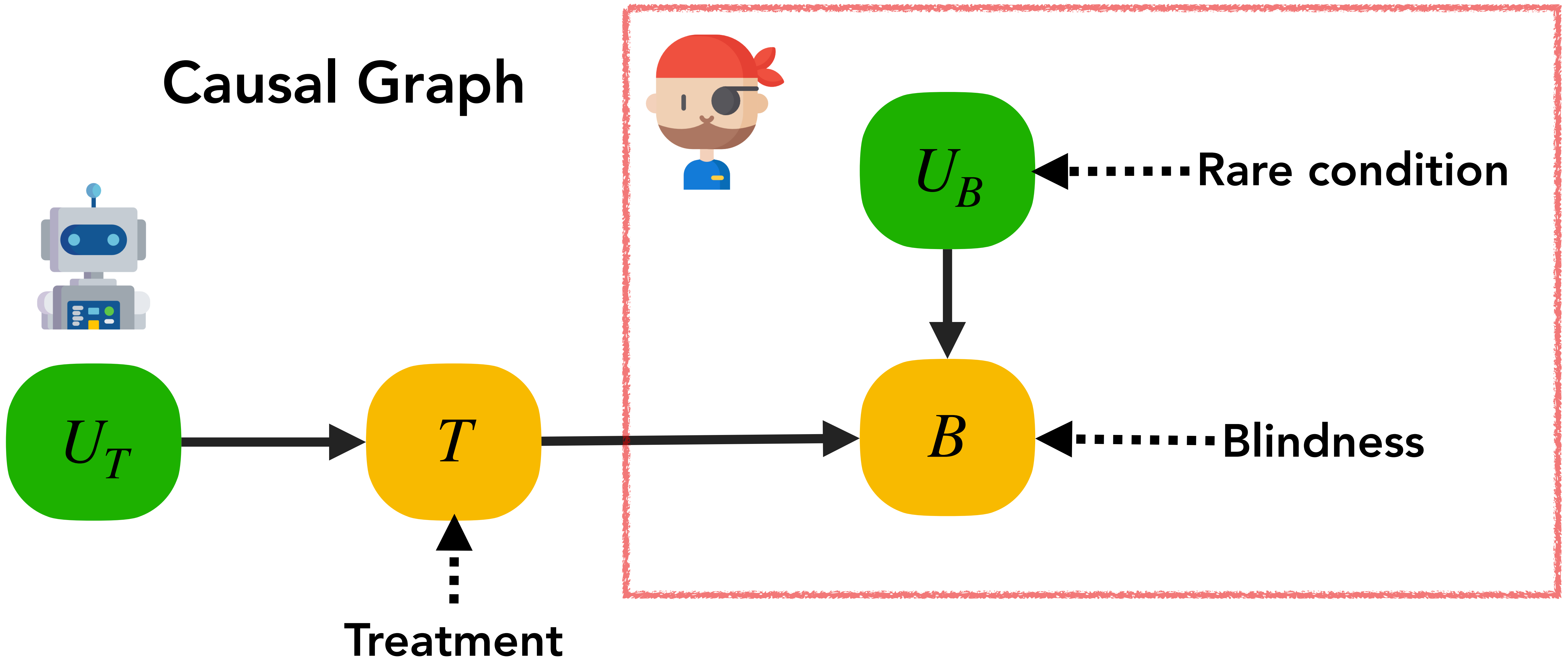
Formally, $\underline{P^{\mathscr{M}\,;\,do(T=1)}(B = 1) = 0.01}$

Example adapted from *Elements of causal inference, MIT Press, 2017*

What kind of (causal) questions can we answer with SCMs?

(1) Observational, (2) Interventional and (3) **Counterfactual** Queries

Causal Graph



Structural Causal Model $\mathscr{M}$

$$T := U_T$$

$$B := T \cdot U_B + (1 - T) \cdot (1 - U_B)$$

$$U_B \sim Ber(0.01), \quad U_T \sim Ber(0.5)$$

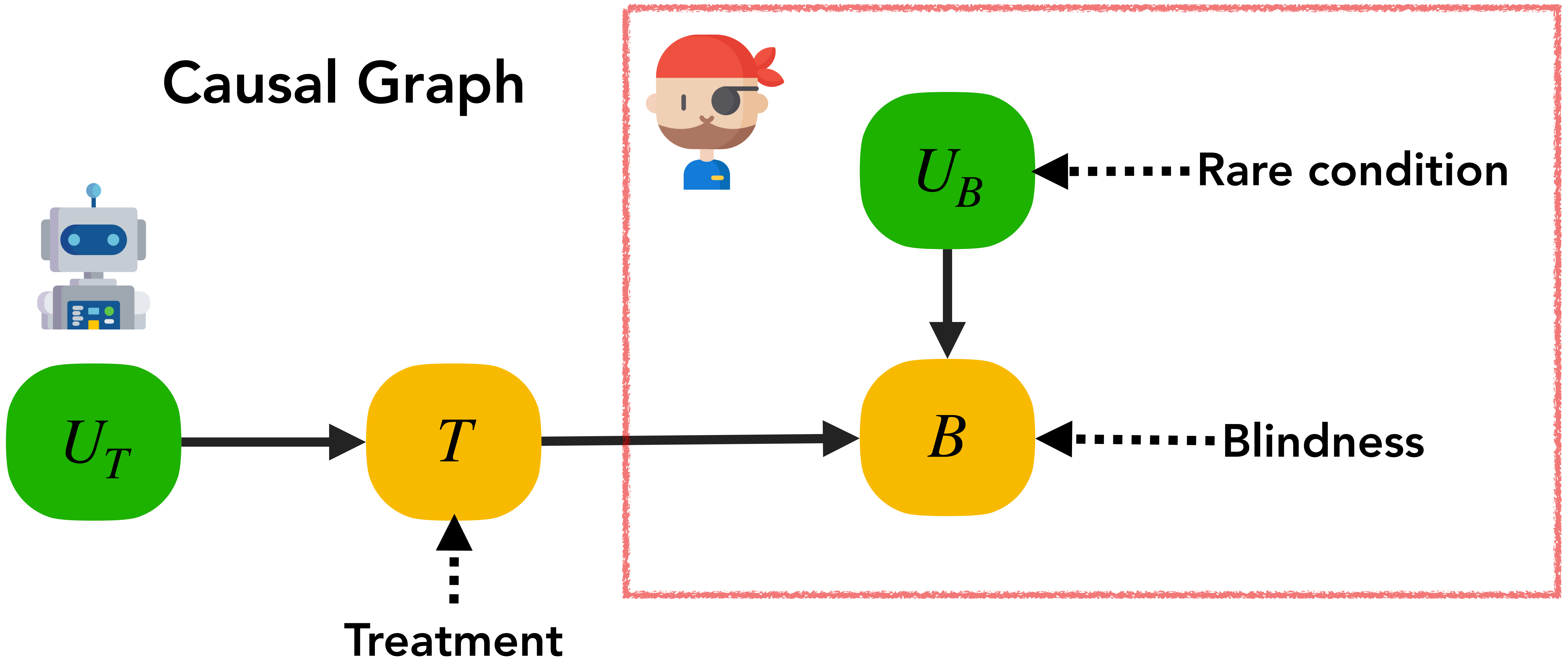
Counterfactual question

The treatment was administered and the patient got blind. What would have happened if the treatment had not been administered?

Example adapted from *Elements of causal inference, MIT Press, 2017*

What kind of (causal) questions can we answer with SCMs?

(1) Observational, (2) Interventional and (3) **Counterfactual** Queries



Modified Structural Causal Model $\mathscr{M}_{T=1,B=1}$

$T := 1$

$B := T$

$U_B = 1$ with prob. 1 ← **Posterior distribution of the noise**

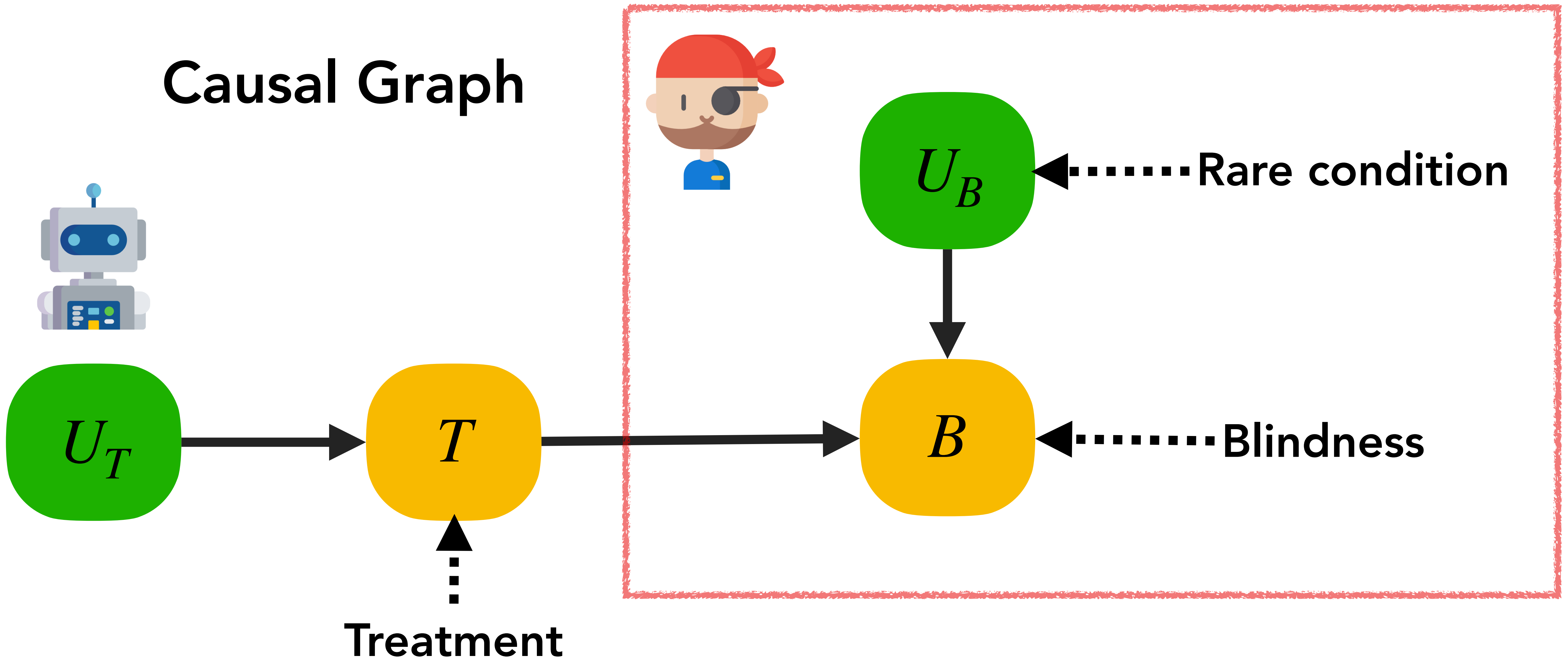
Counterfactual question

The treatment was administered and the patient got blind. What would have happened if the treatment had not been administered?

Example adapted from *Elements of causal inference, MIT Press, 2017*

What kind of (causal) questions can we answer with SCMs?

(1) Observational, (2) Interventional and (3) **Counterfactual** Queries



Modified Structural Causal Model $\mathscr{M}_{T=1,B=1}$

$\xcancel{T := 1}$ $T := 0$

$B := T$

$U_B = 1$ with prob. 1

"imagine"

Counterfactual question

The treatment was administered and the patient got blind. What would have happened if the treatment had not been administered?

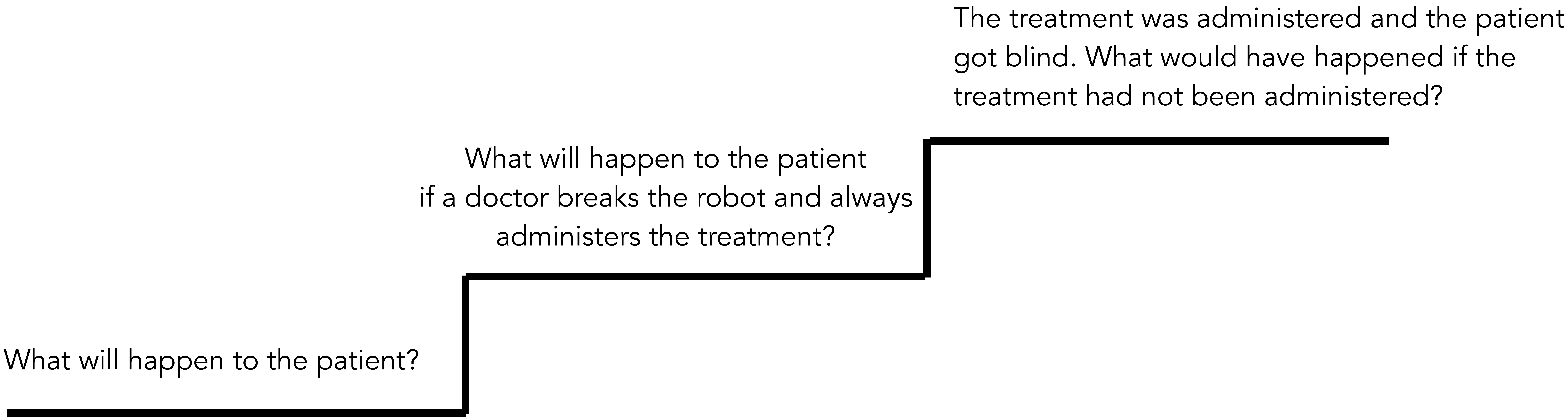
The patient would not have gotten blind ($B = 0$)

Formally, $P^{\mathscr{M}\,|\,T=1,B=1\,;\,do(T=1)}(B = 1) = 0$

Example adapted from *Elements of causal inference, MIT Press, 2017*

The ladder of causation

(1) Observational, (2) Interventional and (3) Counterfactual Queries



What will happen to the patient?

What will happen to the patient
if a doctor breaks the robot and always
administers the treatment?

The treatment was administered and the patient
got blind. What would have happened if the
treatment had not been administered?

It is called **ladder of causation** because questions at level $i \in \{1,2,3\}$ can only be answered if information from level $j \geq i$ is available. Counterfactuals sit at the top of the ladder!

Pearl. "*Causality.*" Cambridge university press, 2009.

Bareinboim et al. "*On Pearl's hierarchy and the foundations of causal inference.*" Probabilistic and causal inference: the works of Judea Pearl, 2022.

Identifiability

Identification of

an interventional probability, e.g., $P^{\mathscr{M}\,;\,do(T=1)}(B)$, or

a counterfactual probability, e.g., $P^{\mathscr{M}\,|\,T=1,\,B=1\,;\,do(T=1)}(B)$

refers to the process of estimating it using (observational) data from $\mathscr{M}$.

Shpitser and Pearl. "*Complete identification methods for the causal hierarchy.*" JMLR, 2008.
Perkovic et al. "*Complete graphical characterization and construction of adjustment sets in markov equivalence classes of ancestral graphs.*" JMLR, 2018.
Shalit et al. "*Estimating individual treatment effect: generalization bounds and algorithms.*" ICML, 2017.
Kallus. "*Treatment effect risk: Bounds and inference.*" Management Science, 2023.

Identifiability

Identification of

an interventional probability, e.g., $P^{\mathscr{M}\,;\,do(T=1)}(B)$, or

a counterfactual probability, e.g., $P^{\mathscr{M}\,|\,T=1,\,B=1\,;\,do(T=1)}(B)$

refers to the process of estimating it using (observational) data from $\mathscr{M}$.

If an interventional or counterfactual probability is not identifiable, then regardless of how much data we have, we will not be able to estimate it.

Shpitser and Pearl. "*Complete identification methods for the causal hierarchy.*" JMLR, 2008.
Perkovic et al. "*Complete graphical characterization and construction of adjustment sets in markov equivalence classes of ancestral graphs.*" JMLR, 2018.
Shalit et al. "*Estimating individual treatment effect: generalization bounds and algorithms.*" ICML, 2017.
Kallus. "*Treatment effect risk: Bounds and inference.*" Management Science, 2023.

Identifiability

Identification of

an interventional probability, e.g., $P^{\mathscr{M}\,;\,do(T=1)}(B)$, or

a counterfactual probability, e.g., $P^{\mathscr{M}\,|\,T=1,\,B=1\,;\,do(T=1)}(B)$

refers to the process of estimating it using (observational) data from $\mathscr{M}$.

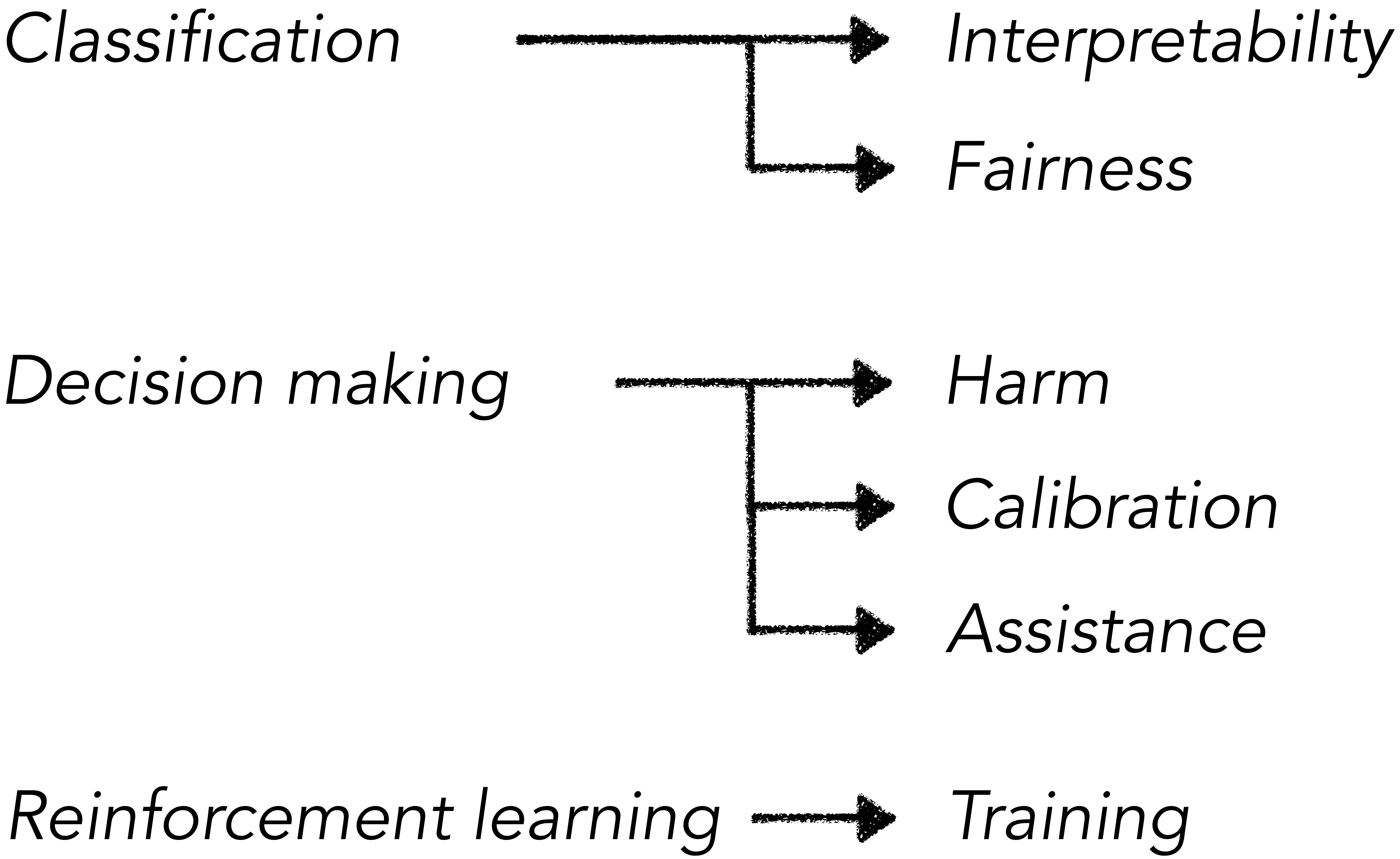
If an interventional or counterfactual probability is not identifiable, then regardless of how much data we have, we will not be able to estimate it.

There exist methods to

(i) determine the identifiability of specific interventional and counterfactual probabilities, and
(ii) estimate (or bound) quantities derived from these probabilities (e.g., individual treatment effects)

Shpitser and Pearl. "*Complete identification methods for the causal hierarchy.*" JMLR, 2008.
Perkovic et al. "*Complete graphical characterization and construction of adjustment sets in markov equivalence classes of ancestral graphs.*" JMLR, 2018.
Shalit et al. "*Estimating individual treatment effect: generalization bounds and algorithms.*" ICML, 2017.
Kallus. "*Treatment effect risk: Bounds and inference.*" Management Science, 2023.

Use cases of counterfactuals in machine learning



Classification

Interpretability

Fairness

Decision making

Harm

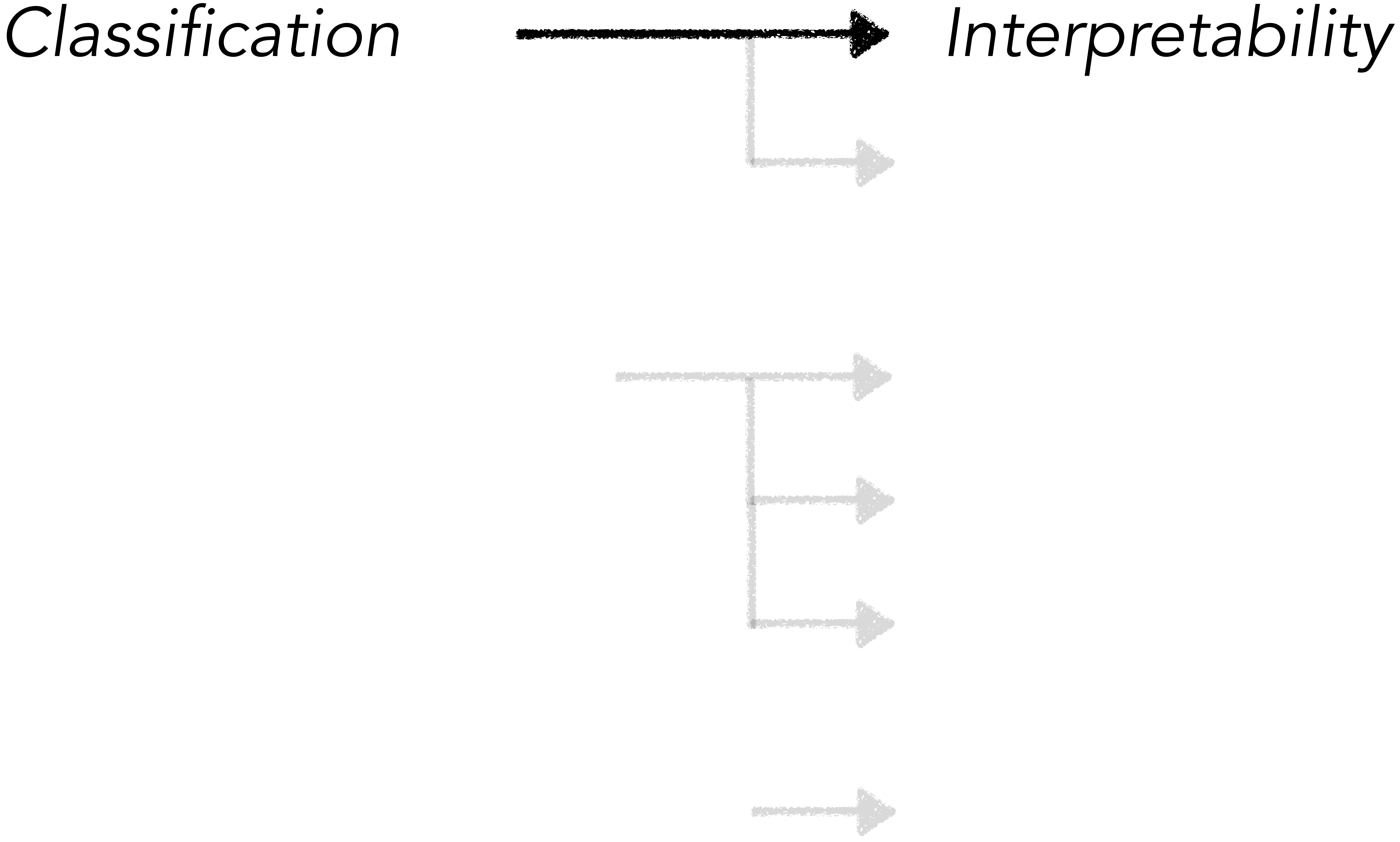
Calibration

Assistance

Reinforcement learning

Training

Use cases of counterfactuals in machine learning



Classification

Interpretability

Counterfactual explanations

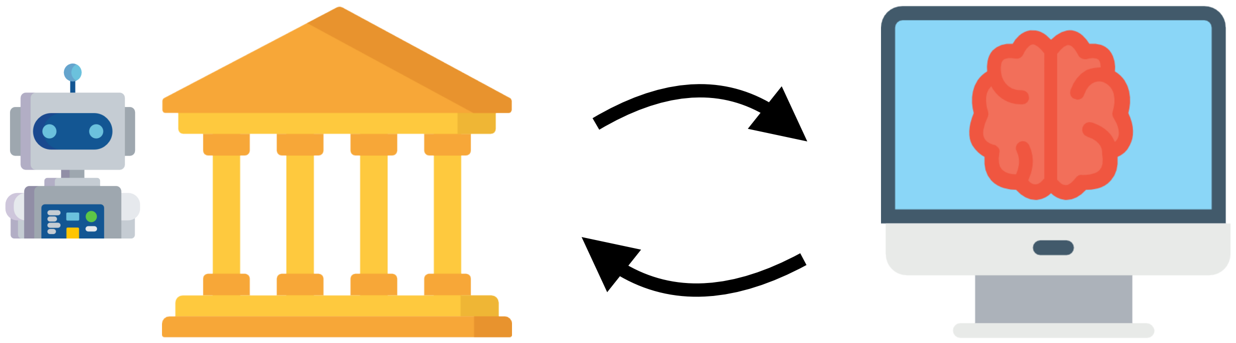
The term counterfactual has arguably become mainstream in the field of machine learning after the seminal work on counterfactual explanations by Wachter et al.

Wachter et al. "*Counterfactual explanations without opening the black box: Automated decisions and the GDPR.*" Harv. JL & Tech., 2017.

Counterfactual explanations

The term counterfactual has arguably become mainstream in the field of machine learning after the seminal work on counterfactual explanations by Wachter et al.

Kate



Bank

ML

Wachter et al. "*Counterfactual explanations without opening the black box: Automated decisions and the GDPR.*" Harv. JL & Tech., 2017.

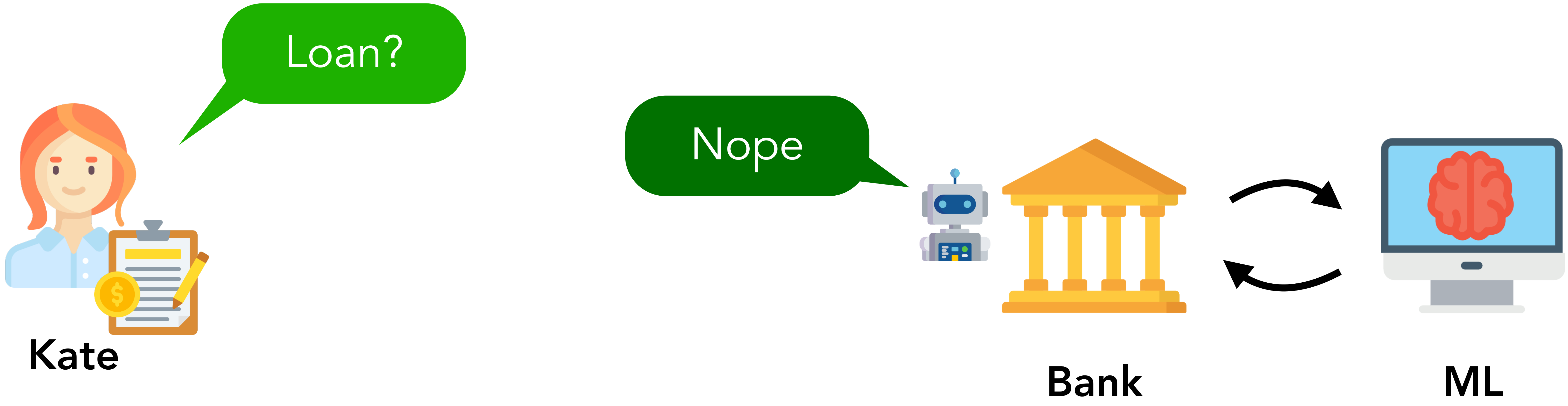
Counterfactual explanations

The term counterfactual has arguably become mainstream in the field of machine learning after the seminal work on counterfactual explanations by Wachter et al.

Wachter et al. "*Counterfactual explanations without opening the black box: Automated decisions and the GDPR.*" Harv. JL & Tech., 2017.

Counterfactual explanations

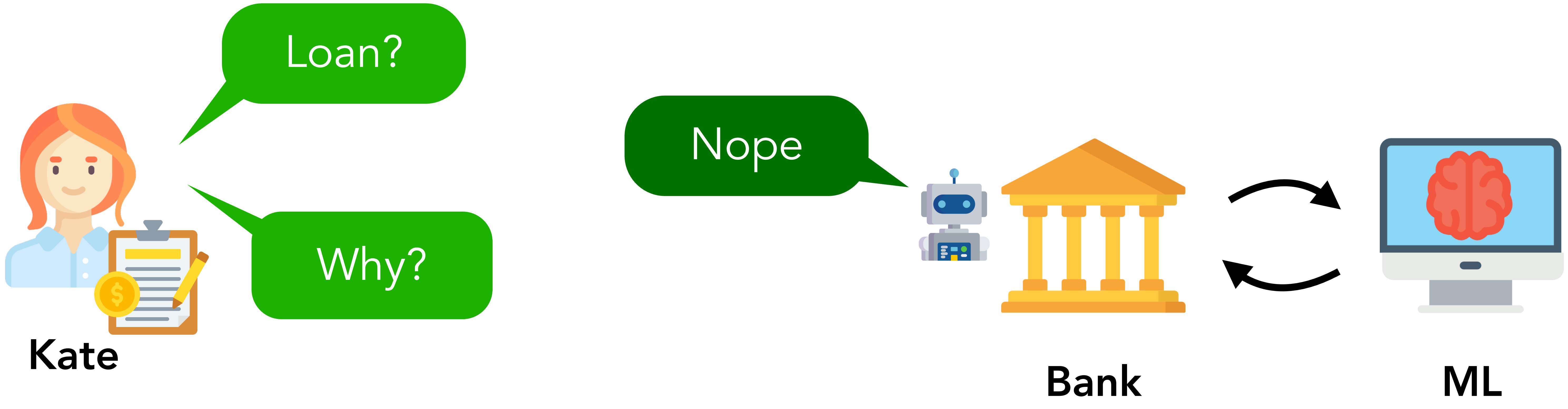
The term counterfactual has arguably become mainstream in the field of machine learning after the seminal work on counterfactual explanations by Wachter et al.



Wachter et al. "*Counterfactual explanations without opening the black box: Automated decisions and the GDPR.*" Harv. JL & Tech., 2017.

Counterfactual explanations

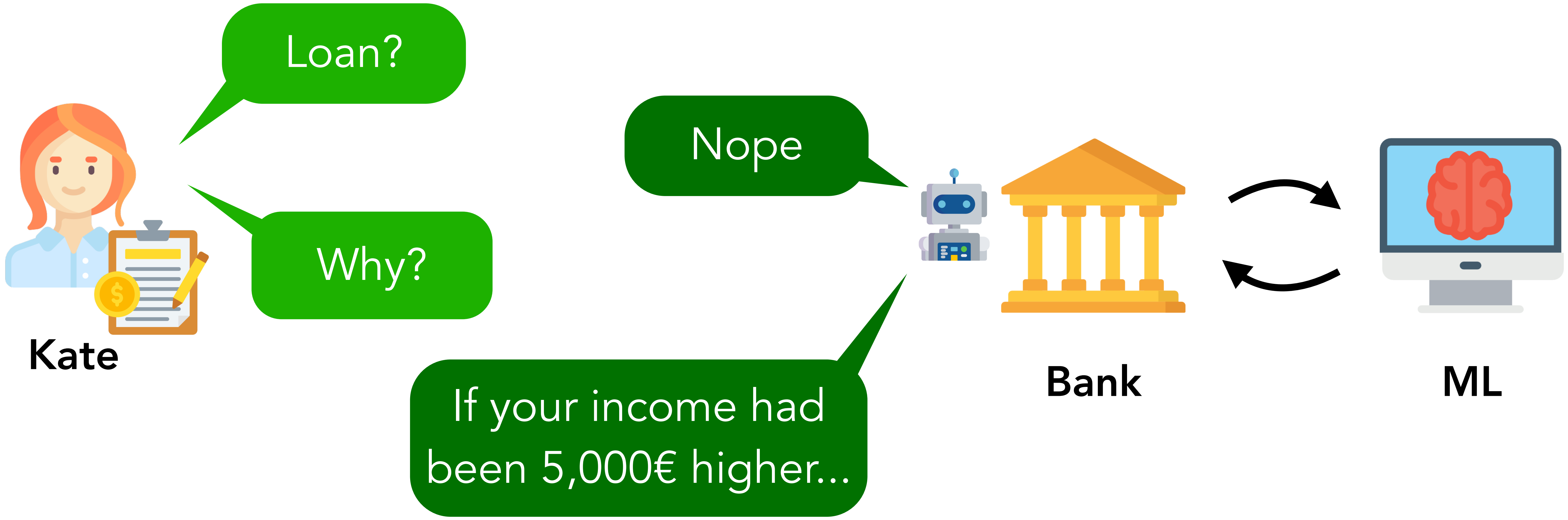
The term counterfactual has arguably become mainstream in the field of machine learning after the seminal work on counterfactual explanations by Wachter et al.



Wachter et al. "*Counterfactual explanations without opening the black box: Automated decisions and the GDPR.*" Harv. JL & Tech., 2017.

Counterfactual explanations

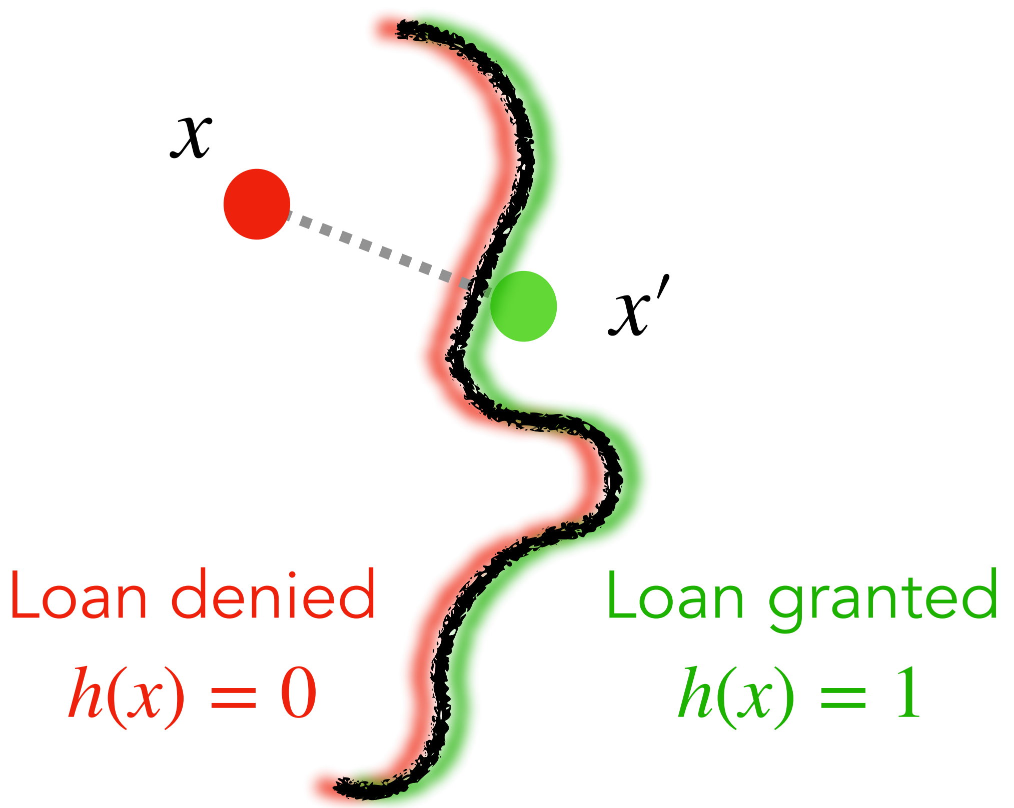
The term counterfactual has arguably become mainstream in the field of machine learning after the seminal work on counterfactual explanations by Wachter et al.



Wachter et al. "*Counterfactual explanations without opening the black box: Automated decisions and the GDPR.*" Harv. JL & Tech., 2017.

Counterfactual explanations

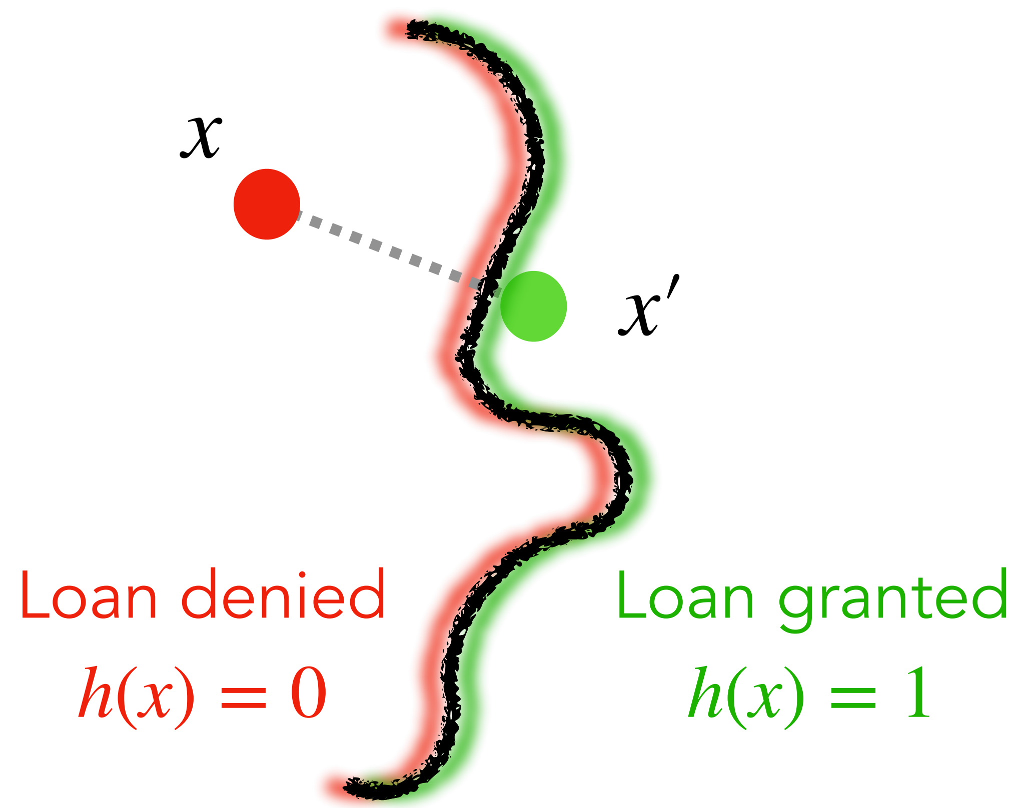
Given a (binary) prediction $h(x)$ by a machine learning model about an individual with features x, a counterfactual explanation is given by the closest feature value x' under which $h(x') \neq h(x)$



Wachter et al. "*Counterfactual explanations without opening the black box: Automated decisions and the GDPR.*" Harv. JL & Tech., 2017.

Counterfactual explanations

Given a (binary) prediction $h(x)$ by a machine learning model about an individual with features x, a counterfactual explanation is given by the closest feature value x' under which $h(x') \neq h(x)$



By showing a feature-perturbed version of an individual, a counterfactual explanations is, in principle, telling the individual what to do to secure a better decision in the future.

Wachter et al. "*Counterfactual explanations without opening the black box: Automated decisions and the GDPR.*" Harv. JL & Tech., 2017.

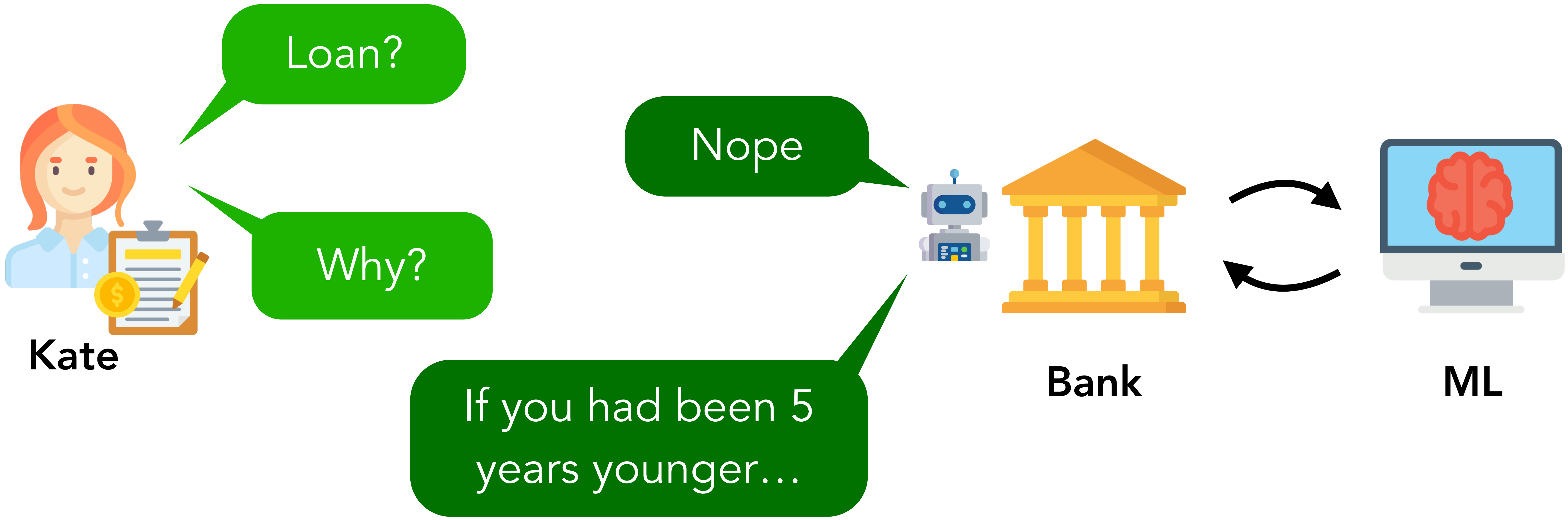
Counterfactual explanations

However, the closest feature value x' may not be actionable, and may not even be plausible.

Verma et al. "*Counterfactual explanations and algorithmic recourses for machine learning: A review.*" ACM Computing Surveys, 2024.

Counterfactual explanations

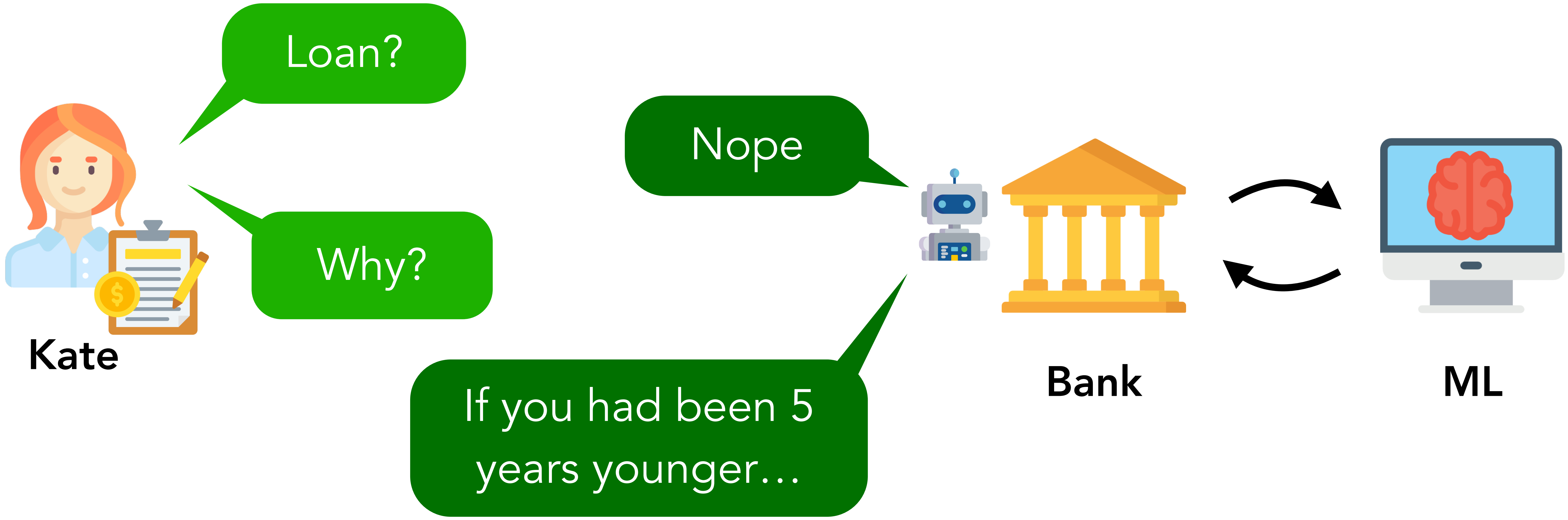
However, the closest feature value x' may not be actionable, and may not even be plausible.



Verma et al. "*Counterfactual explanations and algorithmic recourses for machine learning: A review.*" ACM Computing Surveys, 2024.

Counterfactual explanations

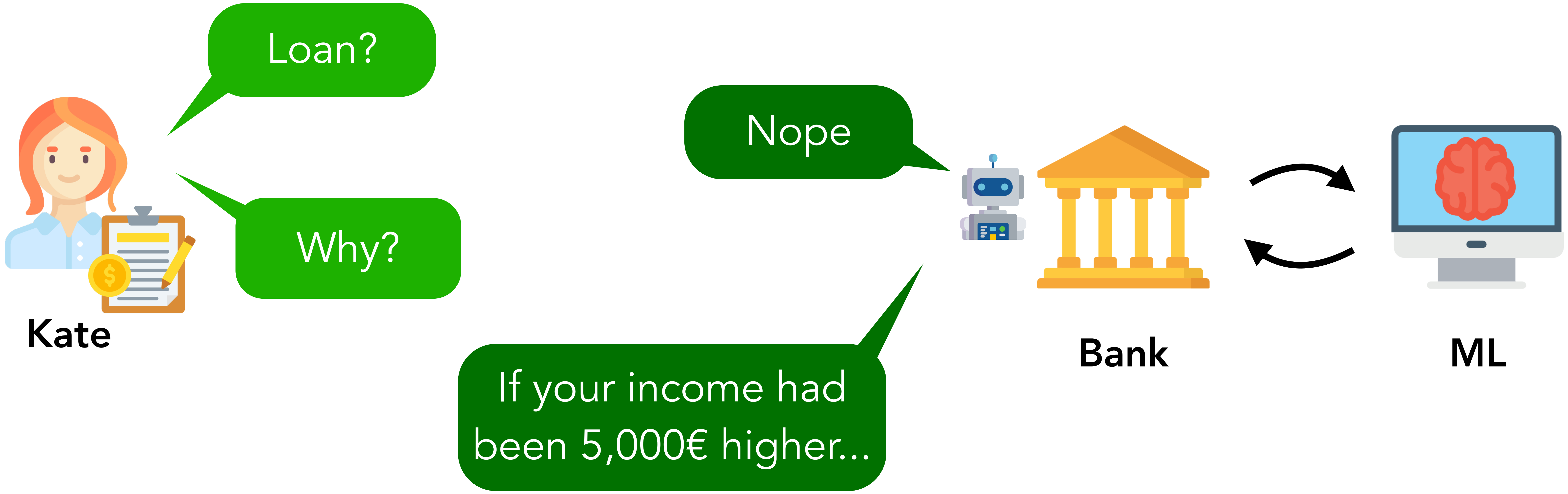
However, the closest feature value x' may not be actionable, and may not even be plausible.



Many follow-up works have addressed this problem by finding the closest feature value subject to a variety of actionability and plausibility constraints.

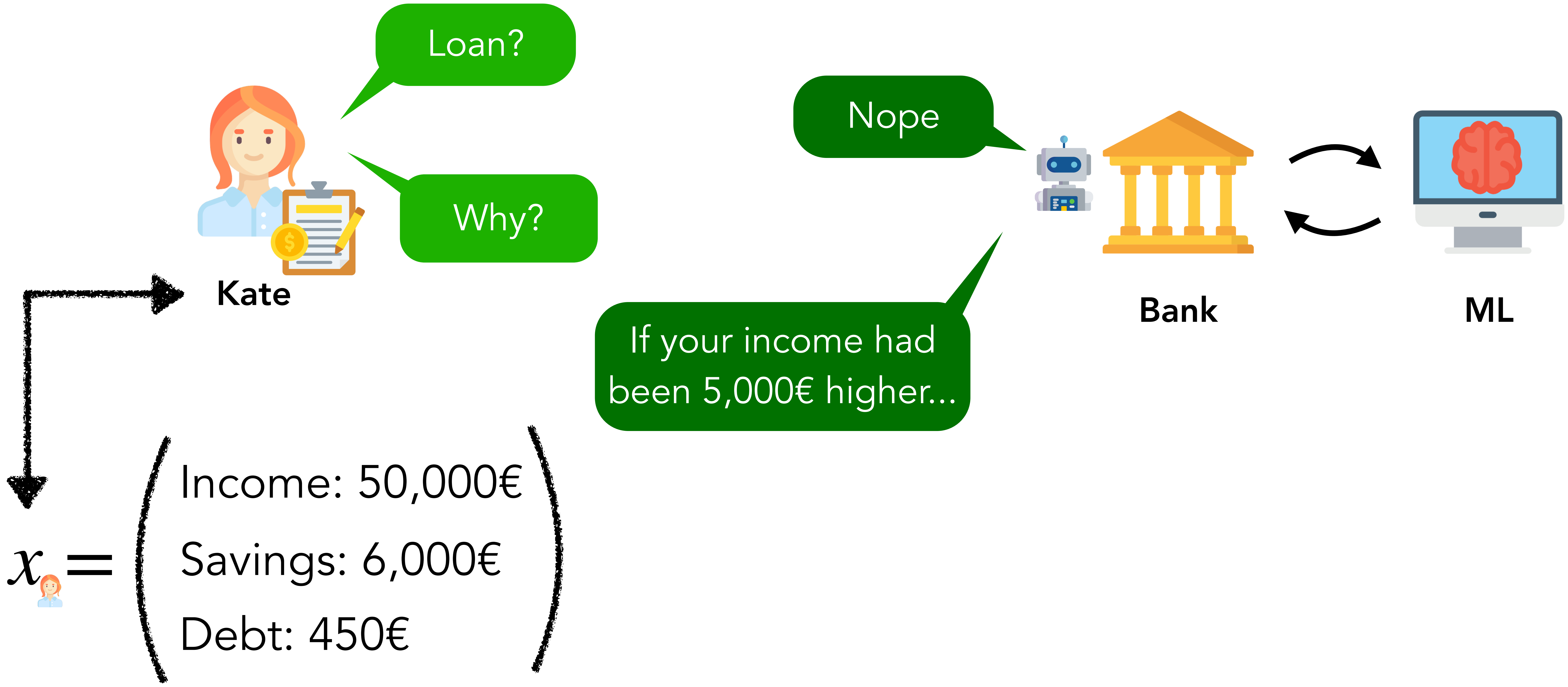
Verma et al. "*Counterfactual explanations and algorithmic recourses for machine learning: A review.*" ACM Computing Surveys, 2024.

Counterfactual explanations ignore causal dependencies



Beckers. "*Causal explanations and xai.*" CLeaR, 2022.
Crupi et al. "*Counterfactual explanations as interventions in latent space.*" DMKD, 2022.

Counterfactual explanations ignore causal dependencies



Loan?

Why?

Kate

Nope

If your income had been 5,000€ higher...

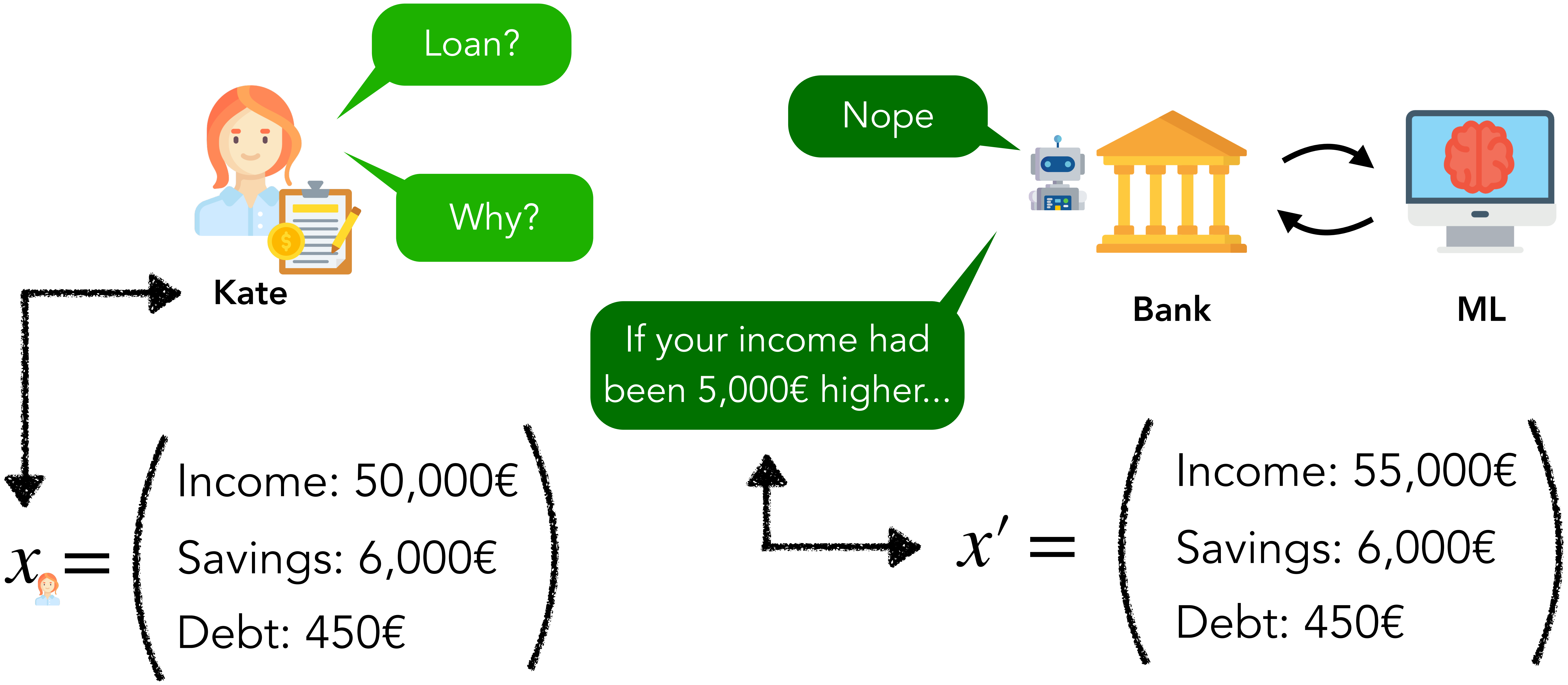
Bank

ML

$$x = \begin{pmatrix} \text{Income: 50,000€} \\ \text{Savings: 6,000€} \\ \text{Debt: 450€} \end{pmatrix}$$

Beckers. "*Causal explanations and xai.*" CLeaR, 2022.
Crupi et al. "*Counterfactual explanations as interventions in latent space.*" DMKD, 2022.

Counterfactual explanations ignore causal dependencies



Loan?

Why?

Kate

Nope

Bank

ML

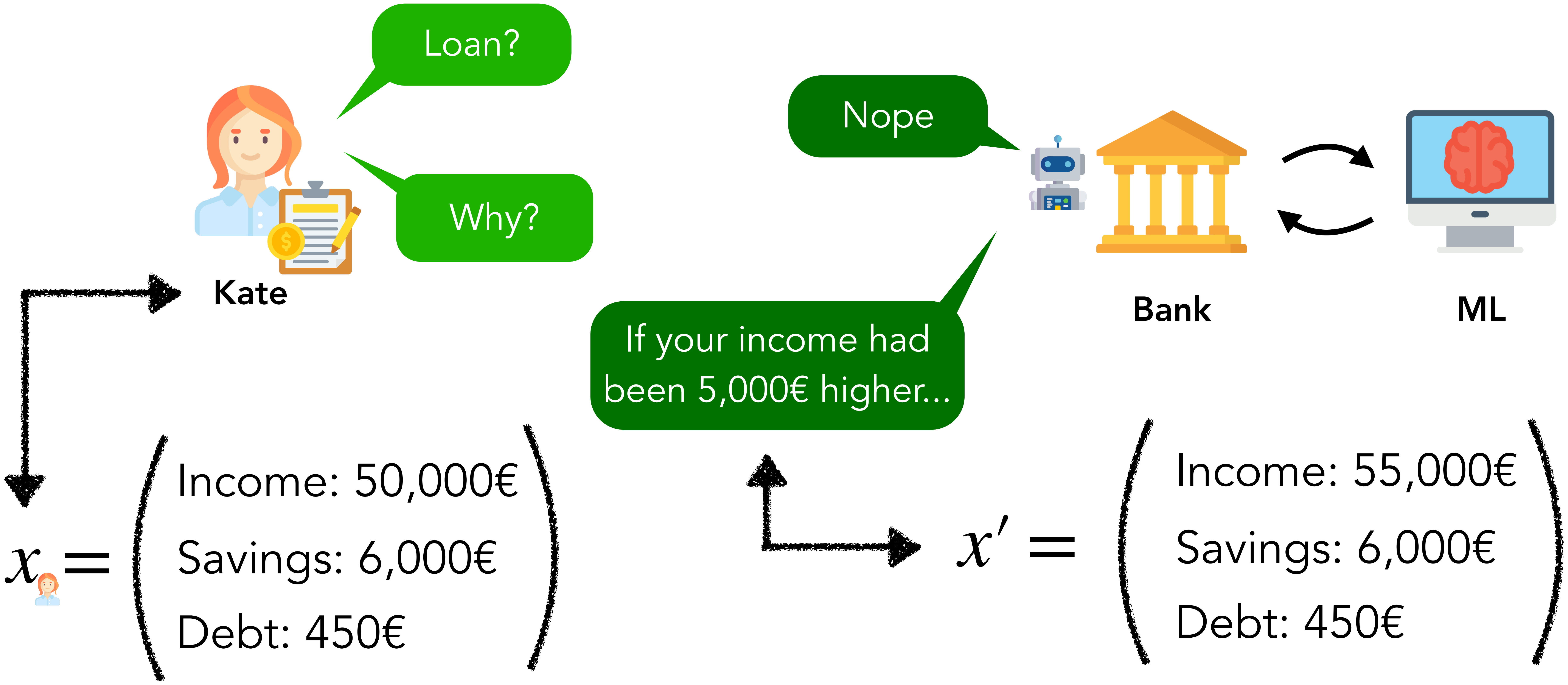
If your income had been 5,000€ higher...

$$x = \begin{pmatrix} \text{Income: } 50{,}000€ \\ \text{Savings: } 6{,}000€ \\ \text{Debt: } 450€ \end{pmatrix}$$

$$x' = \begin{pmatrix} \text{Income: } 55{,}000€ \\ \text{Savings: } 6{,}000€ \\ \text{Debt: } 450€ \end{pmatrix}$$

Beckers. "*Causal explanations and xai.*" CLeaR, 2022.
Crupi et al. "*Counterfactual explanations as interventions in latent space.*" DMKD, 2022.

Counterfactual explanations ignore causal dependencies



$$x_{\text{Kate}} = \begin{pmatrix} \text{Income: 50,000€} \\ \text{Savings: 6,000€} \\ \text{Debt: 450€} \end{pmatrix}$$

$$x' = \begin{pmatrix} \text{Income: 55,000€} \\ \text{Savings: 6,000€} \\ \text{Debt: 450€} \end{pmatrix}$$

If Kate's income had been 5,000€ higher, Kate's savings would have been more than 6,000€!

Beckers. "*Causal explanations and xai.*" CLeaR, 2022.
Crupi et al. "*Counterfactual explanations as interventions in latent space.*" DMKD, 2022.

Counterfactual explanations as interventions

A counterfactual explanation does not answer a counterfactual question but an interventional question.

Counterfactual explanations as interventions

A counterfactual explanation does not answer a counterfactual question but an interventional question.

Structural Causal Model $\mathscr{M}$

$$X_1 := f_{X_1}(D)$$

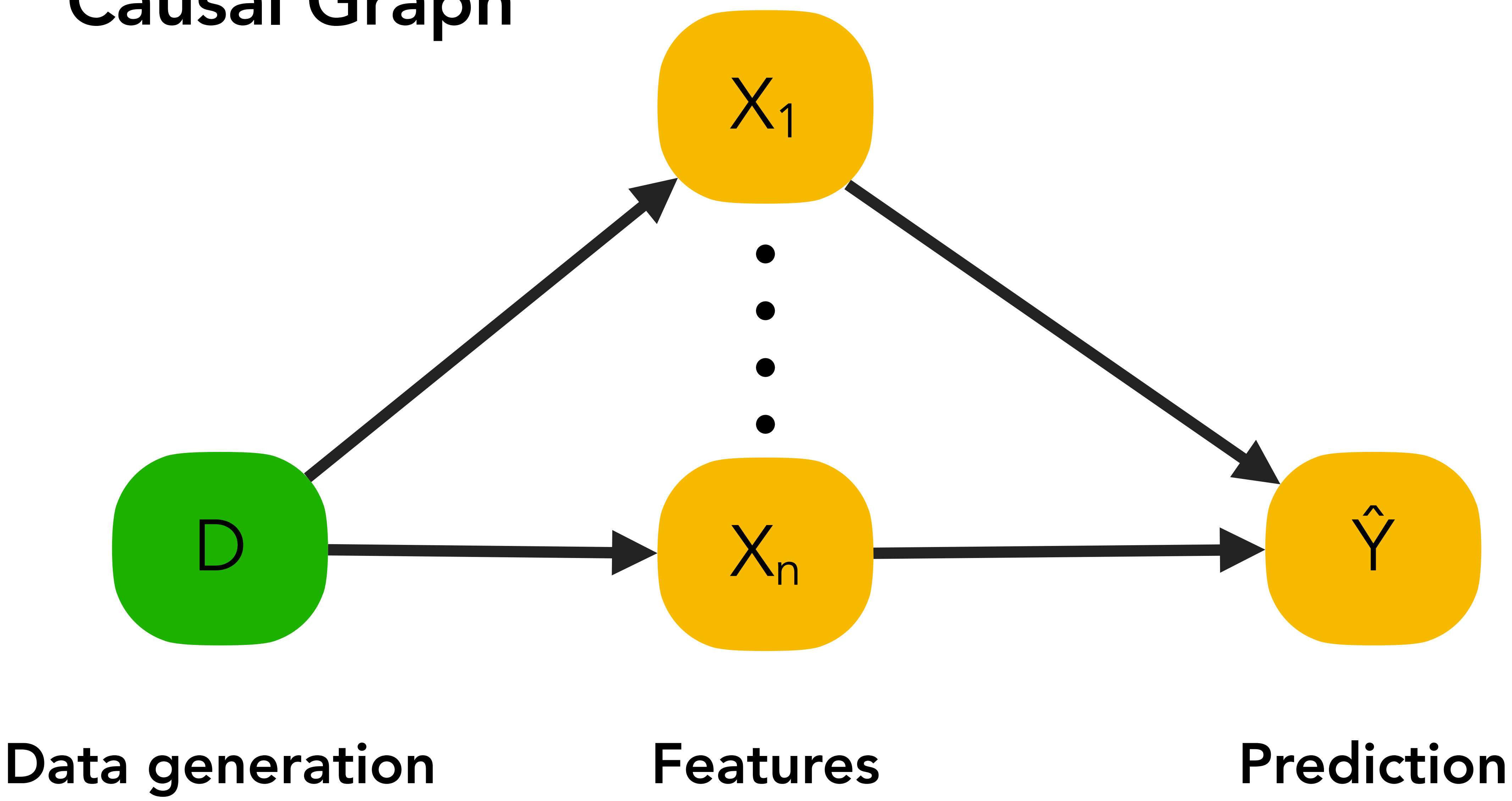
$$\vdots$$

$$X_n := f_{X_n}(D)$$

$$\hat{Y} := h(X)$$

$$D \sim P(D)$$

Causal Graph



Counterfactual explanations as interventions

A counterfactual explanation does not answer a counterfactual question but an interventional question.

Structural Causal Model $\mathscr{M}$

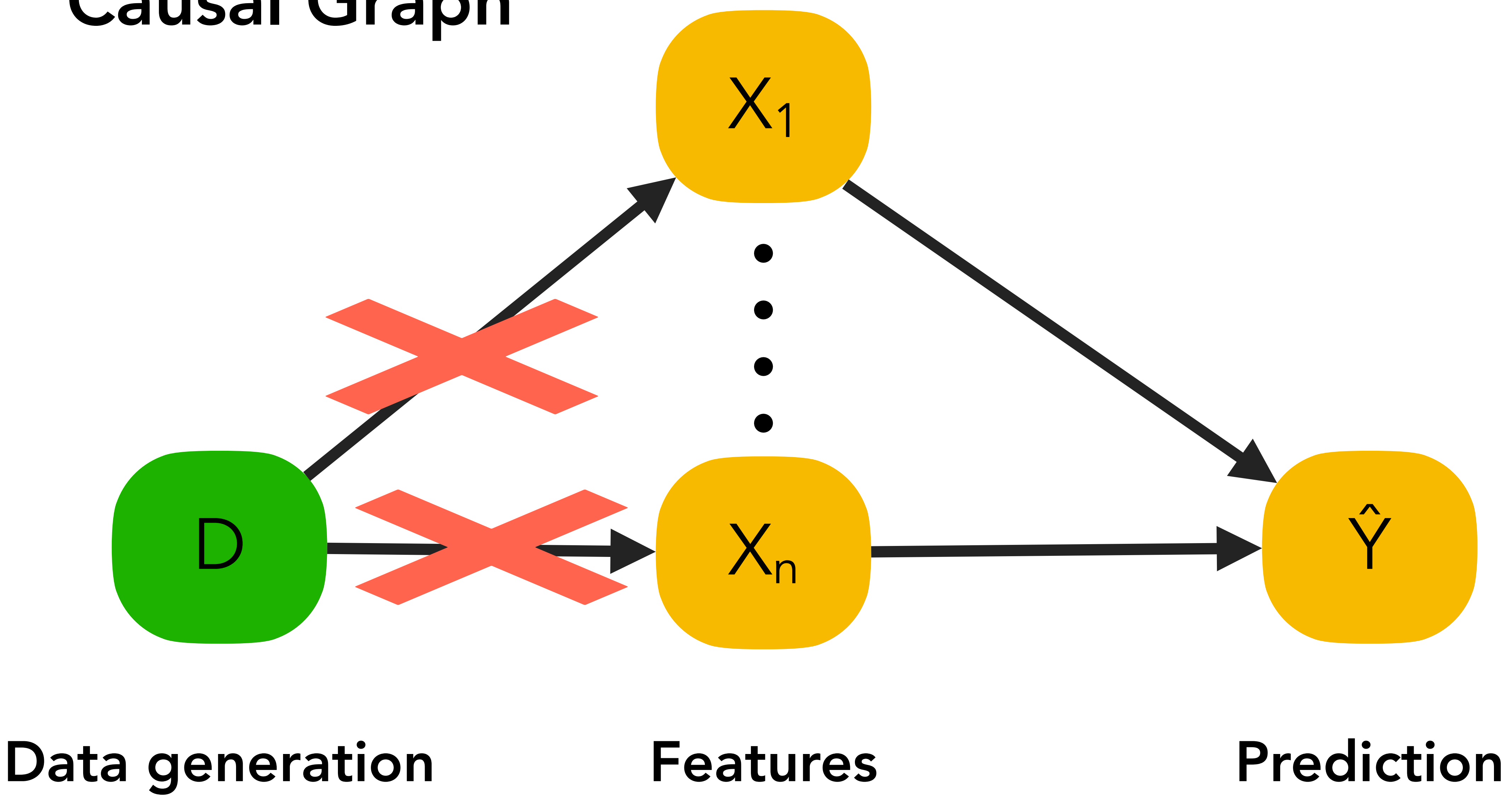
$$\begin{array}{ll} \cancel{X_1 := f_{X_1}(D)} & X_1 := x_1' \\ \vdots & \vdots \\ \cancel{X_n := f_{X_n}(D)} & X_n := x_n' \end{array}$$

$$\hat{Y} := h(X)$$

$$D \sim P(D)$$

Causal Graph

$$\boxed{do(X := x')}$$



Counterfactual explanations as interventions

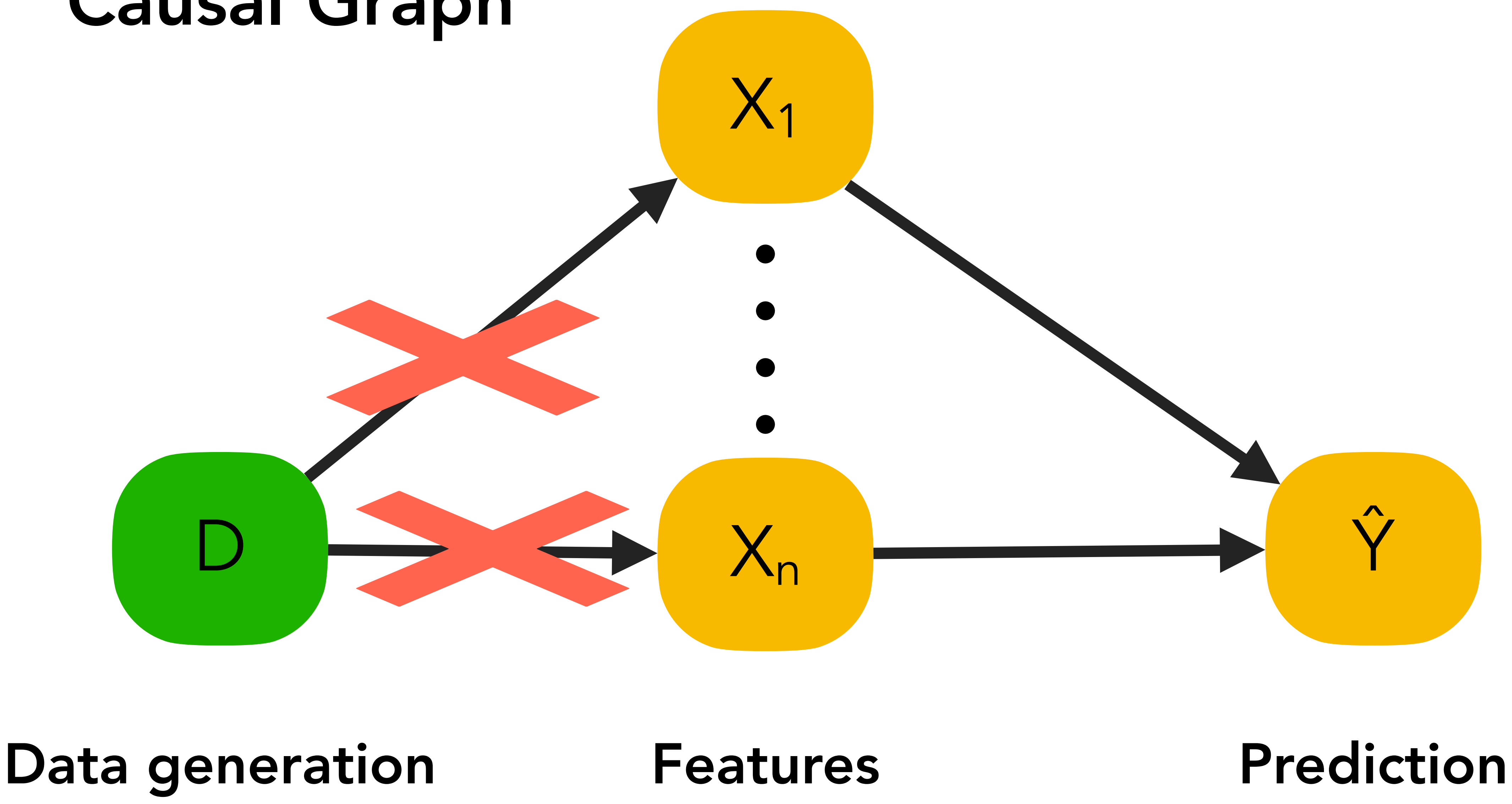
A counterfactual explanation does not answer a counterfactual question but an interventional question.

Structural Causal Model $\mathscr{M}$

$$
\begin{array}{ll}
\cancel{X_1 := f_{X_1}(D)} & X_1 := x_1' \\
\vdots & \vdots \\
\cancel{X_n := f_{X_n}(D)} & X_n := x_n' \\
\hat{Y} := h(X) & \\
D \sim P(D) &
\end{array}
$$

$do(X := x')$

Causal Graph



A counterfactual explanation encourages an individual to change the value of the features x_l such that $x_l \neq x_l'$. However, it does not take into account that such a change may induce changes in features x_l such that $x_l = x_l'$.

Algorithmic recourse

Algorithmic recourse seeks to find the minimal intervention a under which $h(x + a) \neq h(x)$ *while accounting for causal dependencies between features.*

Karimi et al. "*Algorithmic recourse: from counterfactual explanations to interventions.*" FAccT, 2021.
Karimi et al. "*Algorithmic recourse under imperfect causal knowledge: a probabilistic approach.*" NeurIPS, 2020.

Algorithmic recourse

Algorithmic recourse seeks to find the minimal intervention a under which $h(x + a) \neq h(x)$ *while accounting for causal dependencies between features.*

Structural Causal Model $\mathscr{M}$

$$X_1 := f_{X_1}(D)$$

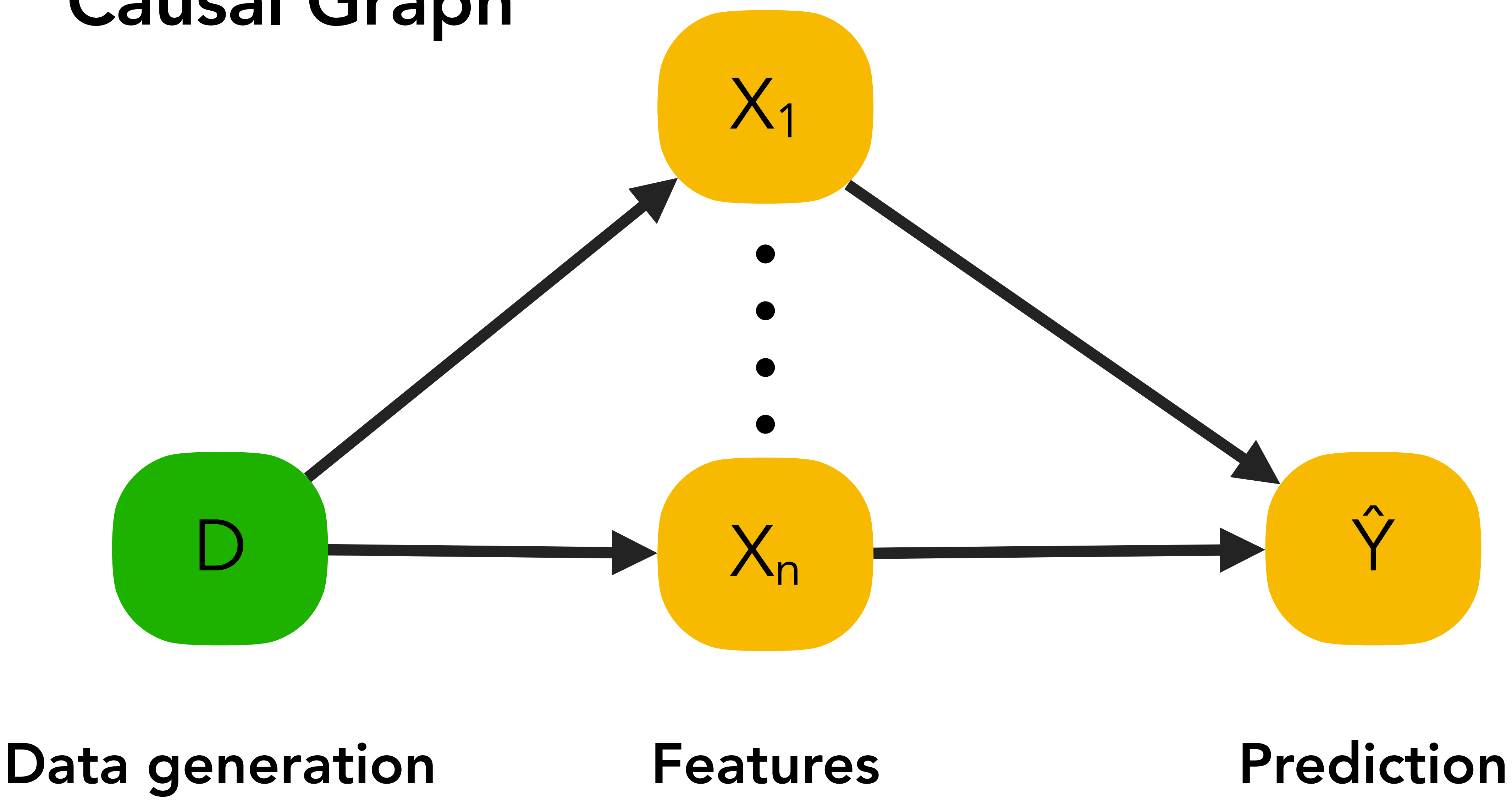
$$\vdots$$

$$X_n := f_{X_n}(D)$$

$$\hat{Y} := h(X)$$

$$D \sim P(D)$$

Causal Graph



Karimi et al. "*Algorithmic recourse: from counterfactual explanations to interventions.*" FAccT, 2021.
Karimi et al. "*Algorithmic recourse under imperfect causal knowledge: a probabilistic approach.*" NeurIPS, 2020.

Algorithmic recourse

Algorithmic recourse seeks to find the minimal intervention a under which $h(x+a) \neq h(x)$ *while accounting for causal dependencies between features.*

Modified Structural Causal Model $\mathscr{M}_{X=x}$

$$X_1 := f_{X_1}(D)$$

$$\vdots$$

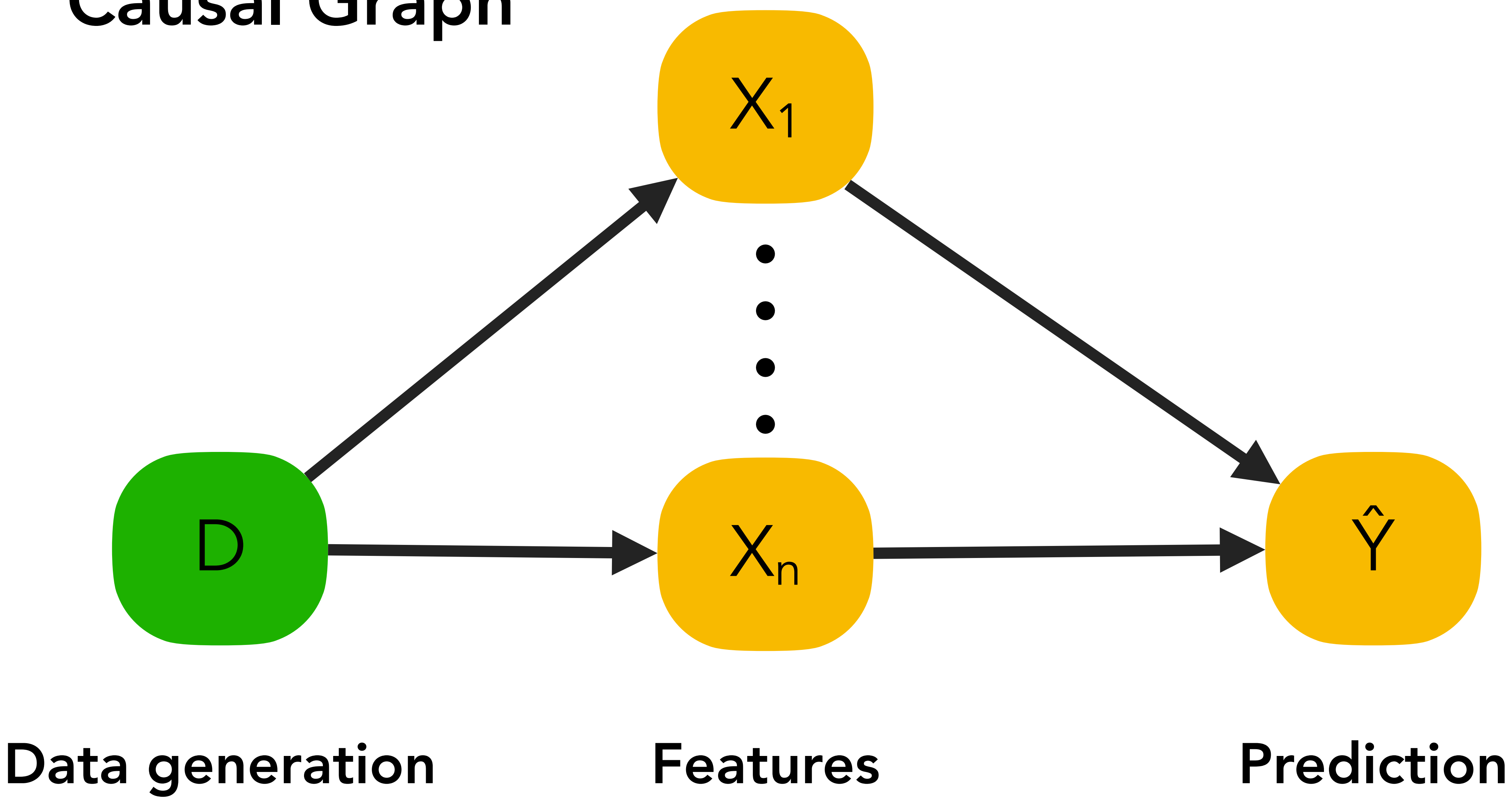
$$X_n := f_{X_n}(D)$$

$$\hat{Y} := h(X)$$

$$D \sim P(D \mid X = x) \dashleftarrow$$

Posterior distribution of the noise

Causal Graph



Karimi et al. "*Algorithmic recourse: from counterfactual explanations to interventions.*" FAccT, 2021.
Karimi et al. "*Algorithmic recourse under imperfect causal knowledge: a probabilistic approach.*" NeurIPS, 2020.

Algorithmic recourse

Algorithmic recourse seeks to find the minimal intervention a under which $h(x + a) \neq h(x)$ *while accounting for causal dependencies between features.*

Modified Structural Causal Model $\mathscr{M}_{X=x}$

$$X_1 := f_{X_1}(D) + a_1$$

$$\vdots$$

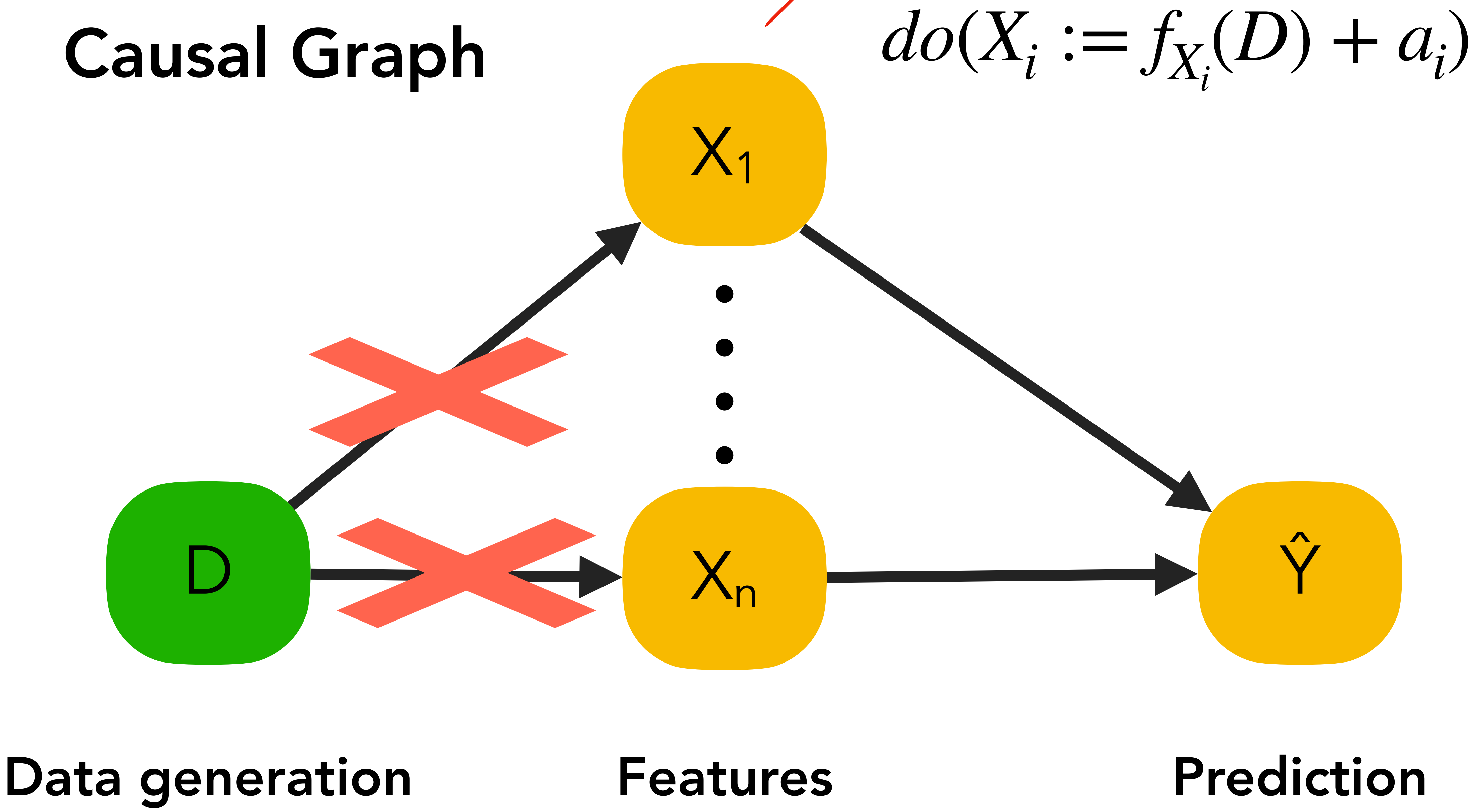
$$X_n := f_{X_n}(D) + a_n$$

$$\hat{Y} := h(X)$$

$$D \sim P(D \mid X = x) \dashleftarrow$$

Posterior distribution of the noise

Causal Graph



Whenever $a_i = 0$, the value of X_i may still change!

Karimi et al. "*Algorithmic recourse: from counterfactual explanations to interventions.*" FAccT, 2021.
Karimi et al. "*Algorithmic recourse under imperfect causal knowledge: a probabilistic approach.*" NeurIPS, 2020.

Counterfactual explanations & performativity

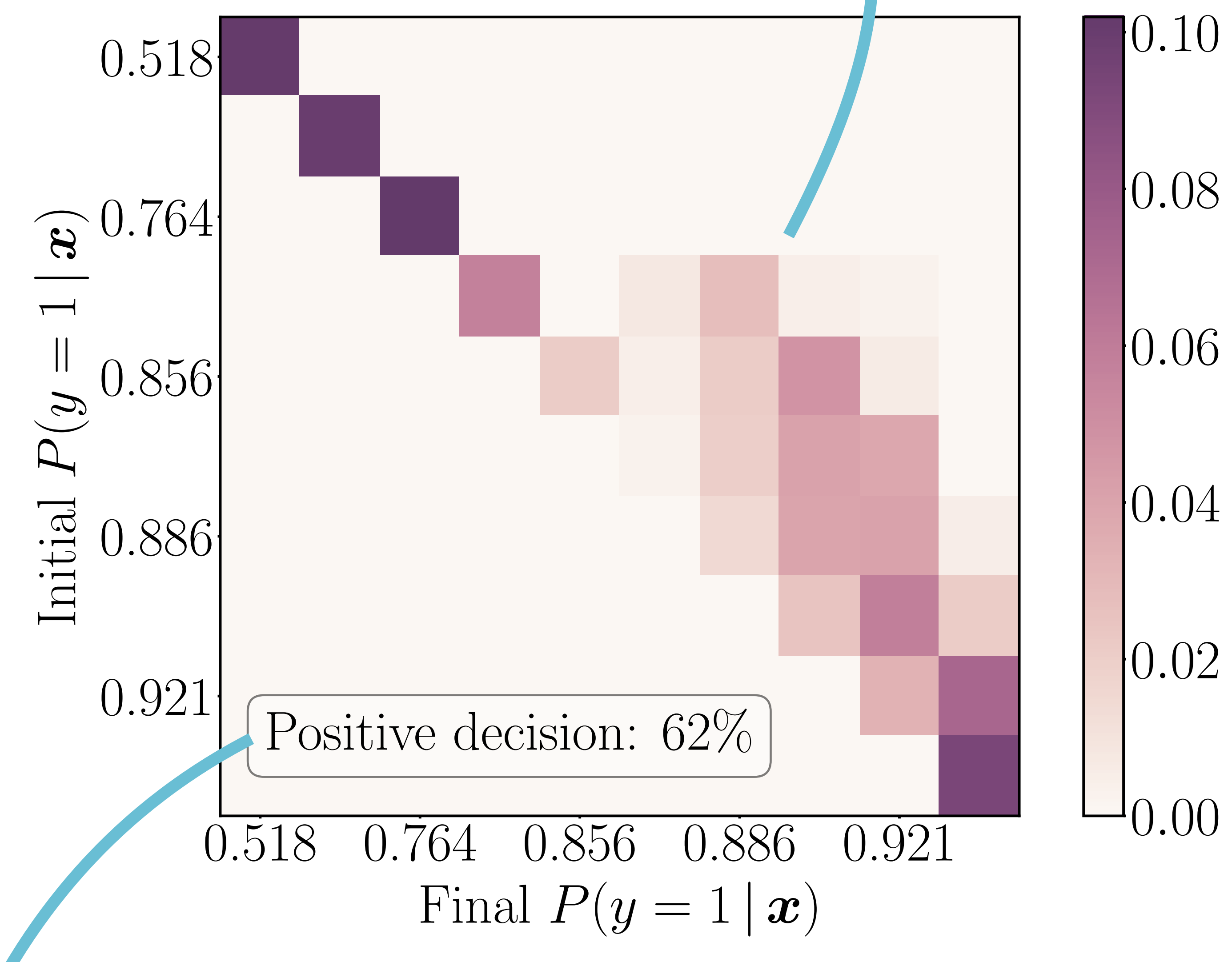
If a sizable number of individuals follow the changes prescribed by counterfactual explanations, the feature distribution $P(X)$ may change.

Tsirtsis and Gomez-Rodriguez. "*Decisions, counterfactual explanations and strategic behavior.*" NeurIPS, 2020.
Perdomo et al. "*Performative prediction.*" ICML, 2020.

Counterfactual explanations & performativity

If a sizable number of individuals follow the changes prescribed by counterfactual explanations, the feature distribution $P(X)$ may change.

Chances of repayment would improve for large part of the population



More people would receive credit

Tsirtsis and Gomez-Rodriguez. "*Decisions, counterfactual explanations and strategic behavior.*" NeurIPS, 2020.
Perdomo et al. "*Performative prediction.*" ICML, 2020.

Counterfactual explanations & performativity

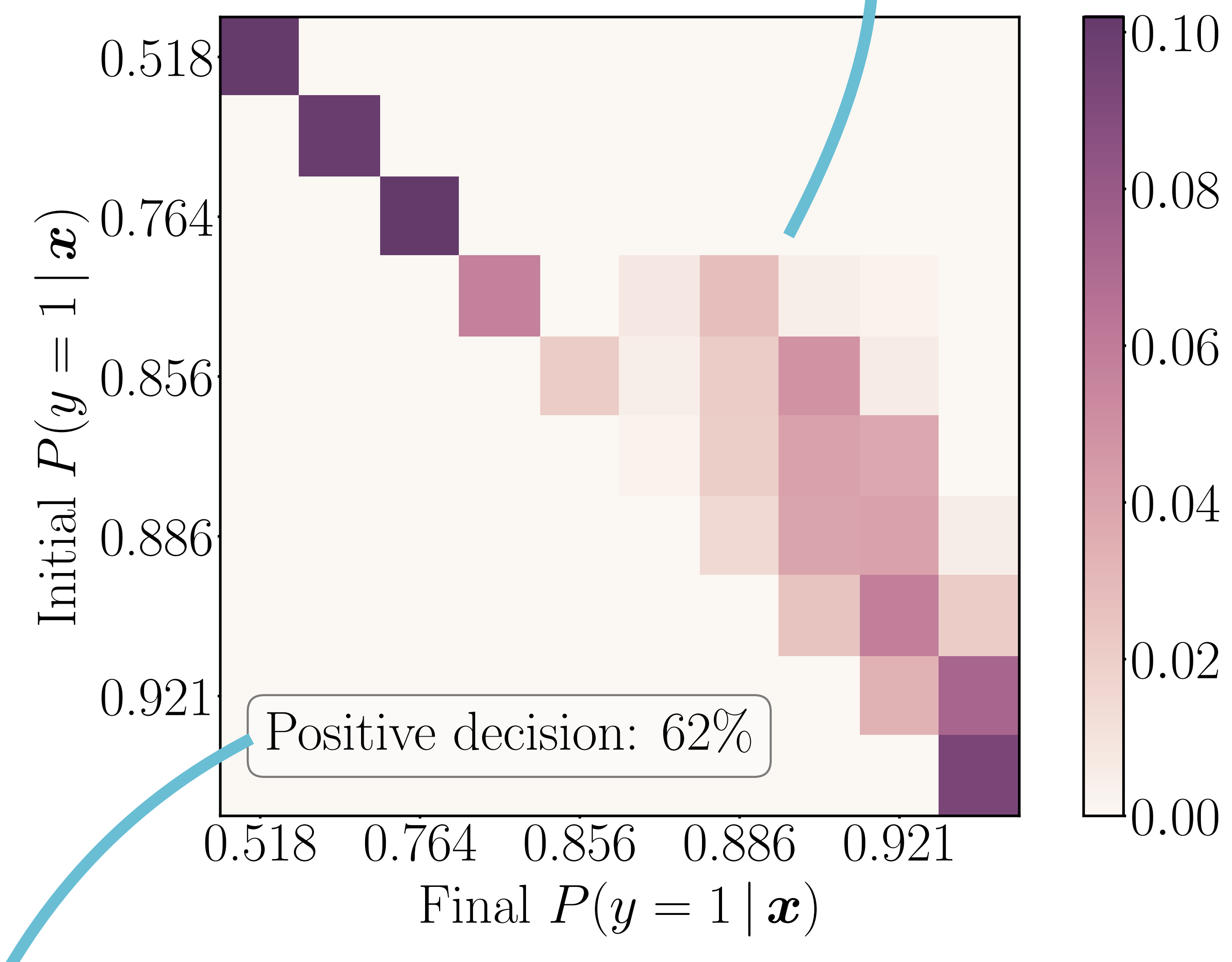
If a sizable number of individuals follow the changes prescribed by counterfactual explanations, the feature distribution $P(X)$ may change.

This raises the question of finding decision policies π *and counterfactual explanations* $\mathscr{A}$ *that are optimal in terms of utility.*

$$\max_{\pi,\mathscr{A}} u(\pi,\mathscr{A}) := \mathbb{E}_{x\sim P(X\,|\,\pi,\mathscr{A})}\left[\pi(x)\big(P(Y=1\,|\,x)-\gamma\big)\right]$$

constant reflecting economic considerations of the decision maker

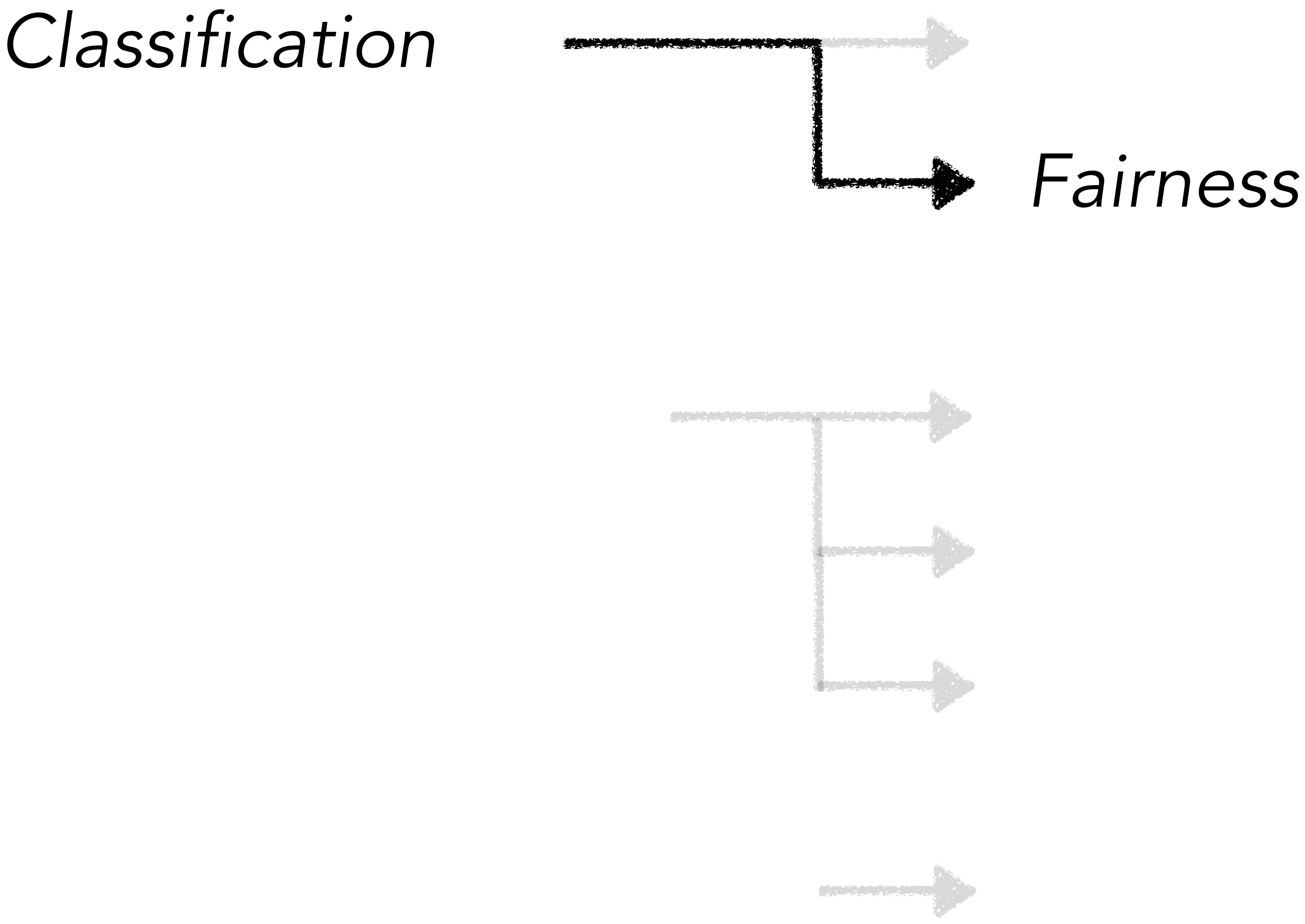
Chances of repayment would improve for large part of the population



More people would receive credit

Tsirtsis and Gomez-Rodriguez. "*Decisions, counterfactual explanations and strategic behavior.*" NeurIPS, 2020.
Perdomo et al. "*Performative prediction.*" ICML, 2020.

Use cases of counterfactuals in machine learning



Counterfactual fairness

Counterfactual fairness captures the intuition that a prediction by a machine learning model is fair towards an individual who belongs to a demographic group $A = a$ if it would have been the same had the individual belonged to a different demographic group $A = a'$.

Kusner et al. “*Counterfactual fairness.*” NeurIPS, 2017.

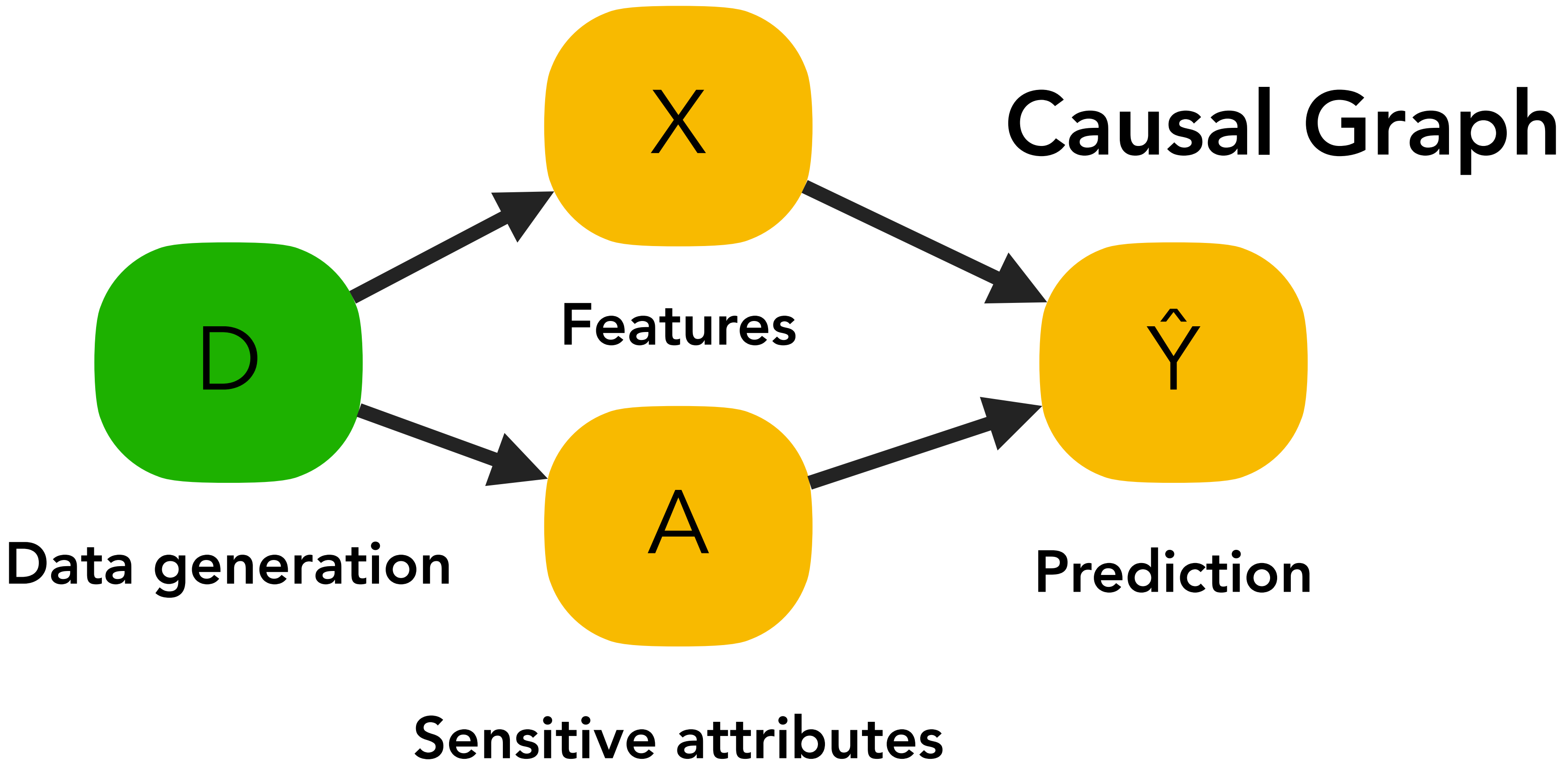
Counterfactual fairness

Counterfactual fairness captures the intuition that a prediction by a machine learning model is fair towards an individual who belongs to a demographic group $A = a$ if it would have been the same had the individual belonged to a different demographic group $A = a'$.

Structural Causal Model $\mathscr{M}$

$X := f_X(D) \qquad \hat{Y} := h(X, A)$

$A := f_A(D) \qquad D \sim P(D)$



Kusner et al. "*Counterfactual fairness.*" NeurIPS, 2017.

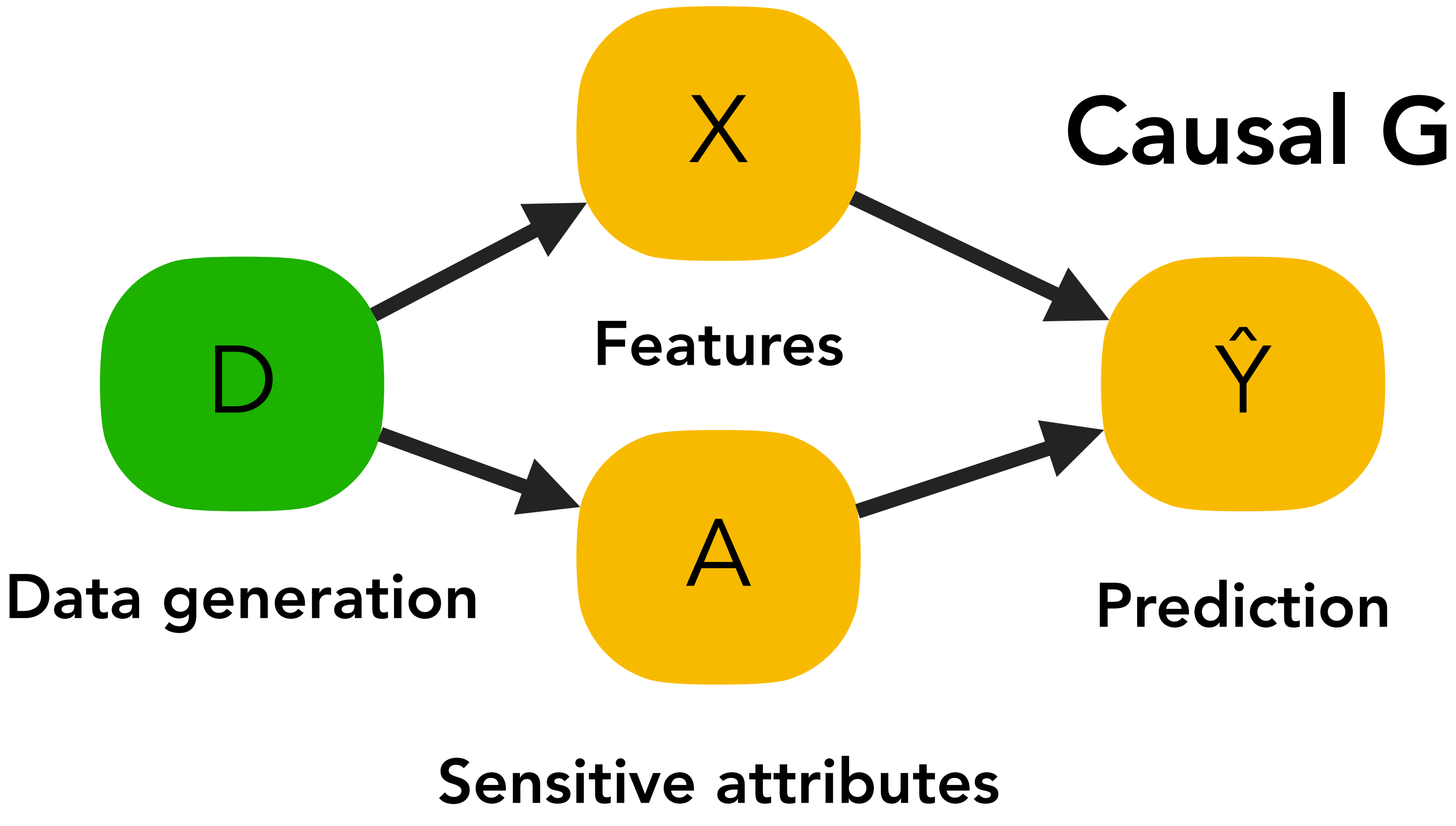
Counterfactual fairness

Counterfactual fairness captures the intuition that a prediction by a machine learning model is fair towards an individual who belongs to a demographic group $A = a$ if it would have been the same had the individual belonged to a different demographic group $A = a'$

Structural Causal Model $\mathscr{M}$

$$X := f_X(D) \qquad \hat{Y} := h(X, A)$$

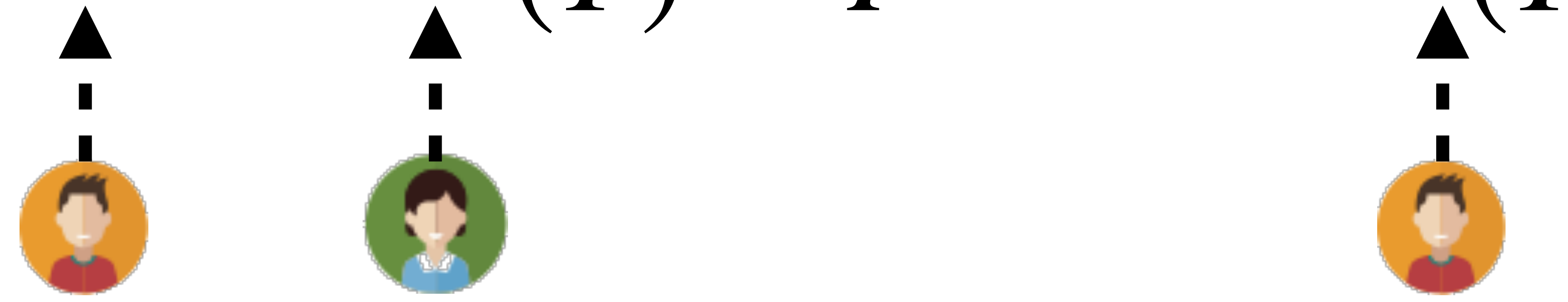
$$A := f_A(D) \qquad D \sim P(D)$$



Causal Graph

Counterfactual fairness

$$P^{\mathscr{M}\,|\,X=x,\,A=a\,;\,do(A=a')}(\hat{Y}) = P^{\mathscr{M}\,|\,X=x,\,A=a}(\hat{Y})$$



Kusner et al. "*Counterfactual fairness.*" NeurIPS, 2017.

Counterfactual fairness can be too restrictive

Counterfactual fairness considers the full effect of the demographic group on the prediction as problematic. However, this is not the case in certain scenarios.

Chiappa. "*Path-specific counterfactual fairness.*" AAAI, 2019.

Counterfactual fairness can be too restrictive

Counterfactual fairness considers the full effect of the demographic group on the prediction as problematic. However, this is not the case in certain scenarios.

Alleged gender bias case at Berkeley

8,442 male applicants for the fall of 1973, 44 percent were admitted,

4,351 female applicants, 35 percent were admitted

Chiappa. "*Path-specific counterfactual fairness.*" AAAI, 2019.

Counterfactual fairness can be too restrictive

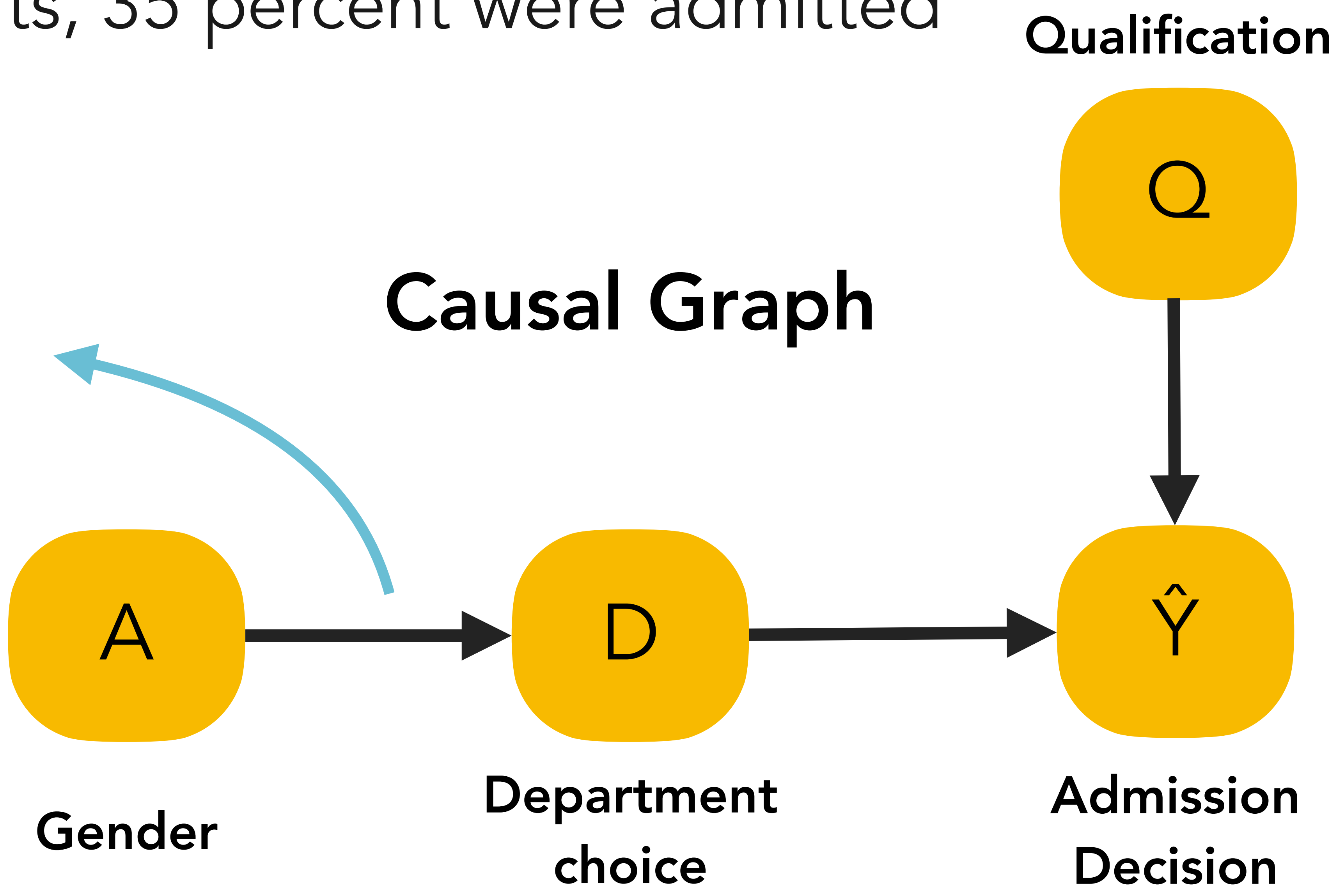
Counterfactual fairness considers the full effect of the demographic group on the prediction as problematic. However, this is not the case in certain scenarios.

Alleged gender bias case at Berkeley

8,442 male applicants for the fall of 1973, 44 percent were admitted,

4,351 female applicants, 35 percent were admitted

Female applied to departments with lower admission rates



Causal Graph

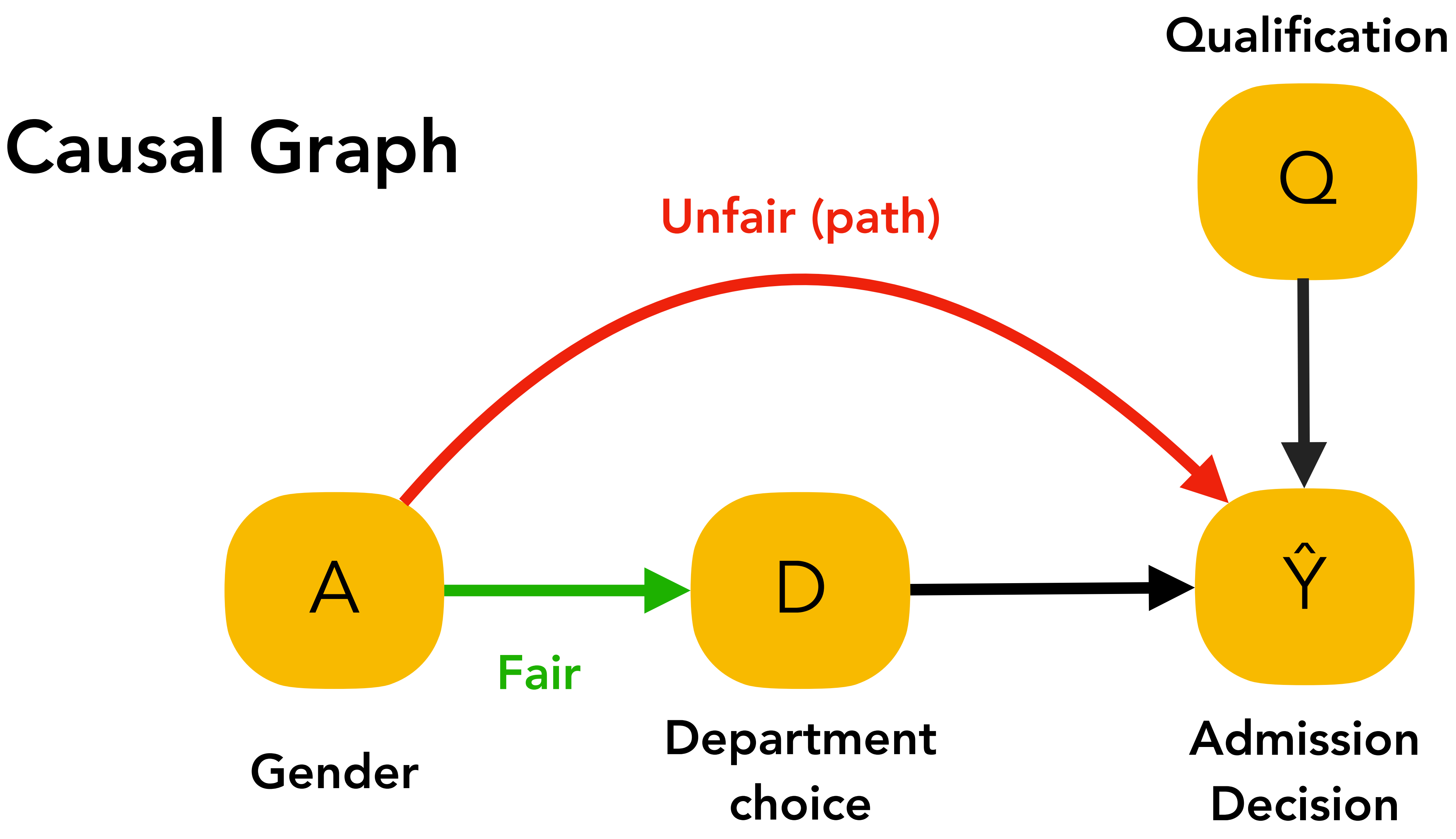
Counterfactual fairness is violated

$$P^{\mathscr{M} \mid Q=q, A=a\,;\, do(A=a')}(\hat{Y}) \neq P^{\mathscr{M} \mid Q=q, A=a}(\hat{Y})$$

Chiappa. "*Path-specific counterfactual fairness.*" AAAI, 2019.

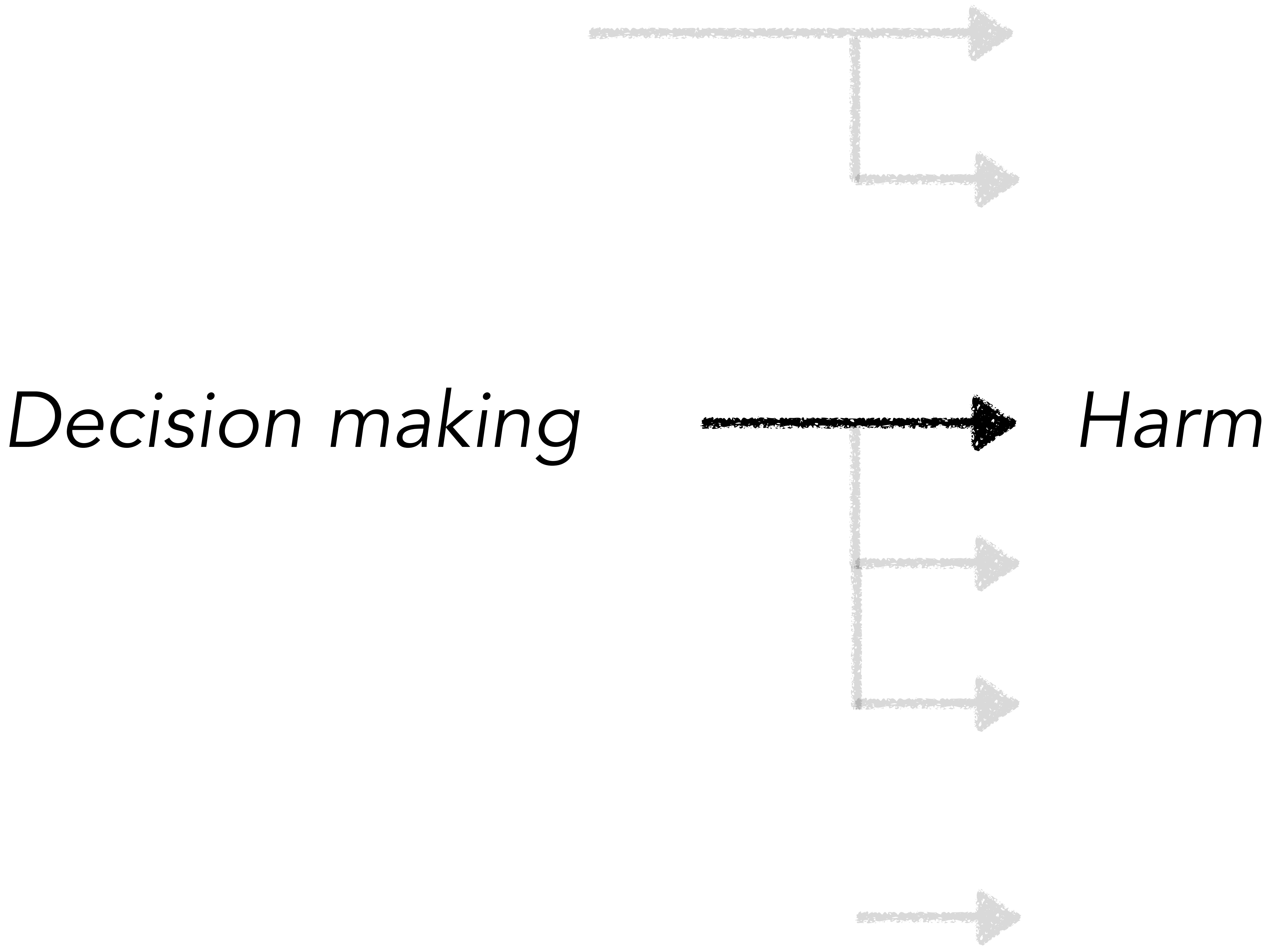
Path-specific counterfactual fairness

Path-specific counterfactual fairness is a more fine-grained fairness criterion that deals with sensitive attributes affecting the prediction along both fair and unfair pathways.



Chiappa. "*Path-specific counterfactual fairness.*" AAAI, 2019.

Use cases of counterfactuals in machine learning



Decision making → *Harm*

Counterfactual harm

Disease
50% mortality rate

Treatment A	60% chance of curing a patient	40% chance of having no effect
Treatment B	80% chance of curing a patient	20% chance of killing a patient

Counterfactual harm

Disease
50% mortality rate

Treatment A

60% chance of curing a patient

40% chance of having no effect

Treatment B

80% chance of curing a patient

20% chance of killing a patient

Treatments A and B have **identical recovery rates**. However, doctors would systematically favor treatment A as it achieves the same recovery rate but never harms the patient.

Counterfactual harm

Disease
50% mortality rate

Treatment A	**60% chance of curing a patient**	**40% chance of having no effect**
Treatment B	**80% chance of curing a patient**	**20% chance of killing a patient**

Treatments A and B have **identical recovery rates**. However, doctors would systematically favor treatment A as it achieves the same recovery rate but never harms the patient.

- Under treatment A, there are no patients that would have survived had they not been treated.
- Under treatment B, there are patients who die following treatment who would have lived had they not been treated.

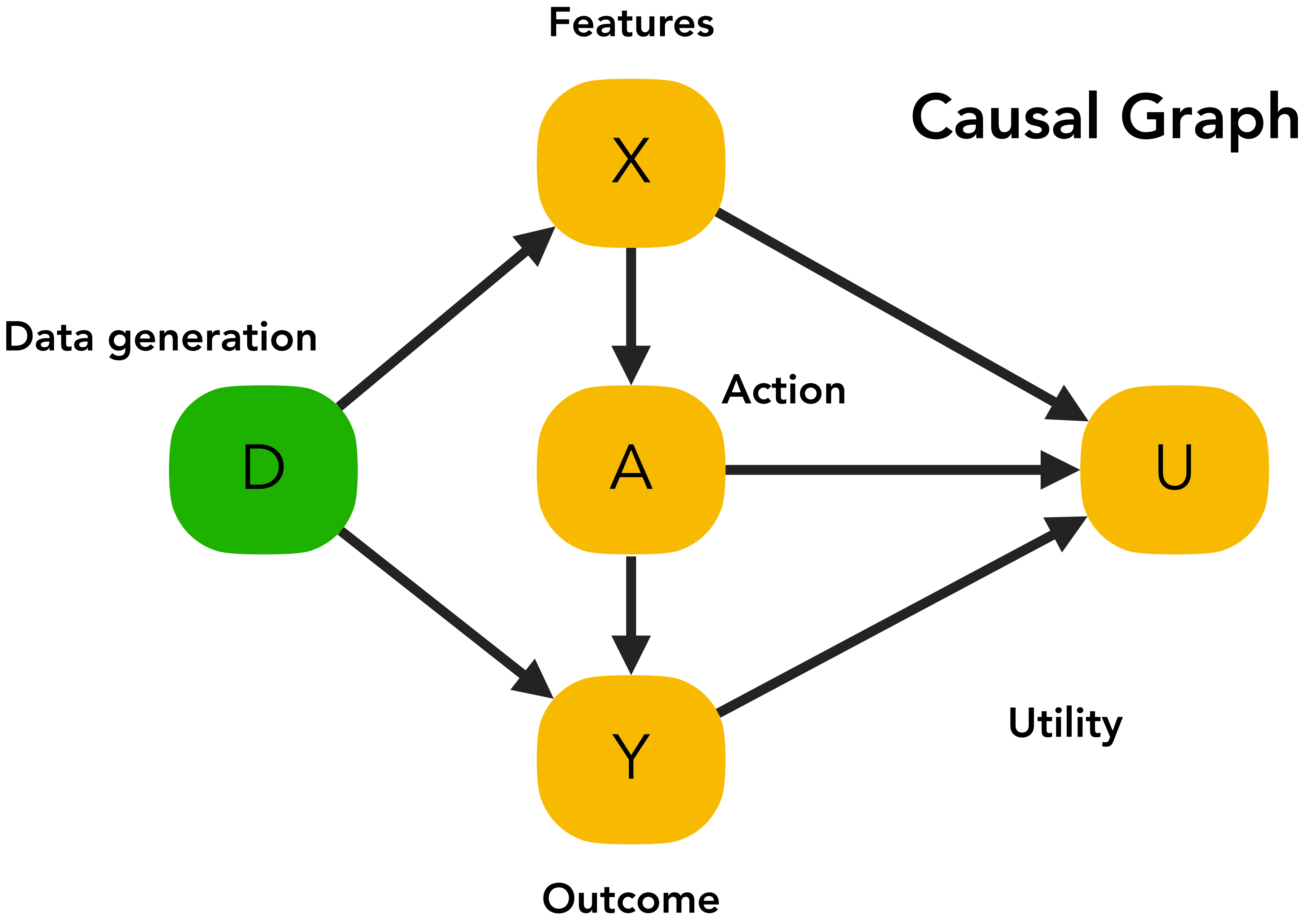
Formalizing counterfactual harm

Structural Causal Model $\mathscr{M}$

$X := f_X(D) \qquad Y := f_Y(D) \qquad D \sim P(D)$

$A := \pi(X)$ ◀- - **Algorithmic policy** 🤖

$U := f_U(A, X, Y)$



Richens et al. "*Counterfactual harm.*" NeurIPS, 2022.

Formalizing counterfactual harm

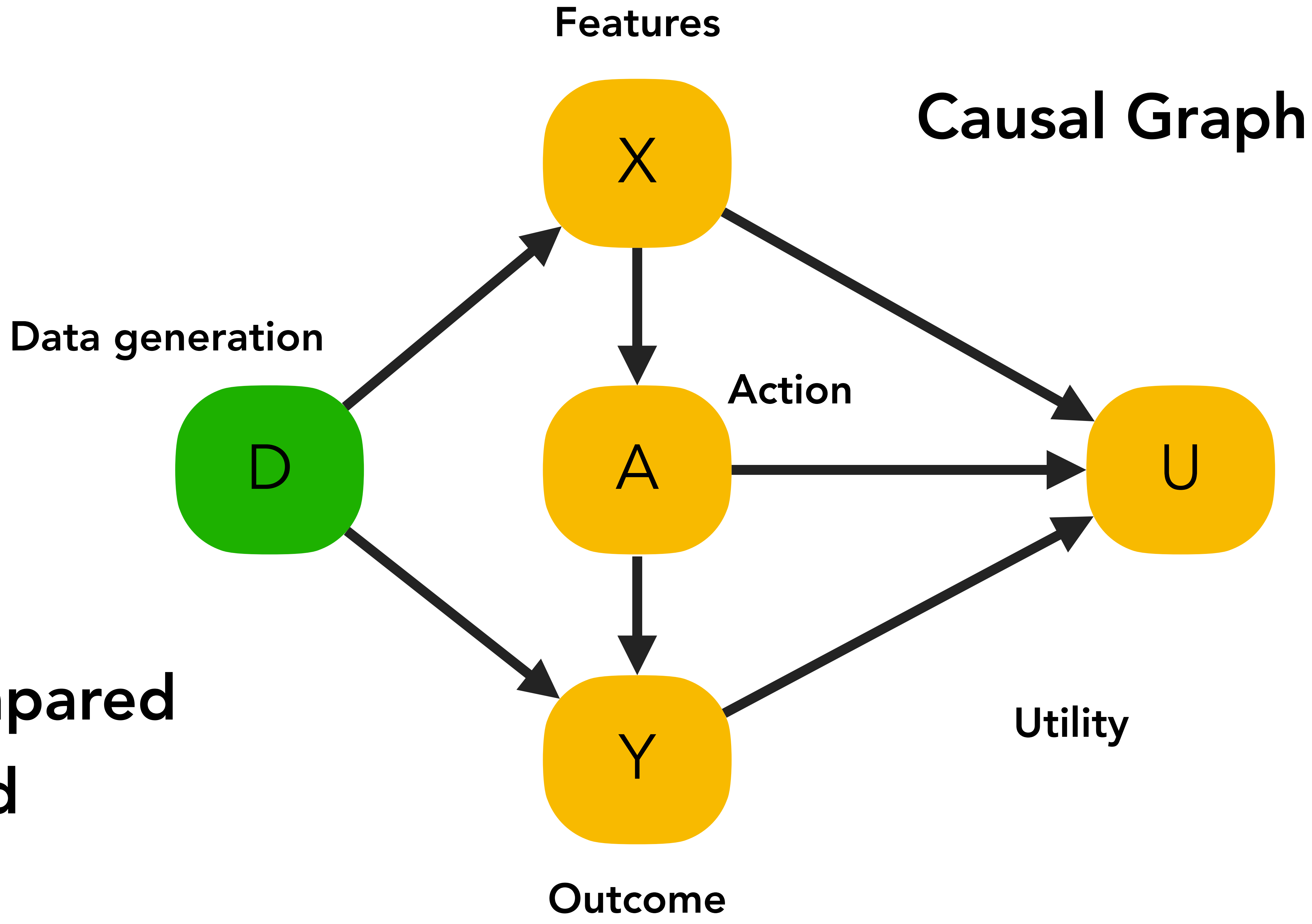
Structural Causal Model $\mathscr{M}$

$X := f_X(D) \qquad Y := f_Y(D) \qquad D \sim P(D)$

$A := \pi(X)$ ◀-- **Algorithmic policy** 🤖

$U := f_U(A, X, Y)$

Harm caused by action a taken by 🤖 compared to default action $\bar{a}$ given context $X = x$ and outcome $Y = y$



Richens et al. "*Counterfactual harm.*" NeurIPS, 2022.

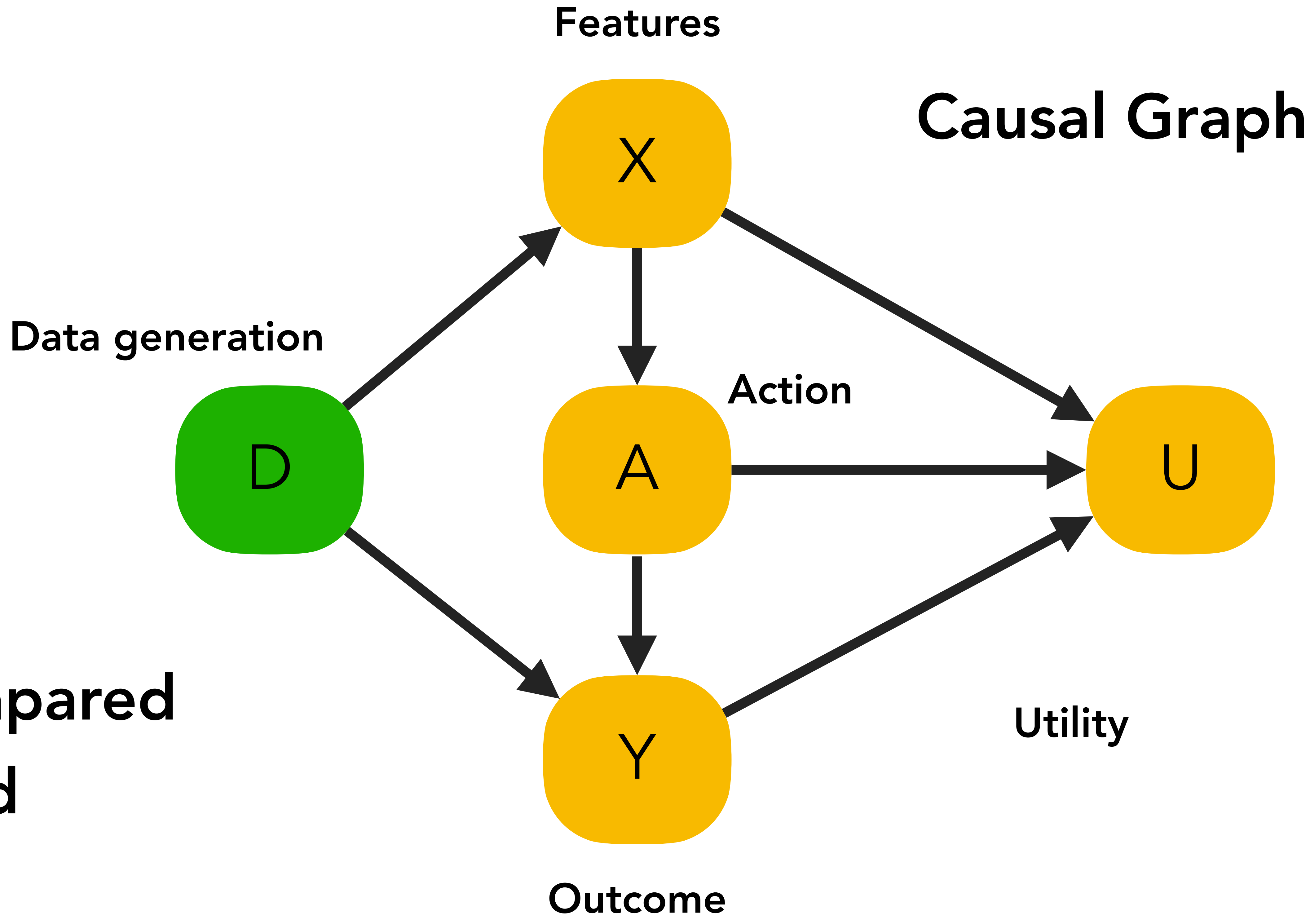
Formalizing counterfactual harm

Structural Causal Model $\mathscr{M}$

$X := f_X(D) \qquad Y := f_Y(D) \qquad D \sim P(D)$

$A := \pi(X)$ ◀- - **Algorithmic policy** 🤖

$U := f_U(A, X, Y)$



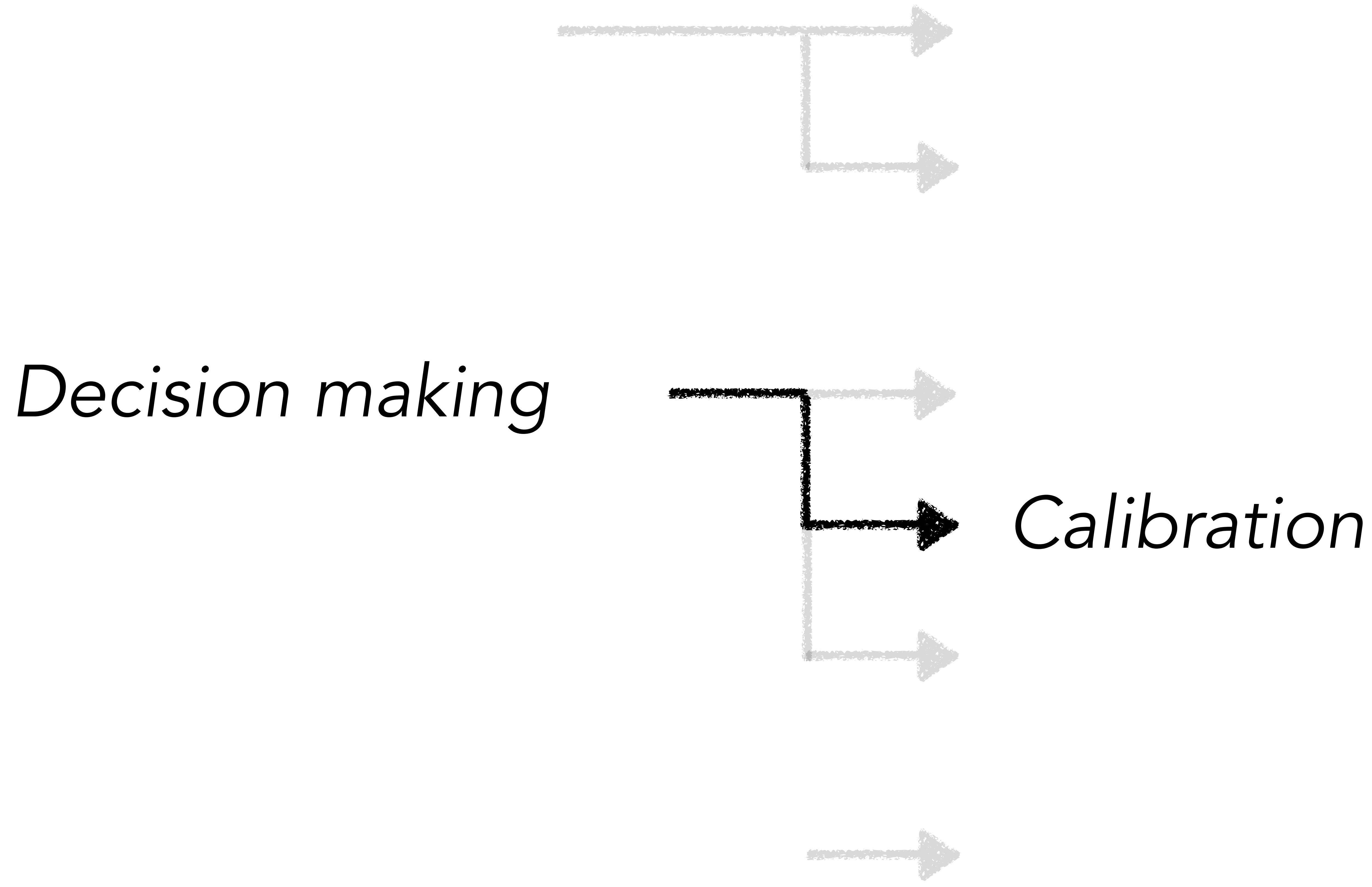
Causal Graph

Harm caused by action a taken by 🤖 compared to default action $\bar{a}$ given context $X = x$ and outcome $Y = y$

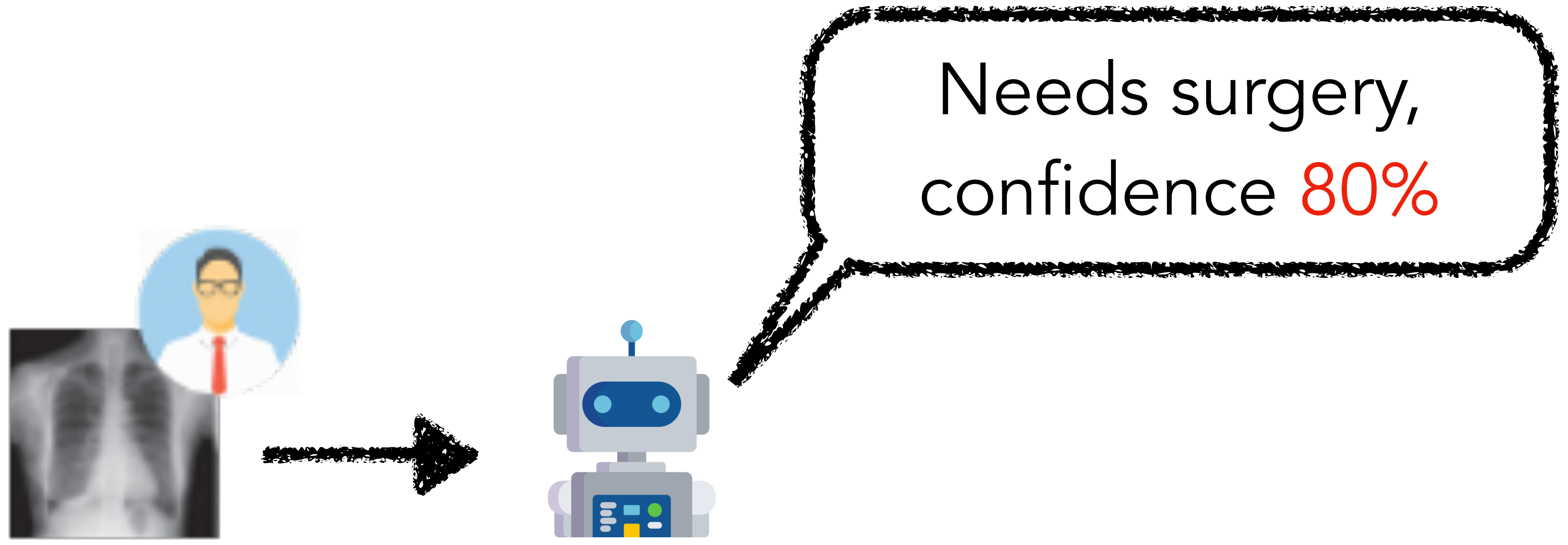
$$h(a, x, y) = \int_{y'} P^{\mathscr{M} \,|\, X=x,\, Y=y,\, A=a\,;\, do(A=\bar{a})}(Y = y') \max\left(0, \underbrace{U(\bar{a}, x, y')}_{\text{Counterfactual utility}} - \underbrace{U(a, x, y)}_{\text{Utility}}\right) dy'$$

Richens et al. "*Counterfactual harm.*" NeurIPS, 2022.

Use cases of counterfactuals in machine learning



Calibration

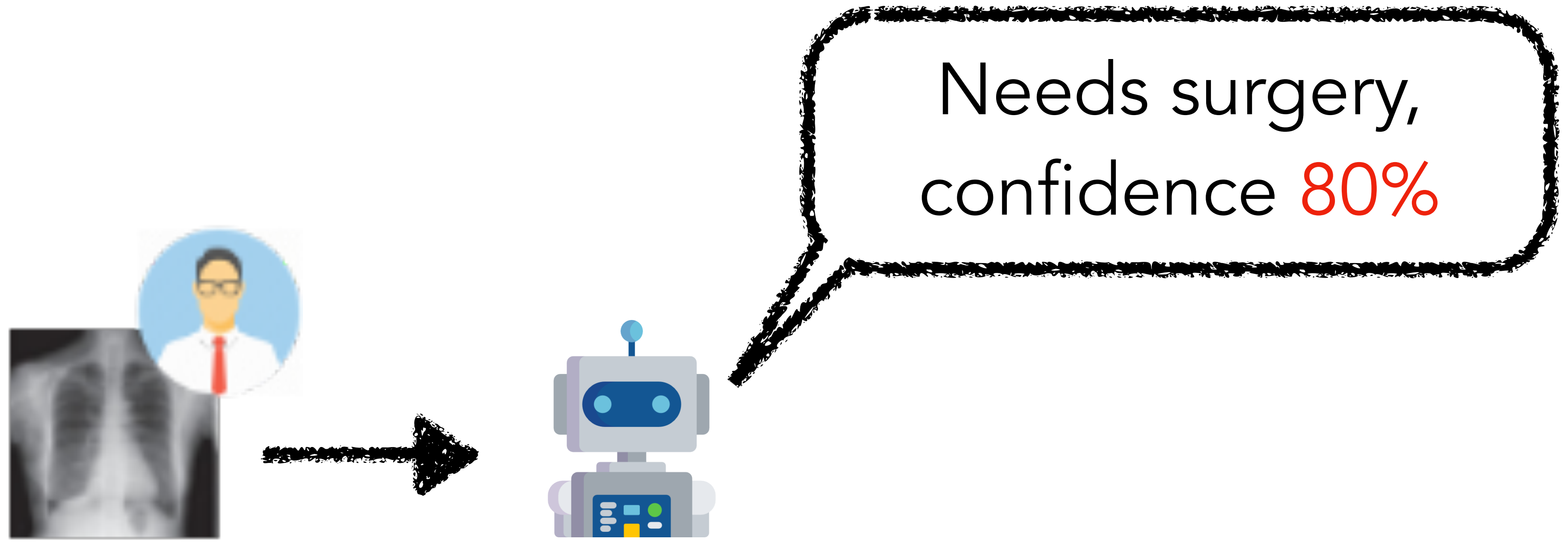


Needs surgery,
confidence 80%

Calibration:
Across all patients who 🤖 predicts there is a 80% chance they need surgery, it truly happens 80% of them needs surgery

Corvelo Benz and Gomez-Rodriguez. "*Human-aligned calibration for ai-assisted decision making.*" NeurIPS, 2023.

Calibration

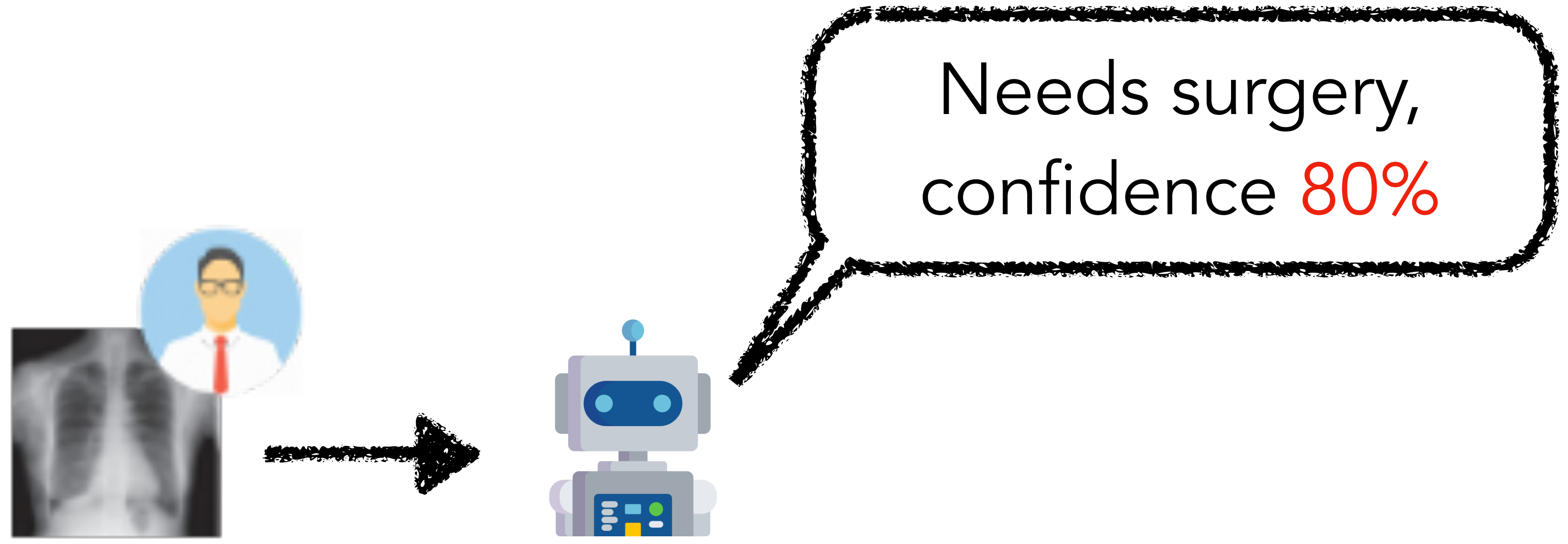


Calibration:
Across all patients who 🤖 predicts there is a 80% chance they need surgery, it truly happens 80% of them needs surgery

Counterfactual reasoning reveals that the way in which machine learning models compute confidence values today is problematic.

Corvelo Benz and Gomez-Rodriguez. "*Human-aligned calibration for ai-assisted decision making.*" NeurIPS, 2023.

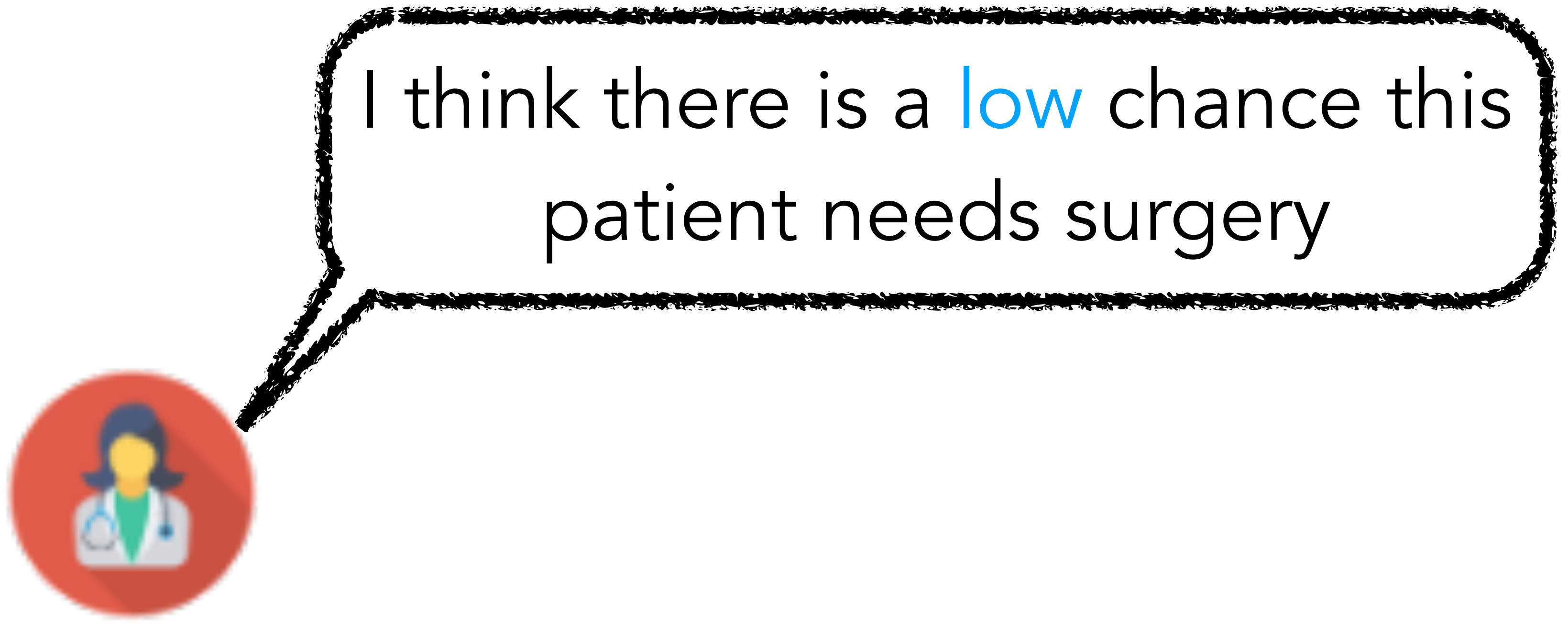
Calibration



Needs surgery, confidence 80%

Calibration:
Across all patients who 🤖 predicts there is a 80% chance they need surgery, it truly happens 80% of them needs surgery

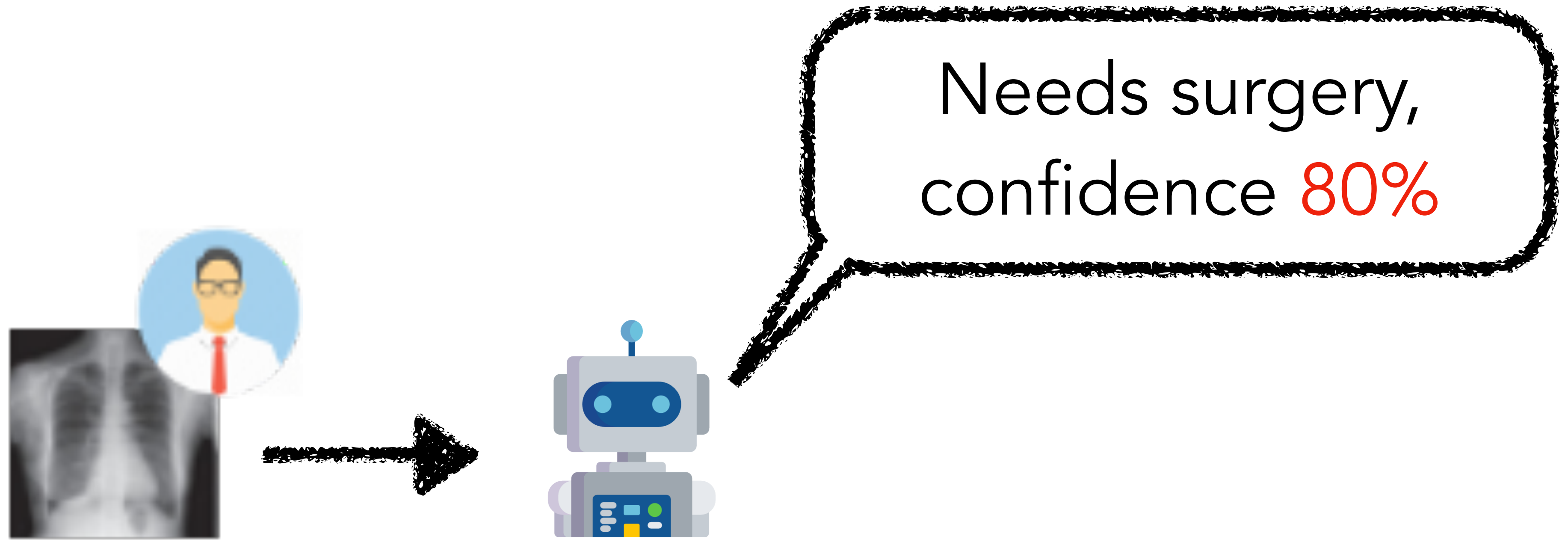
Counterfactual reasoning reveals that the way in which machine learning models compute confidence values today is problematic.



I think there is a low chance this patient needs surgery

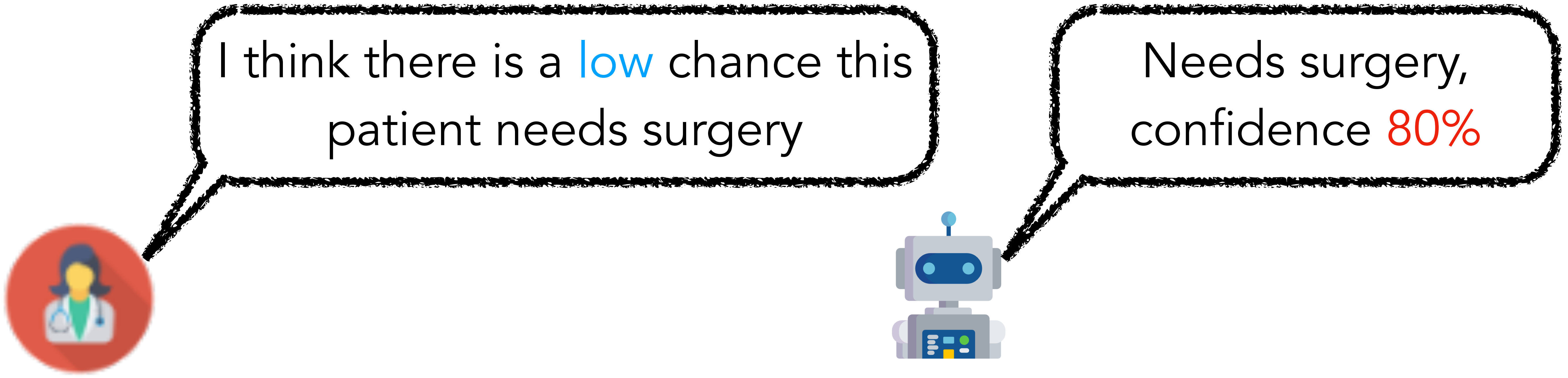
Corvelo Benz and Gomez-Rodriguez. "*Human-aligned calibration for ai-assisted decision making.*" NeurIPS, 2023.

Calibration



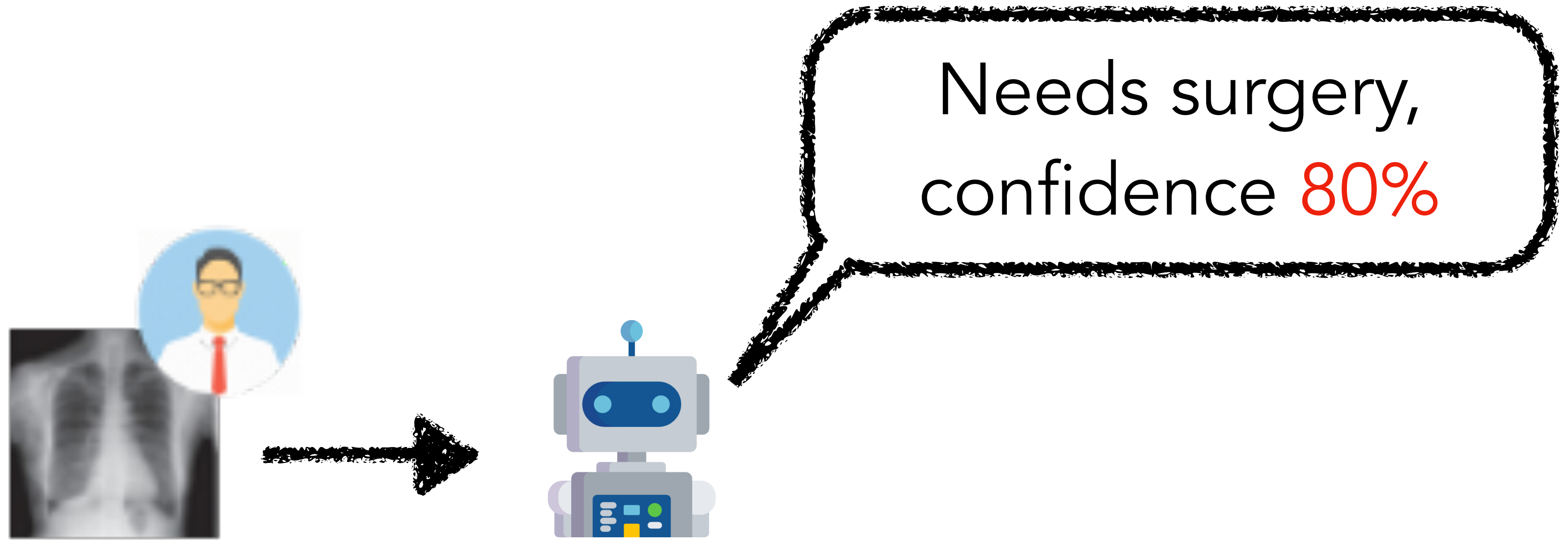
Calibration:
Across all patients who 🤖 predicts there is a 80% chance they need surgery, it truly happens 80% of them needs surgery

Counterfactual reasoning reveals that the way in which machine learning models compute confidence values today is problematic.


I think there is a low chance this patient needs surgery
Needs surgery, confidence 80%

Corvelo Benz and Gomez-Rodriguez. "*Human-aligned calibration for ai-assisted decision making.*" NeurIPS, 2023.

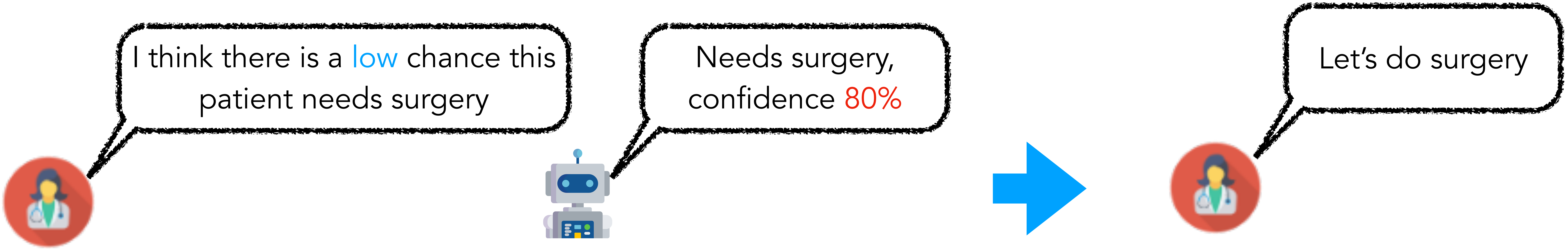
Calibration



Needs surgery, confidence 80%

Calibration:
Across all patients who 🤖 predicts there is a 80% chance they need surgery, it truly happens 80% of them needs surgery

Counterfactual reasoning reveals that the way in which machine learning models compute confidence values today is problematic.



I think there is a low chance this patient needs surgery

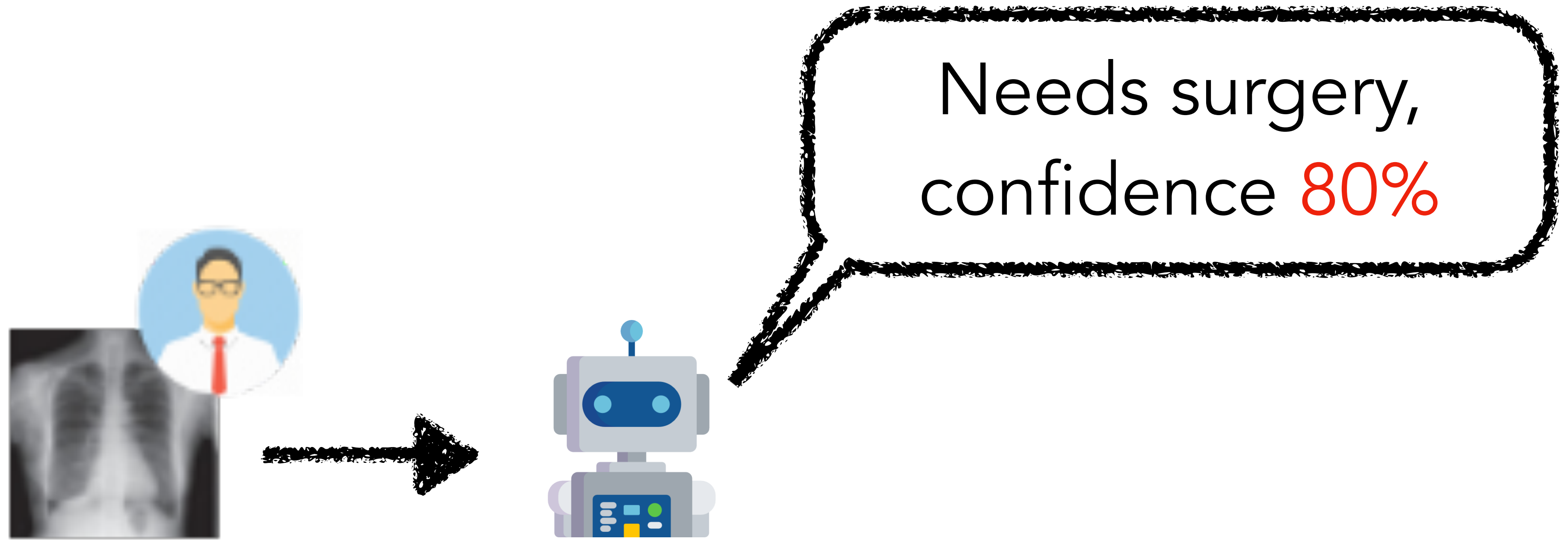
Needs surgery, confidence 80%

Let's do surgery

The doctor decides *optimally*

Corvelo Benz and Gomez-Rodriguez. "*Human-aligned calibration for ai-assisted decision making.*" NeurIPS, 2023.

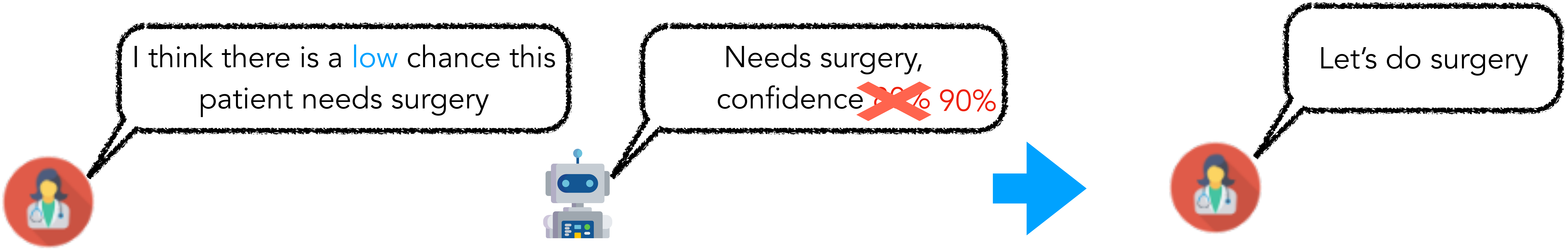
Calibration



Needs surgery, confidence 80%

Calibration:
Across all patients who 🤖 predicts there is a 80% chance they need surgery, it truly happens 80% of them needs surgery

Counterfactual reasoning reveals that the way in which machine learning models compute confidence values today is problematic.



I think there is a low chance this patient needs surgery

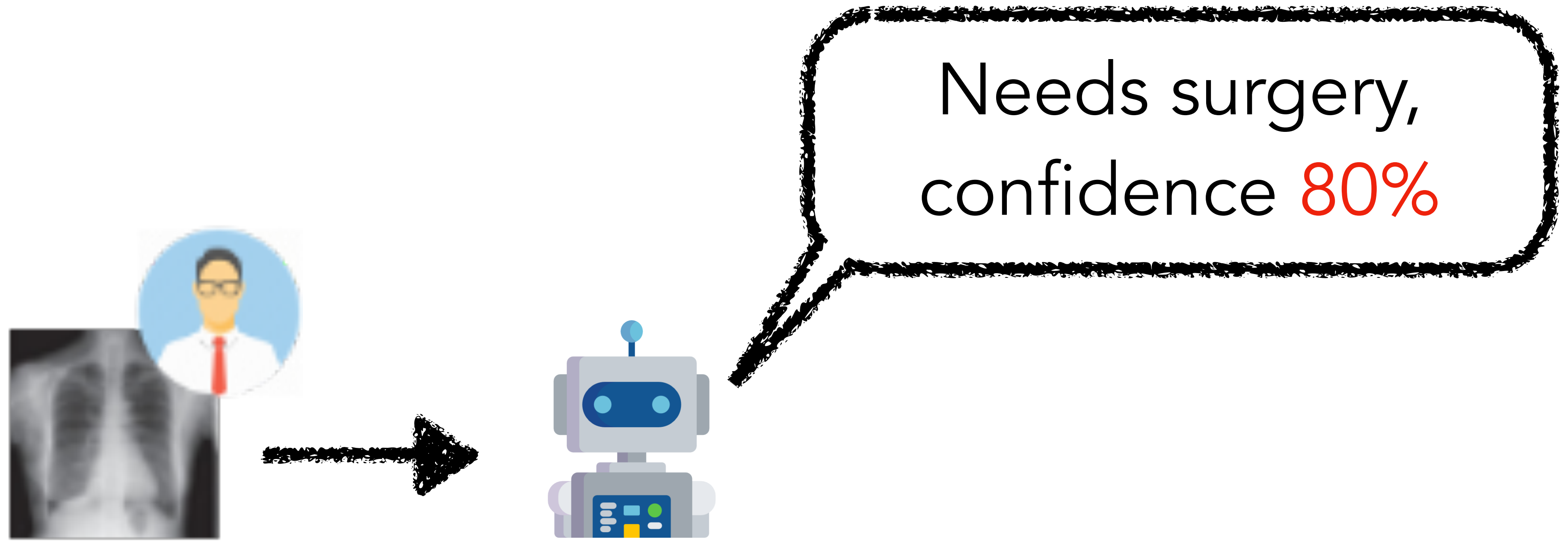
Needs surgery, confidence ~~80%~~ 90%

Let's do surgery

The doctor decides *optimally*

Corvelo Benz and Gomez-Rodriguez. "*Human-aligned calibration for ai-assisted decision making.*" NeurIPS, 2023.

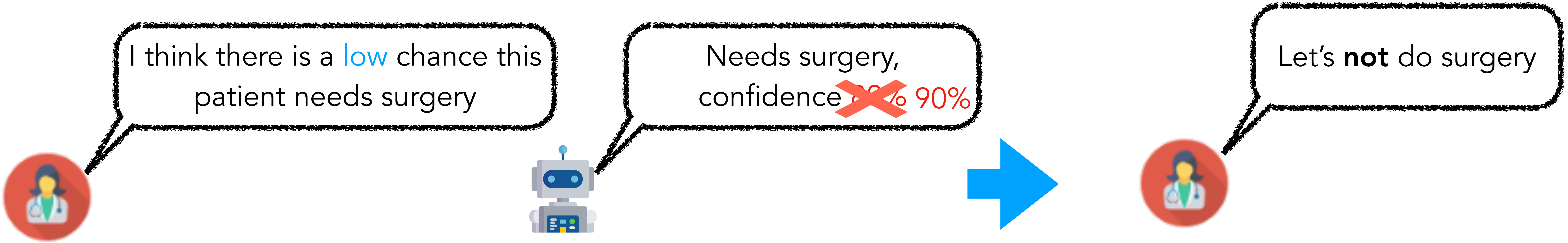
Calibration



Needs surgery, confidence 80%

Calibration:
Across all patients who 🤖 predicts there is a 80% chance they need surgery, it truly happens 80% of them needs surgery

Counterfactual reasoning reveals that the way in which machine learning models compute confidence values today is problematic



I think there is a low chance this patient needs surgery

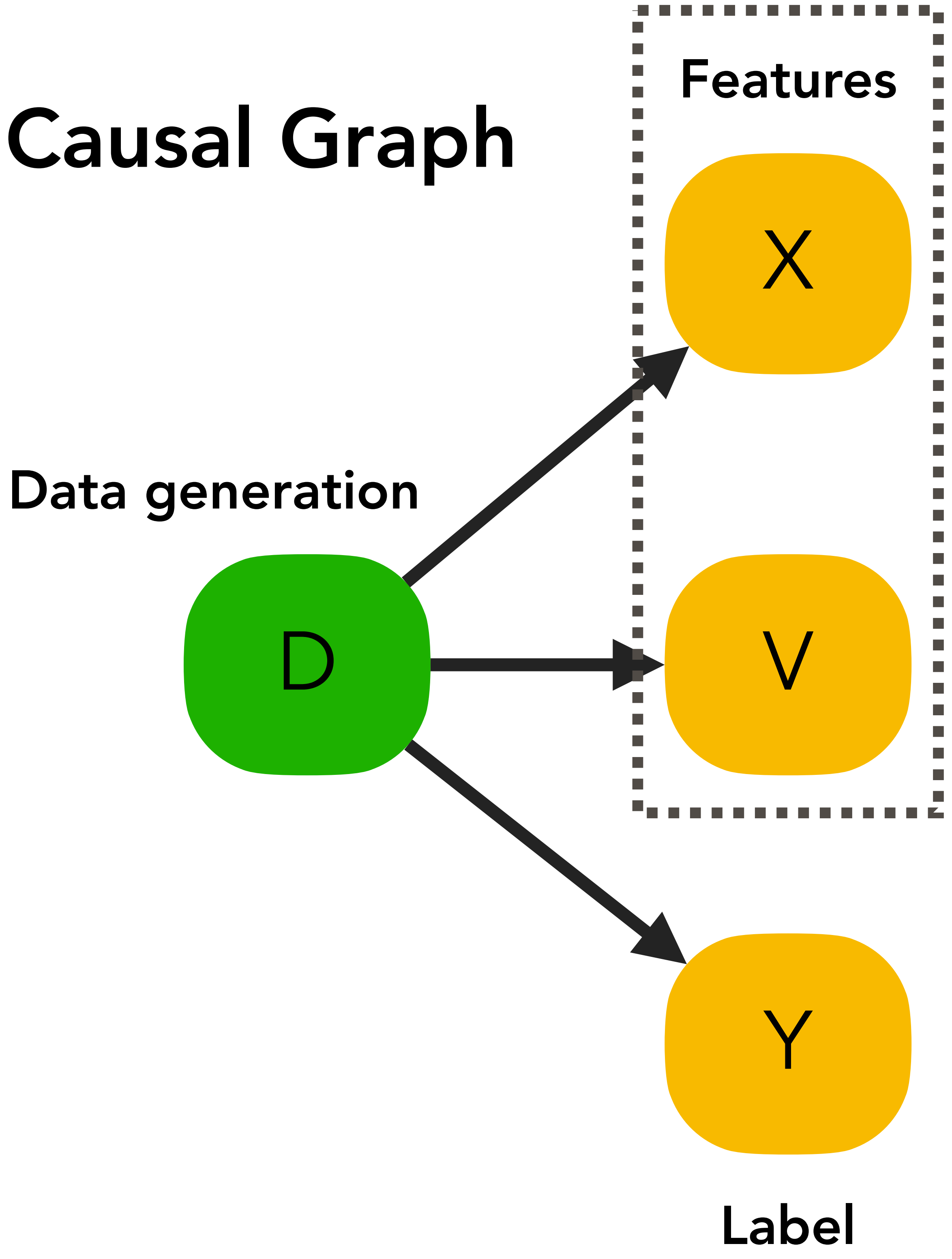
Needs surgery, confidence ~~80%~~ 90%

Let's **not** do surgery

The doctor decides *optimally*

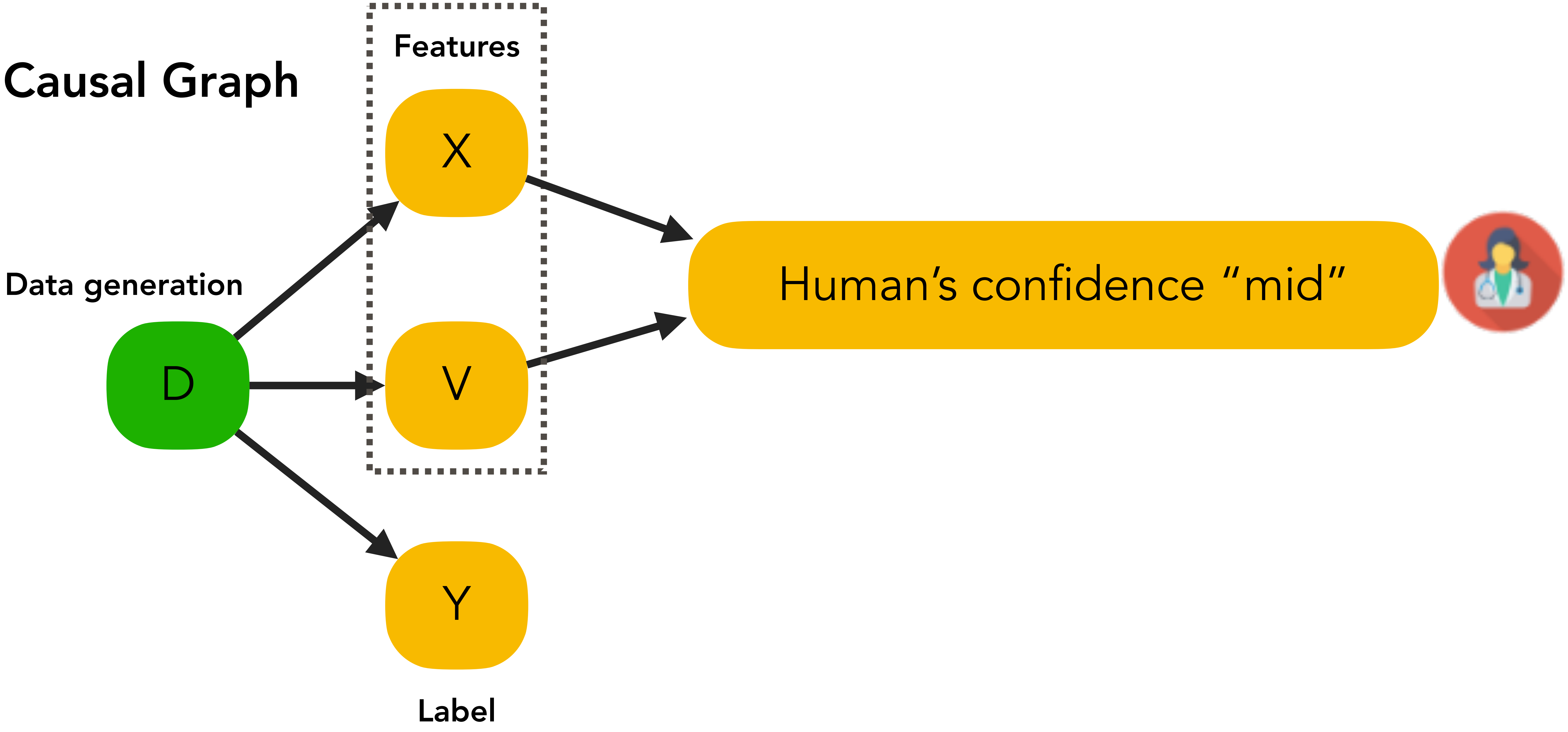
Corvelo Benz and Gomez-Rodriguez. "*Human-aligned calibration for ai-assisted decision making.*" NeurIPS, 2023.

Calibration



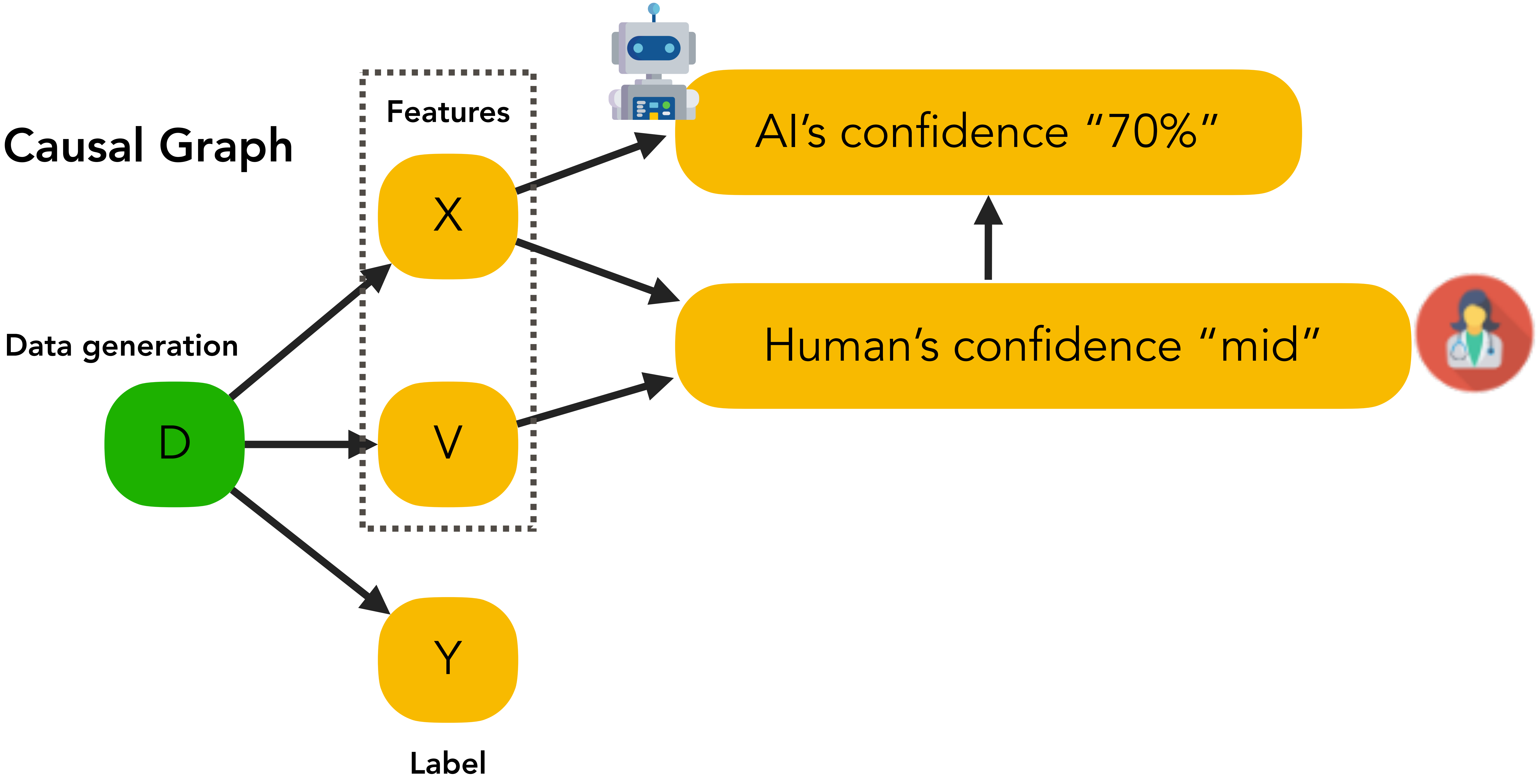
Corvelo Benz and Gomez-Rodriguez. "*Human-aligned calibration for ai-assisted decision making.*" NeurIPS, 2023.

Calibration



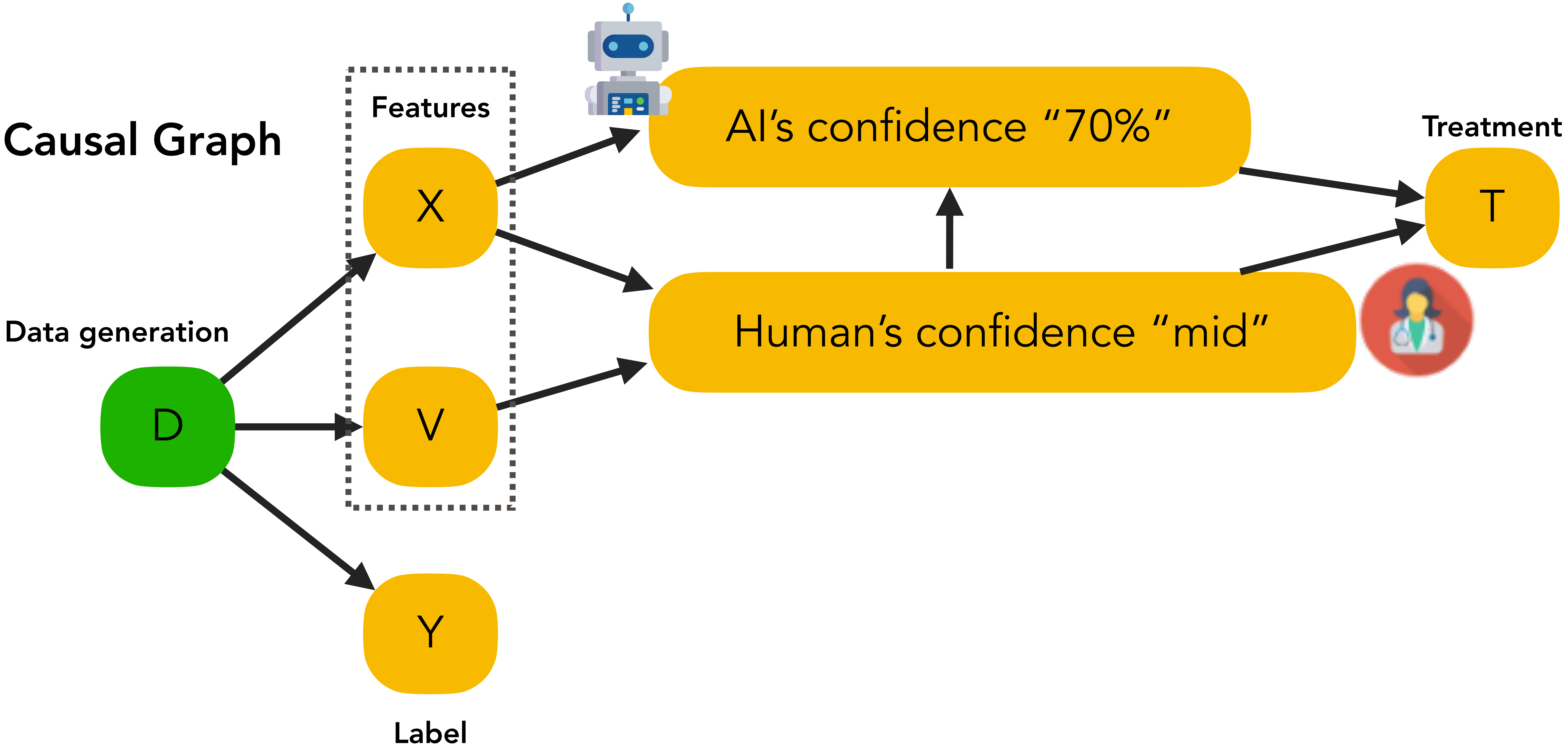
Corvelo Benz and Gomez-Rodriguez. "*Human-aligned calibration for ai-assisted decision making.*" NeurIPS, 2023.

Calibration



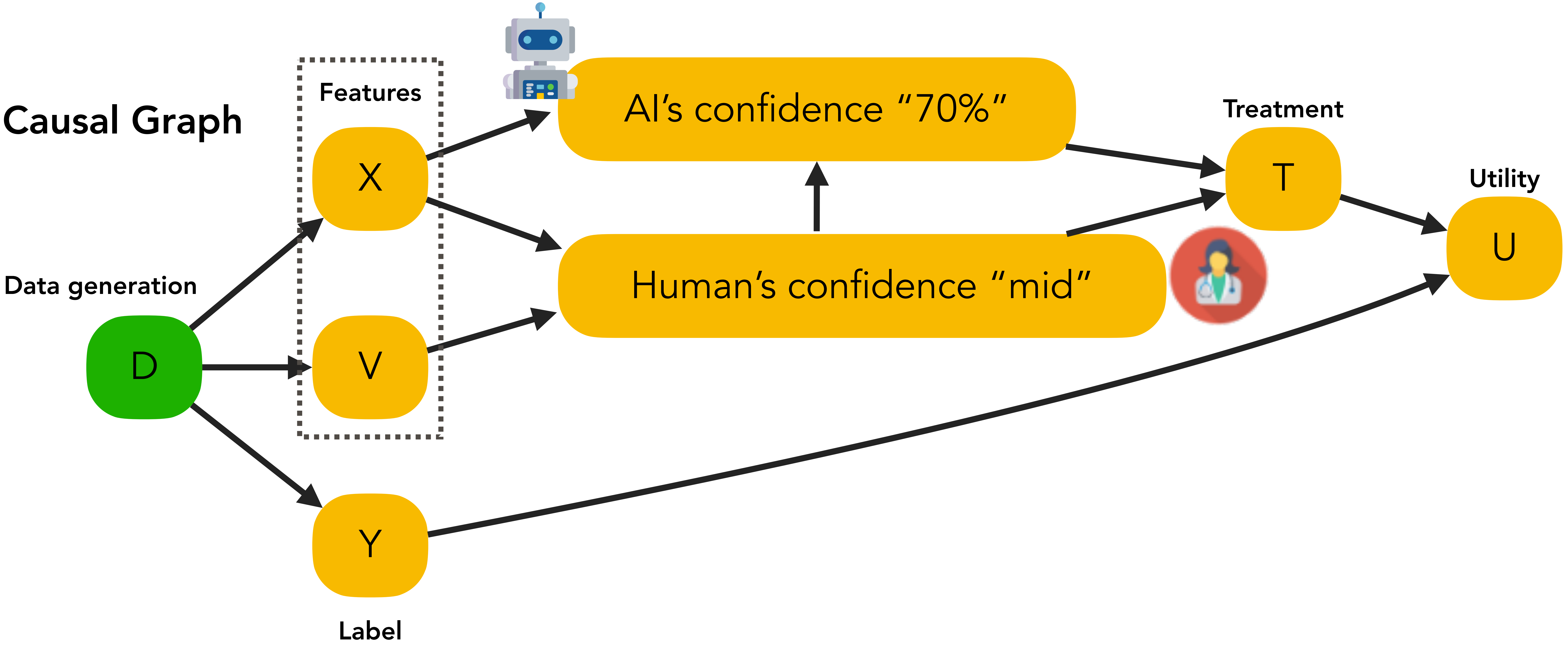
Corvelo Benz and Gomez-Rodriguez. "*Human-aligned calibration for ai-assisted decision making.*" NeurIPS, 2023.

Calibration



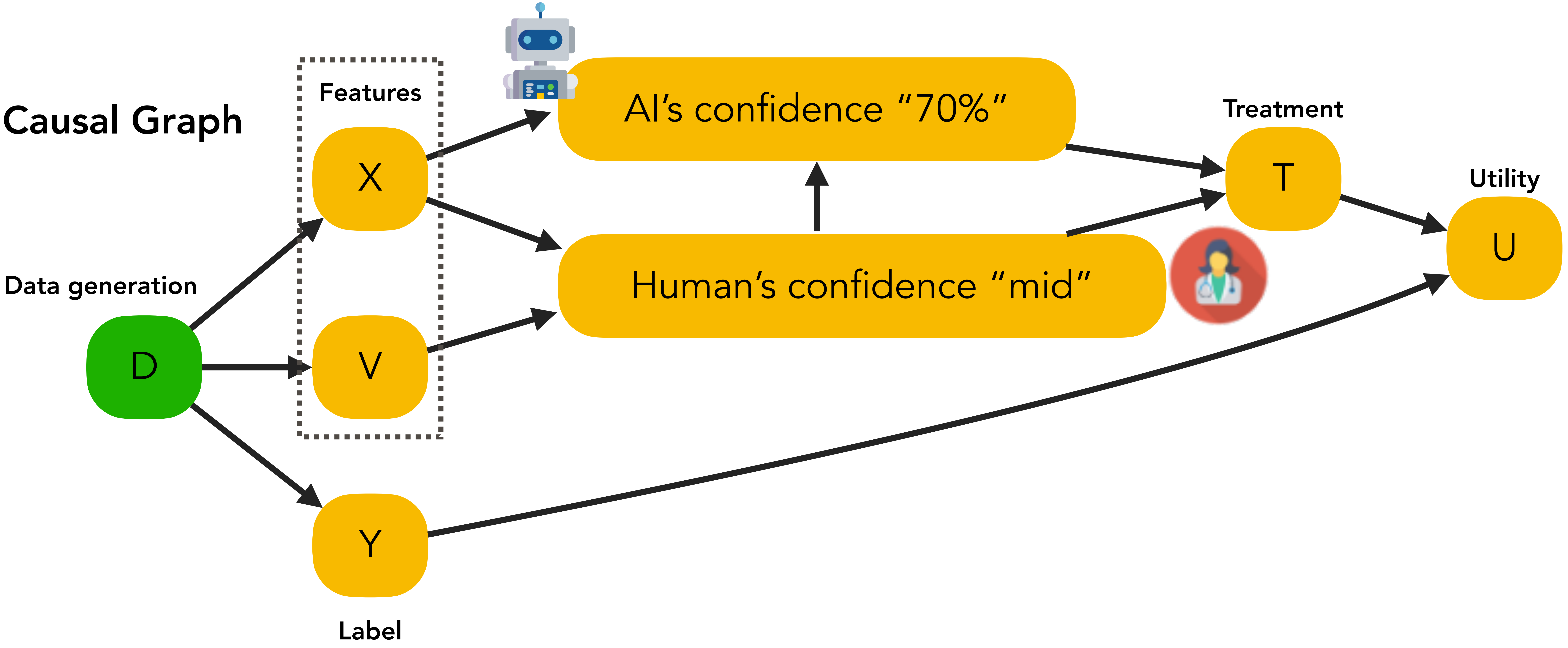
Corvelo Benz and Gomez-Rodriguez. "*Human-aligned calibration for ai-assisted decision making.*" NeurIPS, 2023.

Calibration



Corvelo Benz and Gomez-Rodriguez. "*Human-aligned calibration for ai-assisted decision making.*" NeurIPS, 2023.

Calibration

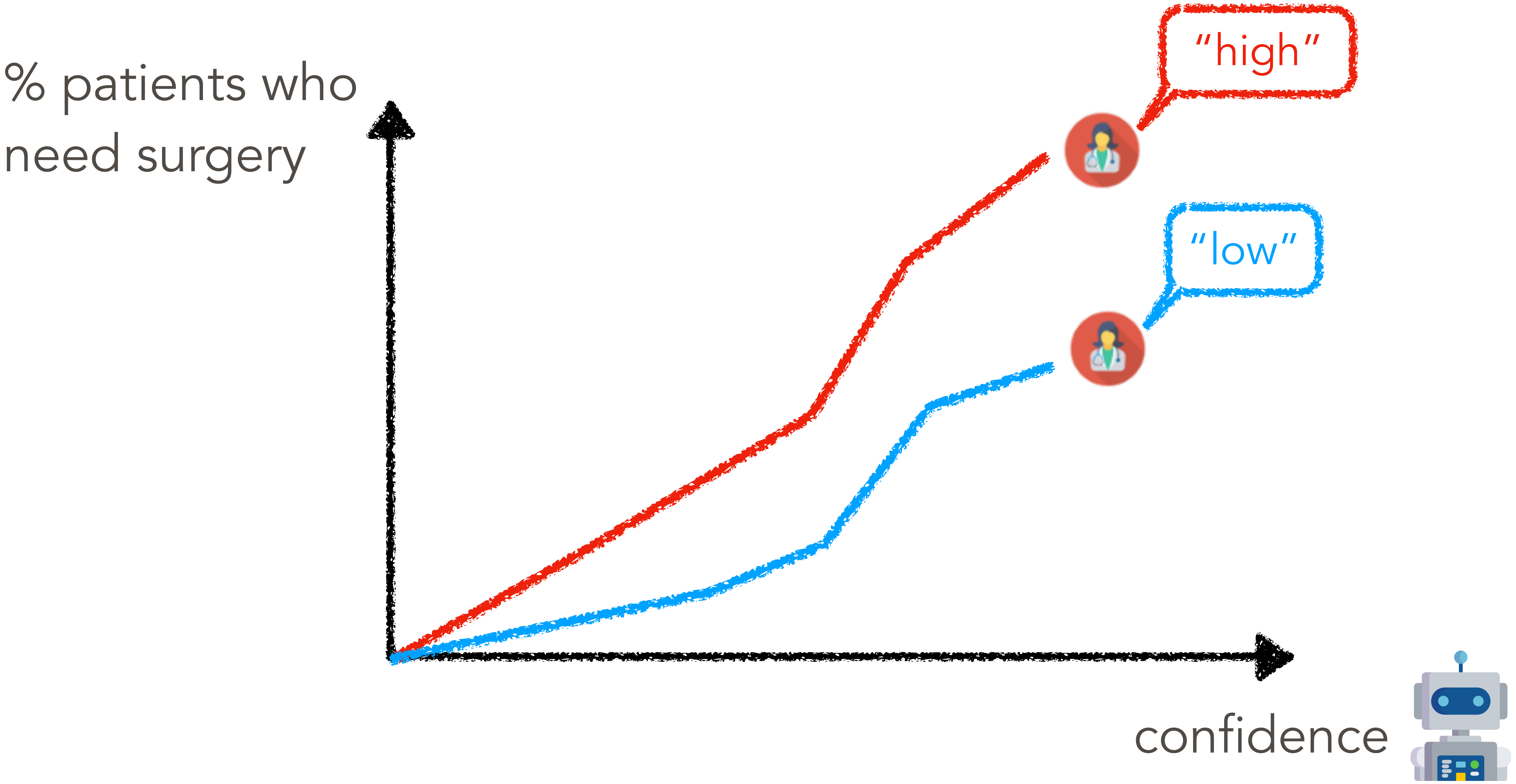


There exist instances of this decision making process in which any monotonic decision policy based on calibrated AI predictions is suboptimal.

Corvelo Benz and Gomez-Rodriguez. "*Human-aligned calibration for ai-assisted decision making.*" NeurIPS, 2023.

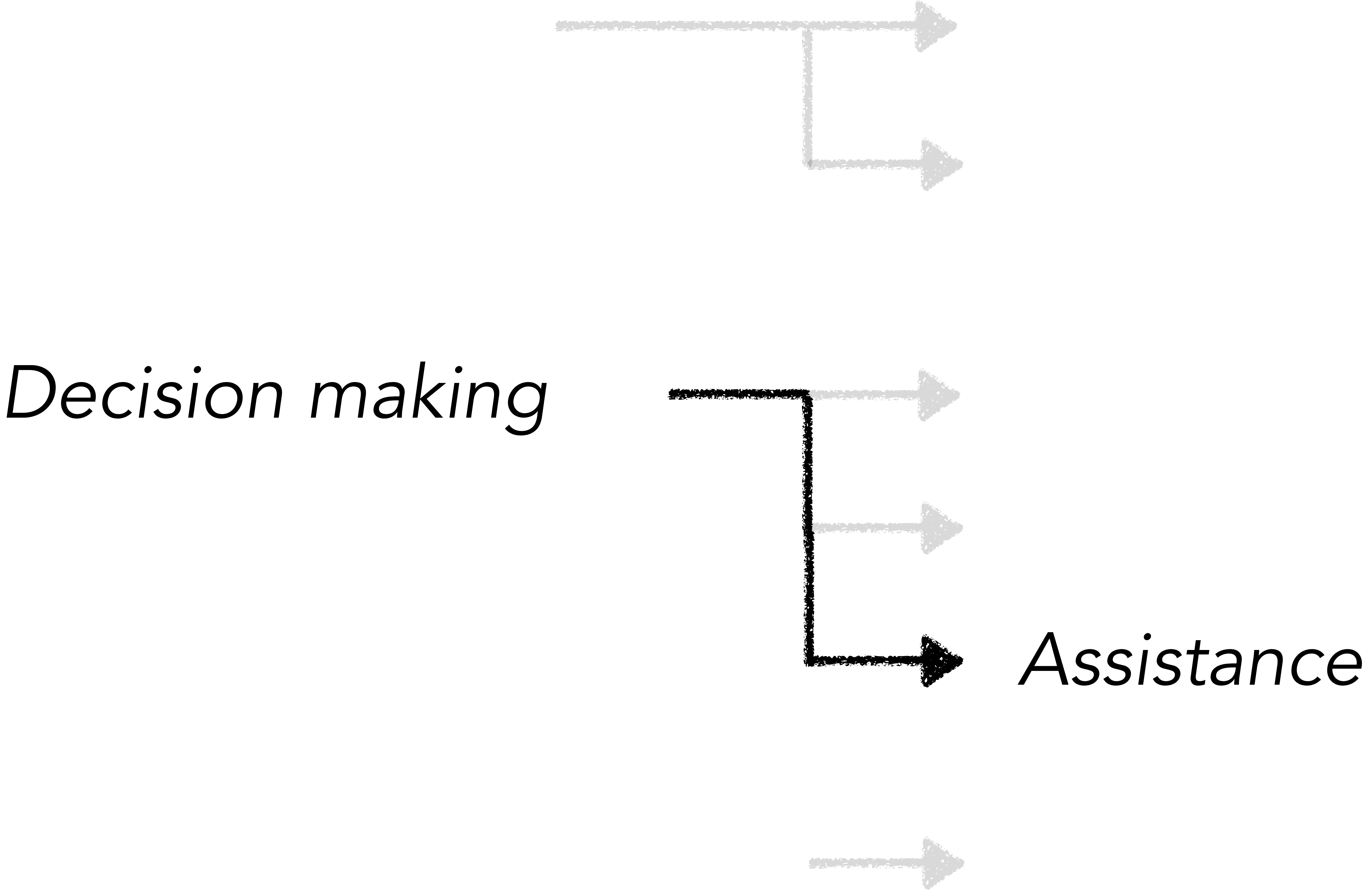
Calibration

To make sure the level of trust the optimal decision maker needs to place on predictions is (always) monotone on the confidence values, one can use **multicalibration.**



Corvelo Benz and Gomez-Rodriguez. "*Human-aligned calibration for ai-assisted decision making.*" NeurIPS, 2023.

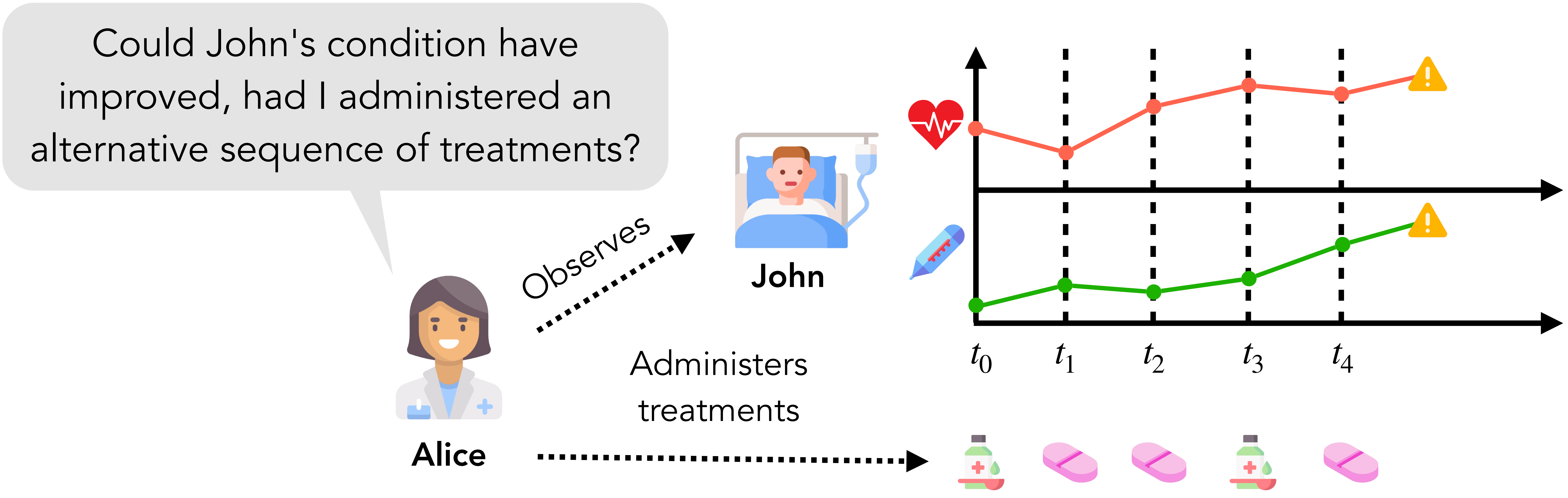
Use cases of counterfactuals in machine learning



Decision making

Assistance

AI-assisted counterfactuals in sequential decision making



Could John's condition have improved, had I administered an alternative sequence of treatments?

Tsirtsis et al. "*Counterfactual Explanations in Sequential Decision Making Under Uncertainty.*" NeurIPS, 2021.

Alternative sequence of treatments as counterfactuals

Structural Causal Model $\mathscr{M}$

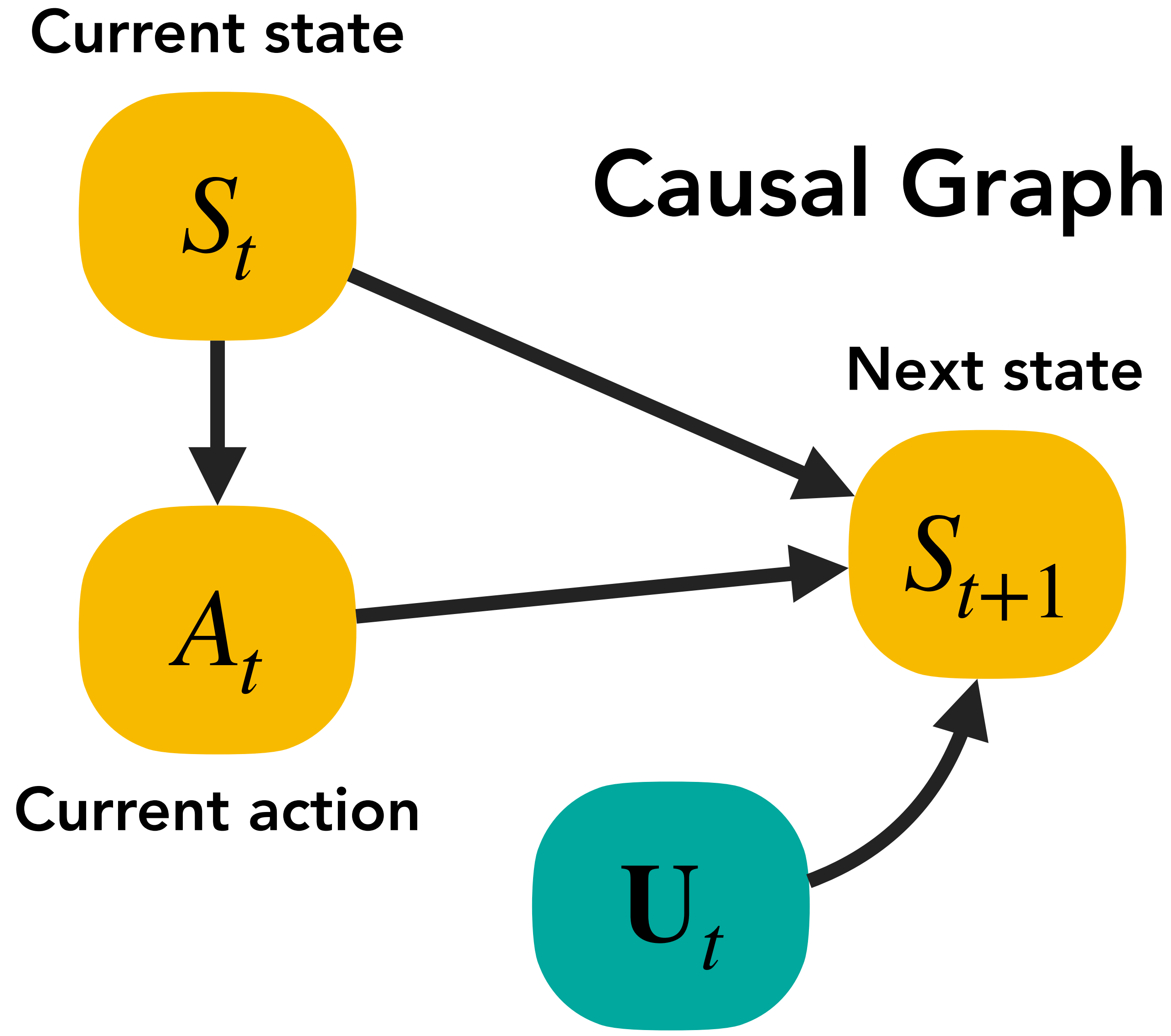
$$S_{t+1} := g_S(S_t, A_t, \mathbf{U}_t)$$

$$A_t := g_A(S_t, \mathbf{V}_t)$$

$$\mathbf{U}_t \sim P(\mathbf{U})$$

$$\mathbf{V}_t \sim P(\mathbf{V})$$

Causal Graph



...

Tsirtsis et al. "*Counterfactual Explanations in Sequential Decision Making Under Uncertainty.*" NeurIPS, 2021.

Alternative sequence of treatments as counterfactuals

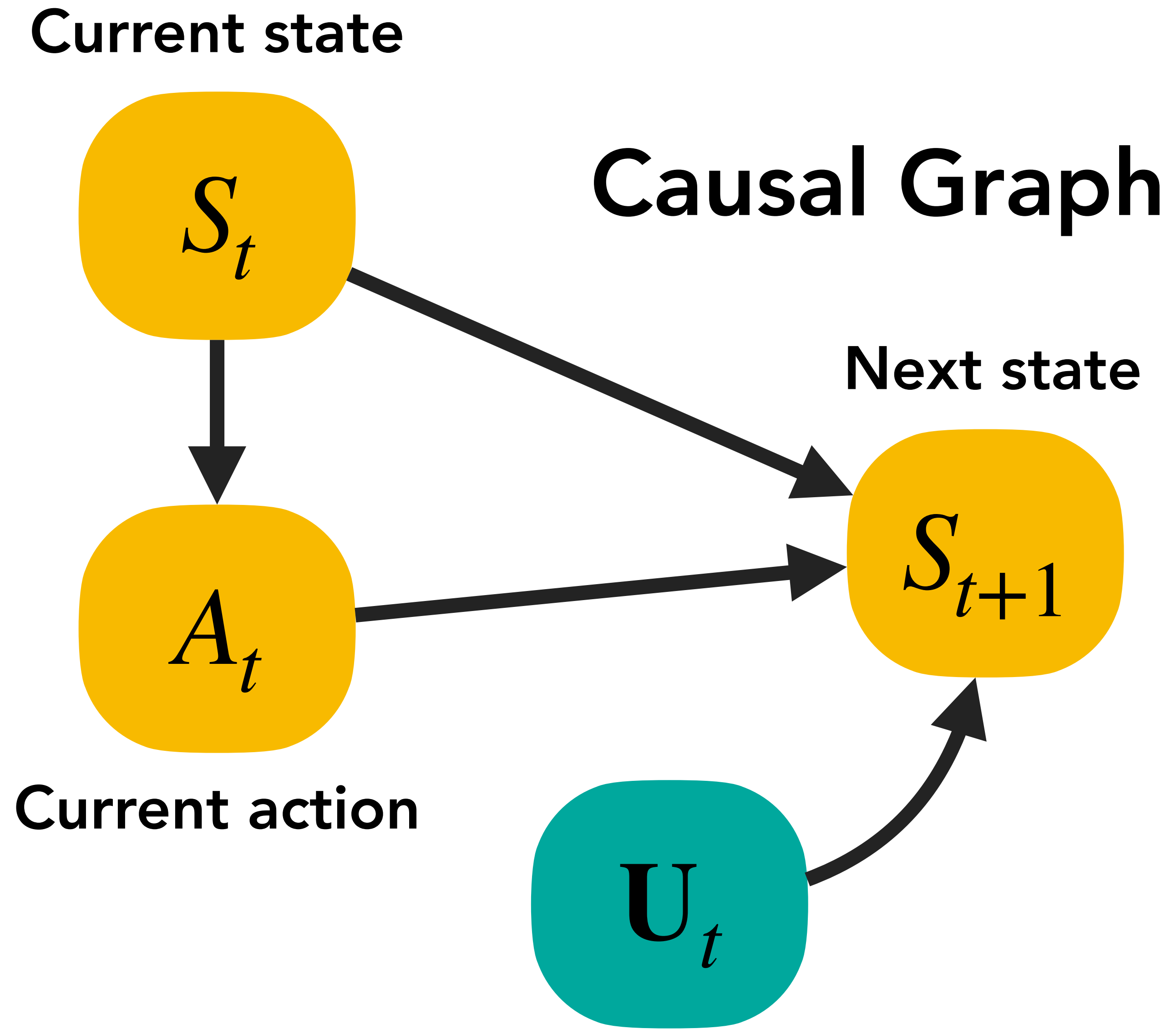
Structural Causal Model $\mathscr{M}$

$$S_{t+1} := g_S(S_t, A_t, \mathbf{U}_t)$$

$$A_t := g_A(S_t, \mathbf{V}_t)$$

$$\mathbf{U}_t \sim P(\mathbf{U})$$

$$\mathbf{V}_t \sim P(\mathbf{V})$$



At state $S_t = s_t$, the doctor took action $A_t = a_t$, what would have happened had the doctor taken action $a' \neq a_t$?

...

Tsirtsis et al. "*Counterfactual Explanations in Sequential Decision Making Under Uncertainty.*" NeurIPS, 2021.

Alternative sequence of treatments as counterfactuals

Modified Structural Causal Model $\mathscr{M}_{\{S_t=s_t,\, A_t=a_t\}}$

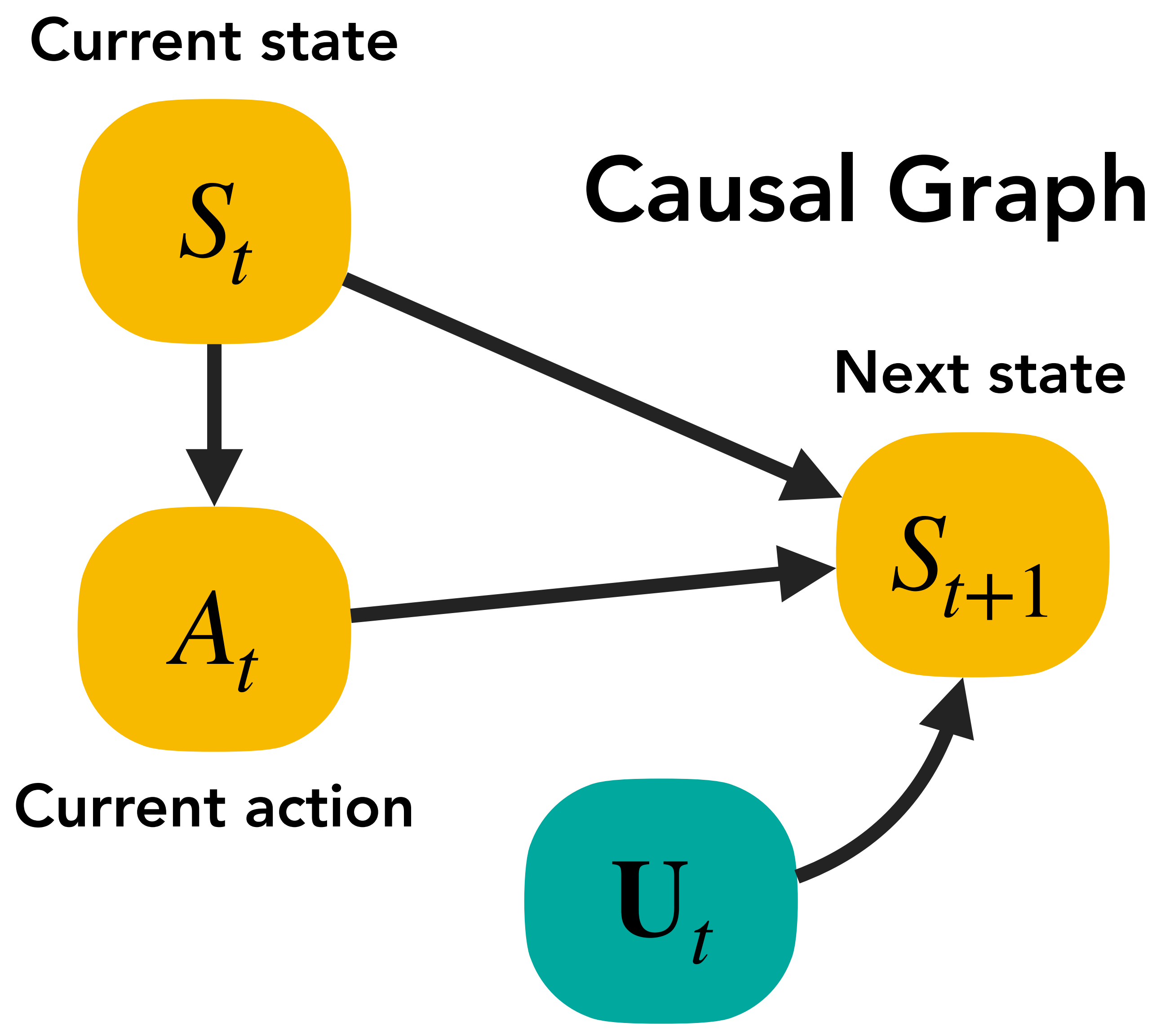
$$S_{t+1} := g_S(S_t, A_t, \mathbf{U}_t)$$

$$A_t := g_A(S_t, \mathbf{V}_t)$$

$$\mathbf{U}_t \sim P(\mathbf{U} \mid S_t = s_t, A_t = a_t)$$

$$\mathbf{V}_t \sim P(\mathbf{V} \mid S_t = s_t)$$

Posterior distribution of the noises



At state $S_t = s_t$, the doctor took action $A_t = a_t$, what would have happened had the doctor taken action $a' \neq a_t$?

...

Tsirtsis et al. "*Counterfactual Explanations in Sequential Decision Making Under Uncertainty.*" NeurIPS, 2021.

Alternative sequence of treatments as counterfactuals

Modified Structural Causal Model $\mathscr{M}_{\{S_t = s_t,\, A_t = a_t\}}$

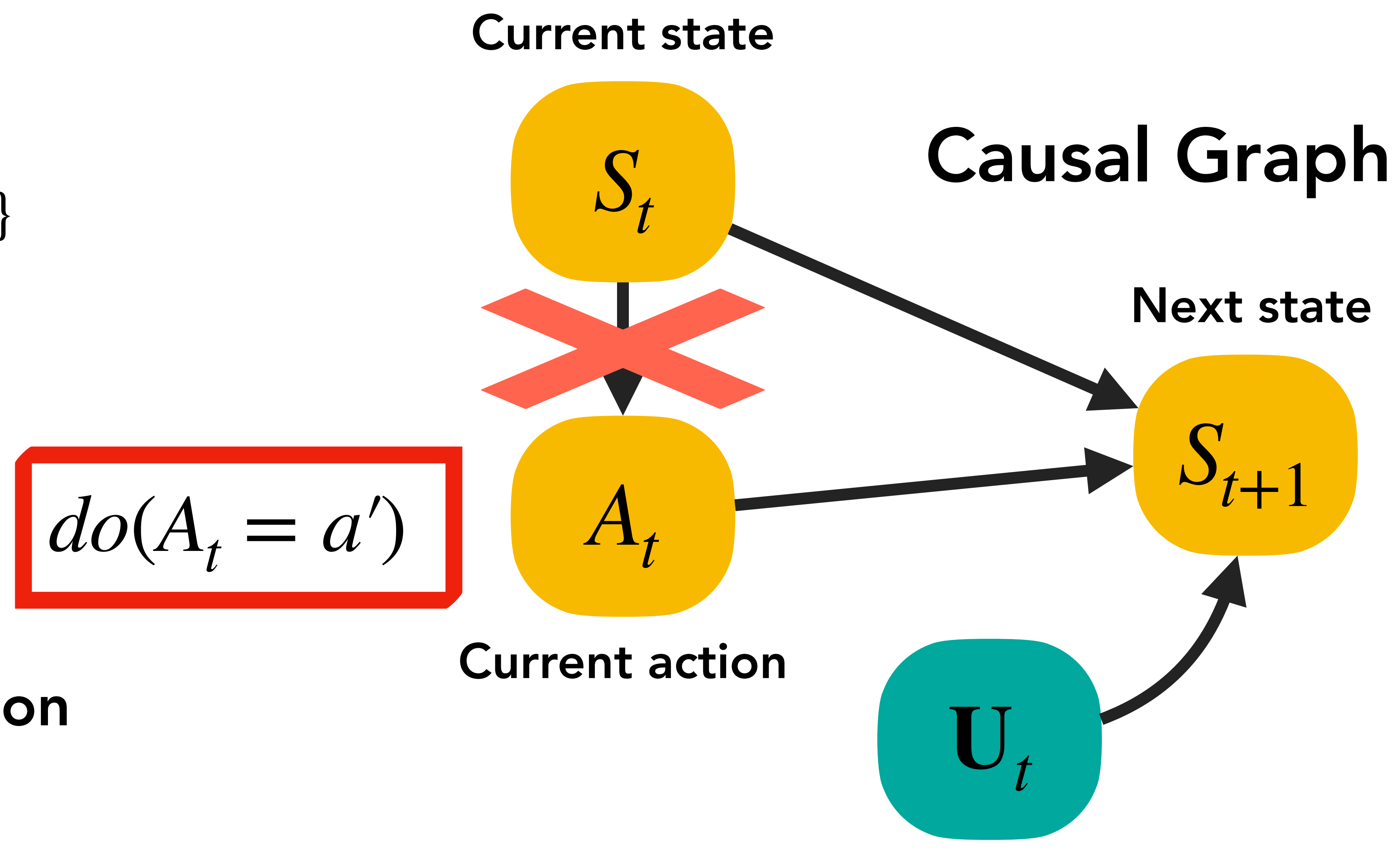
$$S_{t+1} := g_S(S_t, A_t, \mathbf{U}_t)$$

$$A_t := \cancel{g_A(S_t,} \mathbf{V}_t) \qquad A_t := a'$$

$$\mathbf{U}_t \sim P(\mathbf{U} \mid S_t = s_t, A_t = a_t)$$

$$\mathbf{V}_t \sim P(\mathbf{V} \mid S_t = s_t)$$

Posterior distribution of the noises



At state $S_t = s_t$, the doctor took action $A_t = a_t$, what would have happened had the doctor taken action $a' \neq a_t$?

...

Tsirtsis et al. "*Counterfactual Explanations in Sequential Decision Making Under Uncertainty.*" NeurIPS, 2021.

Alternative sequence of treatments as counterfactuals

Modified Structural Causal Model $\mathscr{M}_{\{S_t=s_t,\, A_t=a_t\}}$

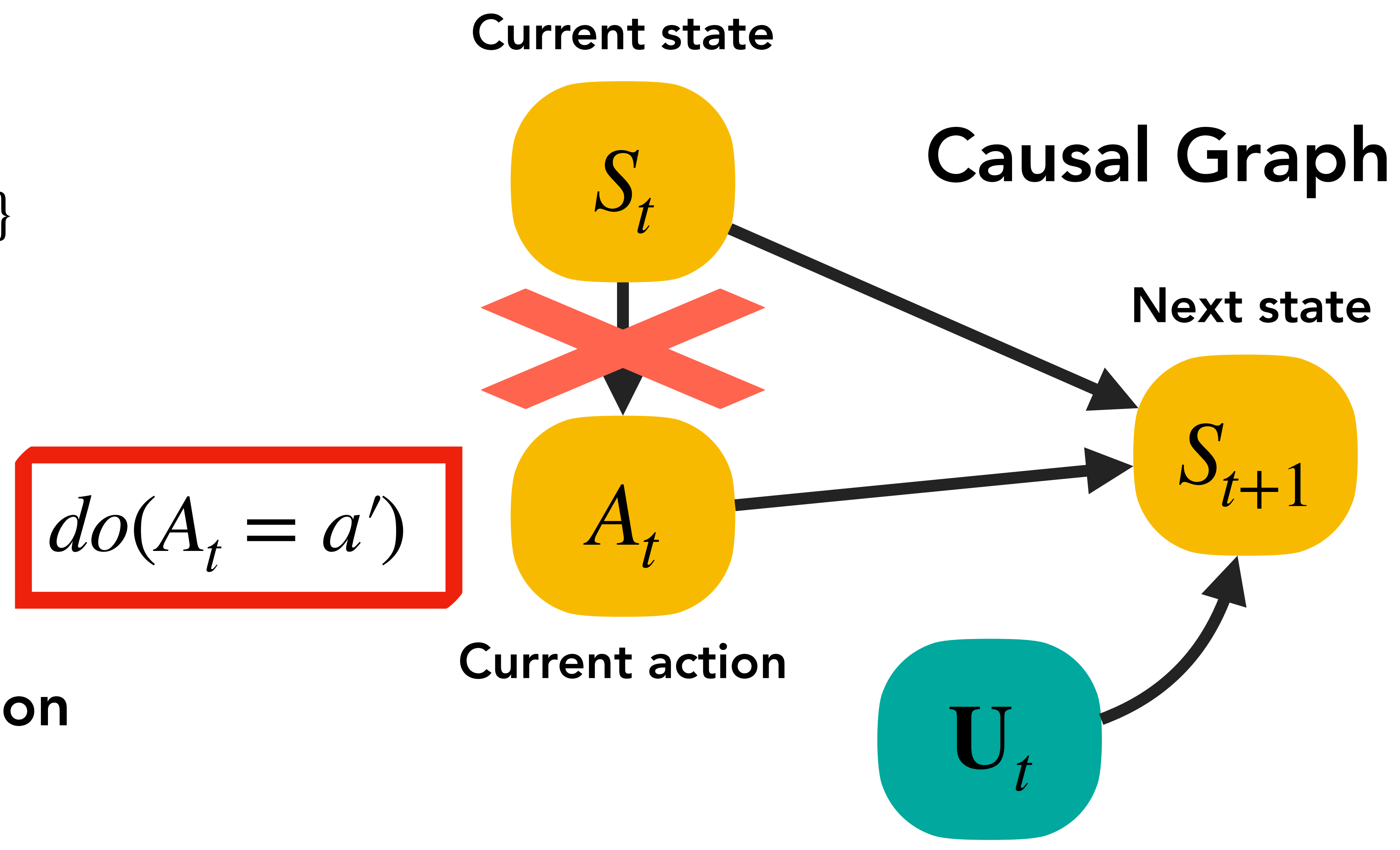
$$S_{t+1} := g_S(S_t, A_t, \mathbf{U}_t)$$

$$A_t \cancel{:= g_A(S_t,} \mathbf{V}_t) \qquad A_t := a'$$

$$\mathbf{U}_t \sim P(\mathbf{U} \mid S_t = s_t, A_t = a_t)$$

$$\mathbf{V}_t \sim P(\mathbf{V} \mid S_t = s_t)$$

Posterior distribution of the noises



At state $S_t = s_t$, the doctor took action $A_t = a_t$, what would have happened had the doctor taken action $a' \neq a_t$?

$$S_{t+1} \sim P^{\mathscr{M} \mid S_t=s_t,\, A_t=a_t\,;\, do(A_t=a')}\left(S_{t+1}\right)$$

...

Tsirtsis et al. "*Counterfactual Explanations in Sequential Decision Making Under Uncertainty.*" NeurIPS, 2021.

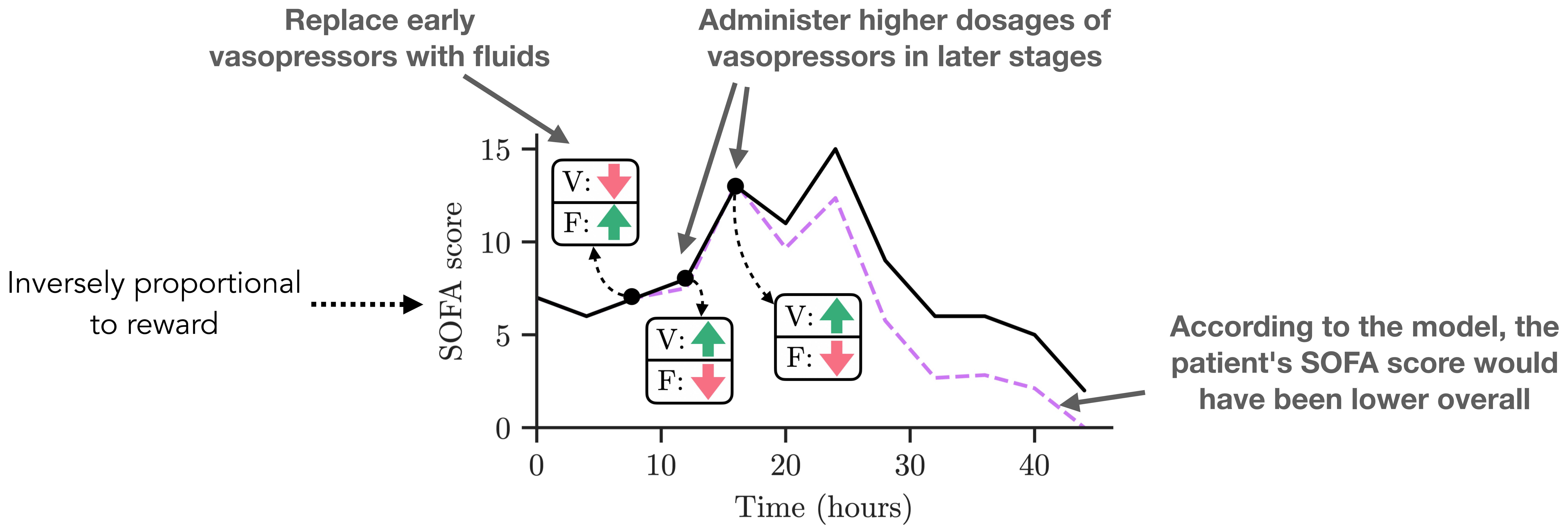
Counterfactually optimal action sequences

Given the counterfactual transition probabilities $S_{t+1} \sim P^{\mathscr{M} \mid S_t = s_t, A_t = a_t \,;\, do(A_t = a')}\left(S_{t+1}\right)$ and a reward function $r(s, a)$, one may find alternative sequence of actions $a'_1, \ldots, a'_{T-1}$ close to the observed actions $a_1, \ldots, a_{T-1}$ that maximizes the average counterfactual reward.

Tsirtsis et al. "*Counterfactual Explanations in Sequential Decision Making Under Uncertainty.*" NeurIPS, 2021.

Counterfactually optimal action sequences

Given the counterfactual transition probabilities $S_{t+1} \sim P^{\mathcal{M} \mid S_t=s_t, A_t=a_t \,;\, do(A_t=a')}\left(S_{t+1}\right)$ and a reward function $r(s, a)$, one may find alternative sequence of actions $a'_1, \ldots, a'_{T-1}$ close to the observed actions $a_1, \ldots, a_{T-1}$ that maximizes the average counterfactual reward.



Replace early vasopressors with fluids

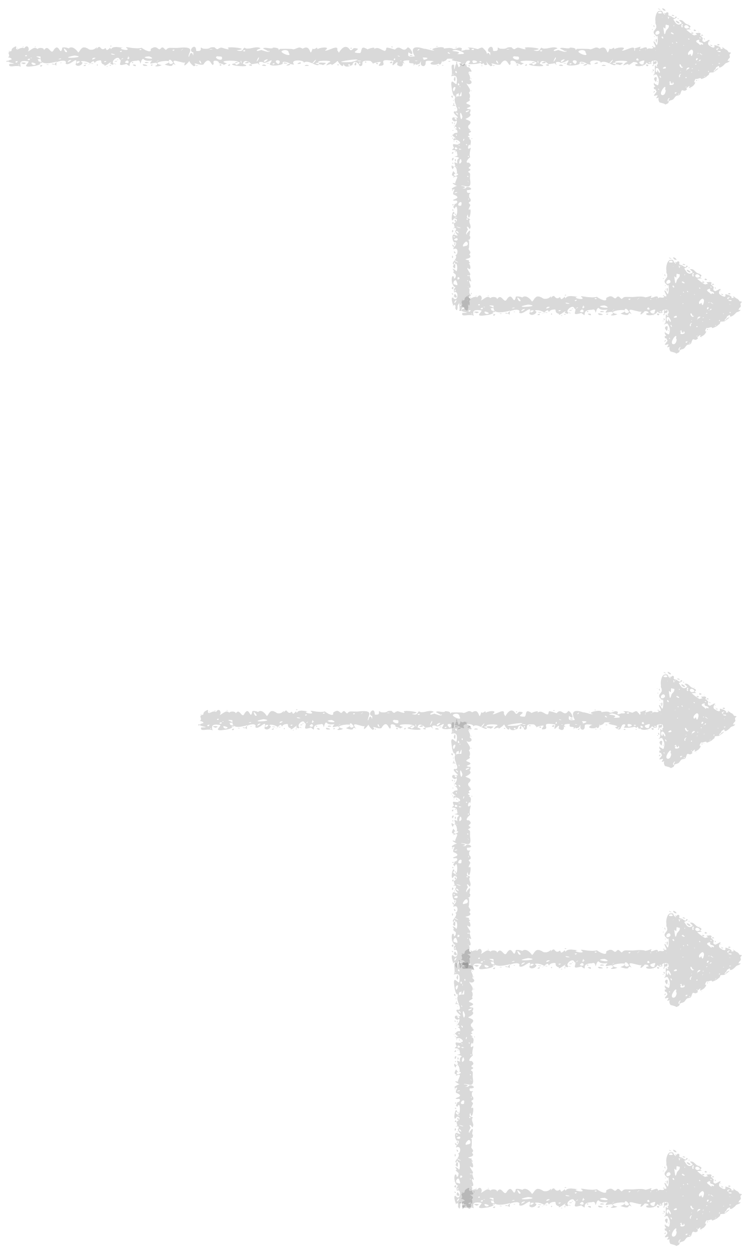
Administer higher dosages of vasopressors in later stages

Inversely proportional to reward

According to the model, the patient's SOFA score would have been lower overall

Tsirtsis et al. “*Counterfactual Explanations in Sequential Decision Making Under Uncertainty.*” NeurIPS, 2021.

Use cases of counterfactuals in machine learning



Reinforcement learning → *Training*

Counterfactually-guided training in reinforcement learning

In reinforcement learning, given a transition probability $P(s' \mid s, a)$ and a reward function $r(s, a)$, the goal is to design an action policy $a := \pi(s)$ with the highest average reward, i.e.

$$\pi^* = \text{argmax}_\pi \mathbb{E}_{\tau \sim \pi, P}\left[R(\tau)\right] \quad \text{where} \quad R(\tau) = \sum_{t=1}^{T} R(s_t, a_t)$$

Buesing et al. "*Woulda, coulda, shoulda: Counterfactually-guided policy search.*" ICLR, 2018.

Counterfactually-guided training in reinforcement learning

In reinforcement learning, given a transition probability $P(s' \mid s, a)$ and a reward function $r(s, a)$, the goal is to design an action policy $a := \pi(s)$ with the highest average reward, i.e.

$$\pi^* = \text{argmax}_{\pi} \mathbb{E}_{\tau \sim \pi, P} \left[R(\tau) \right] \quad \text{where} \quad R(\tau) = \sum_{t=1}^{T} R(s_t, a_t)$$

Counterfactually-guided training refers to the evaluation of the above expectation using data gathered via an action policy $\pi' \neq \pi$ and counterfactual reasoning

Buesing et al. "*Woulda, coulda, shoulda: Counterfactually-guided policy search.*" ICLR, 2018.

Counterfactually-guided training in reinforcement learning

In reinforcement learning, given a transition probability $P(s' \mid s, a)$ and a reward function $r(s, a)$, the goal is to design an action policy $a := \pi(s)$ with the highest average reward, i.e.

$$\pi^* = \text{argmax}_\pi \mathbb{E}_{\tau \sim \pi, P}\left[R(\tau)\right] \quad \text{where} \quad R(\tau) = \sum_{t=1}^{T} R(s_t, a_t)$$

Counterfactually-guided training refers to the evaluation of the above expectation using data gathered via an action policy $\pi' \neq \pi$ and counterfactual reasoning

Structural Causal Model $\mathscr{M}$

$$S_{t+1} := g_S(S_t, A_t, \mathbf{U}_t)$$

$$A_t := \pi'(S_t)$$

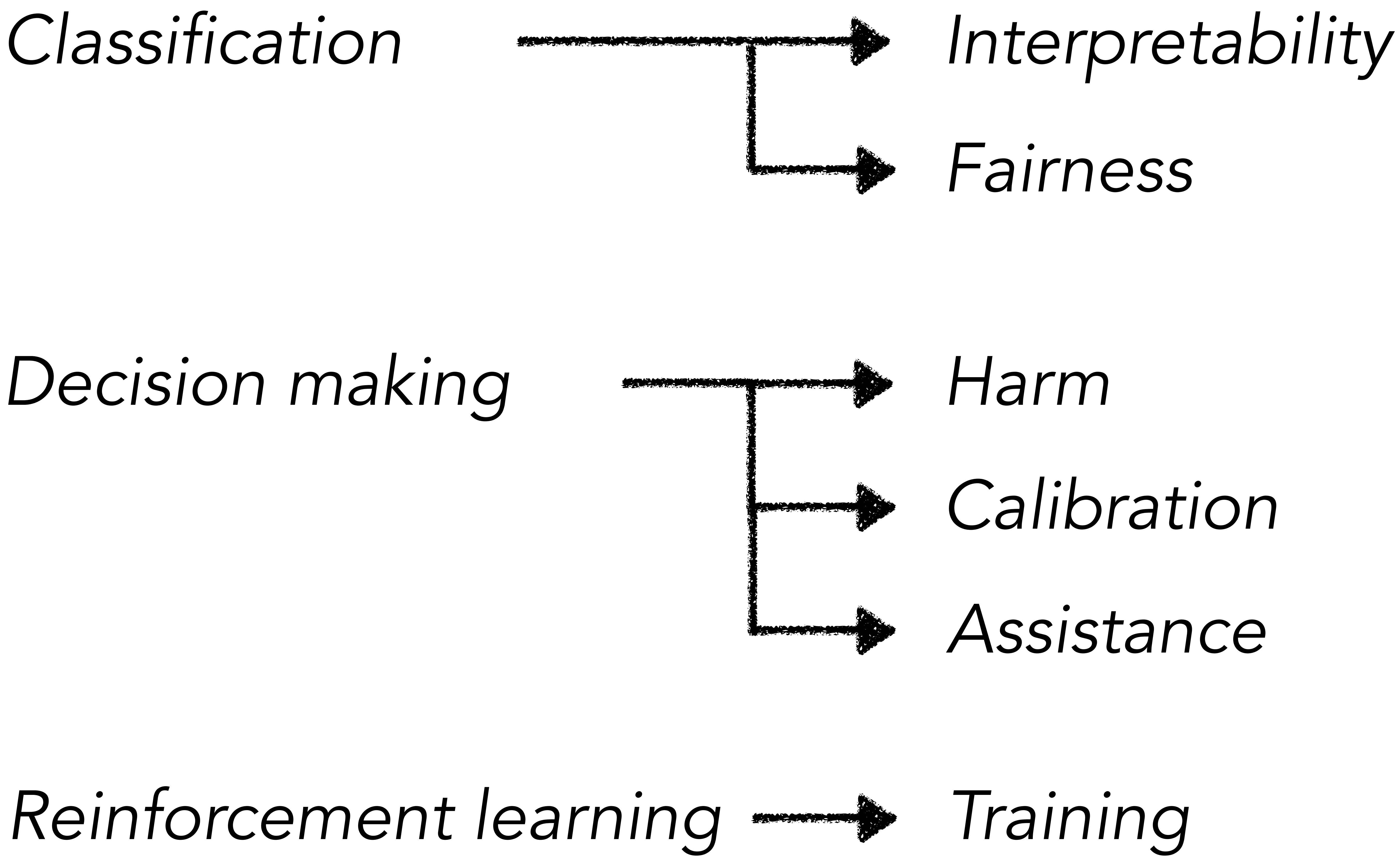
$$\mathbf{U}_t \sim P(\mathbf{U})$$

Key idea:

$$E_{S_t, a_t \sim \underbrace{P^{\mathscr{M}}}_{\text{Observational probability}}}\left[\underbrace{P^{\mathscr{M} \mid S_t=s_t, A_t=a_t \,;\, do(A_t=\pi(S_t))}}_{\text{Counterfactual probability}}\right] = \underbrace{P^{\mathscr{M} \,;\, do(A_t=\pi(S_t))}}_{\text{Interventional probability}}$$

Buesing et al. "*Woulda, coulda, shoulda: Counterfactually-guided policy search.*" ICLR, 2018.

Use cases of counterfactuals in machine learning



Classification → *Interpretability*, *Fairness*

Decision making → *Harm*, *Calibration*, *Assistance*

Reinforcement learning → *Training*