Robust Explainable AI: the Case of Counterfactual Explanations

ECAI 2023

Francesco Leofante
f.leofante@imperial.ac.uk

About me

Research Fellow
Centre for Explainable AI
Imperial College London

Contacts:

- f.leofante@imperial.ac.uk
- https://fraleo.github.io/

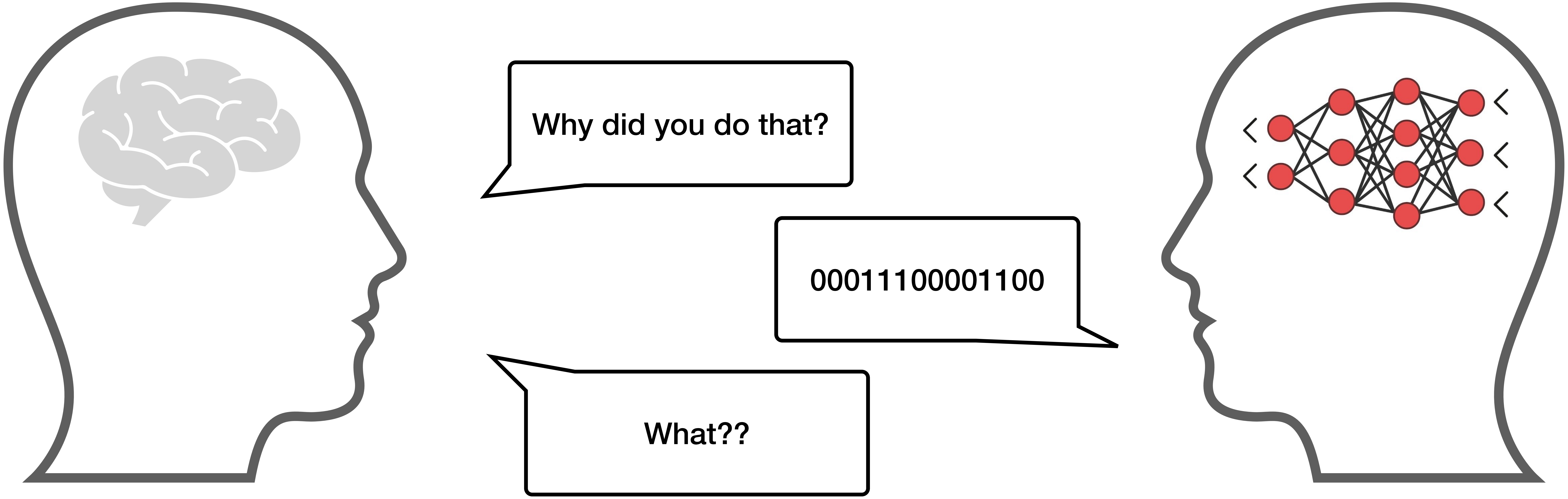
SCAN ME

Agenda

- Explainable AI
- Counterfactual explanations and recourse
- Robustness
    - **what** does it mean?
    - **why** is it needed?
    - **how** can we achieve it?
- Open discussion: robustness and other areas of CS

Explainable AI (XAI)

Techniques and methods that **make AI decisions understandable** by humans



Explainable AI (XAI)

XAI methods span a wide range of topics within AI and beyond, e.g.

- automated planning
- machine learning (ML)
- human computer interaction

Explainable AI (XAI)

XAI methods span a wide range of topics within AI and beyond, e.g.

- automated planning
- machine learning (ML)
- human computer interaction

Today we will focus on **explaining deep neural networks (DNNs)**

- **high-level** concepts rather than specific algorithms
- **fictional** use case and explanations

Supervised learning

Training set

- Age: 25
- Amount: £40K
- Duration: 36M

denied

- Age: 32
- Amount: £20K
- Duration: 24M

accepted

- Age: 82
- Amount: £26K
- Duration: 34M

denied

- Age: 54
- Amount: £14K
- Duration: 24M

accepted

Supervised learning

Training set

- Age: 25
- Amount: £40K
- Duration: 36M

denied

- Age: 32
- Amount: £20K
- Duration: 24M

accepted

- Age: 82
- Amount: £26K
- Duration: 34M

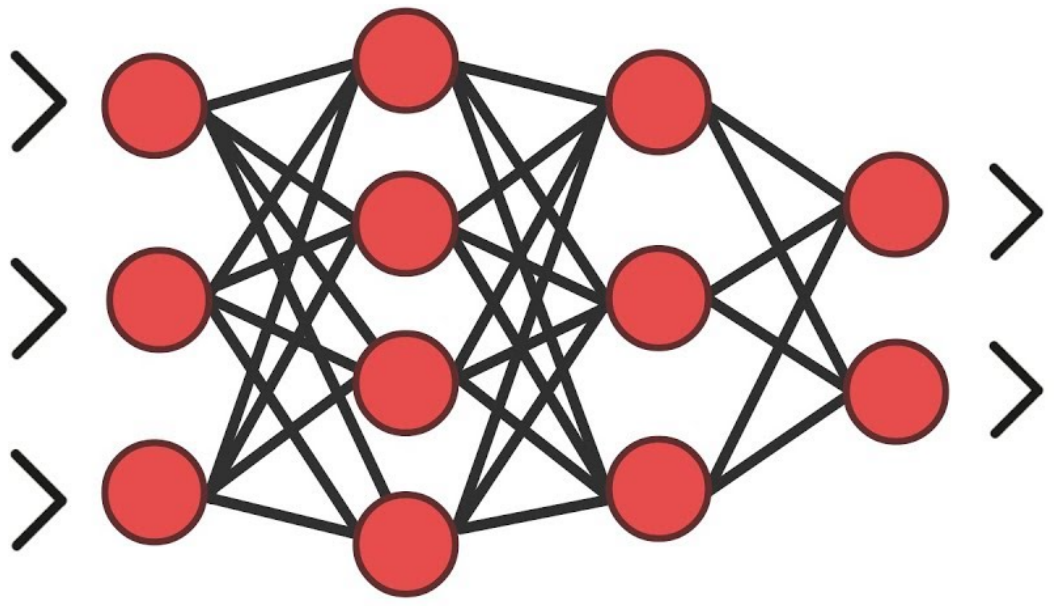
denied

- Age: 54
- Amount: £14K
- Duration: 24M

accepted

Deep neural network

(using your favourite algorithm)



Supervised learning

Training set

- Age: 25
- Amount: £40K
- Duration: 36M

denied

- Age: 32
- Amount: £20K
- Duration: 24M

accepted

- Age: 82
- Amount: £26K
- Duration: 34M

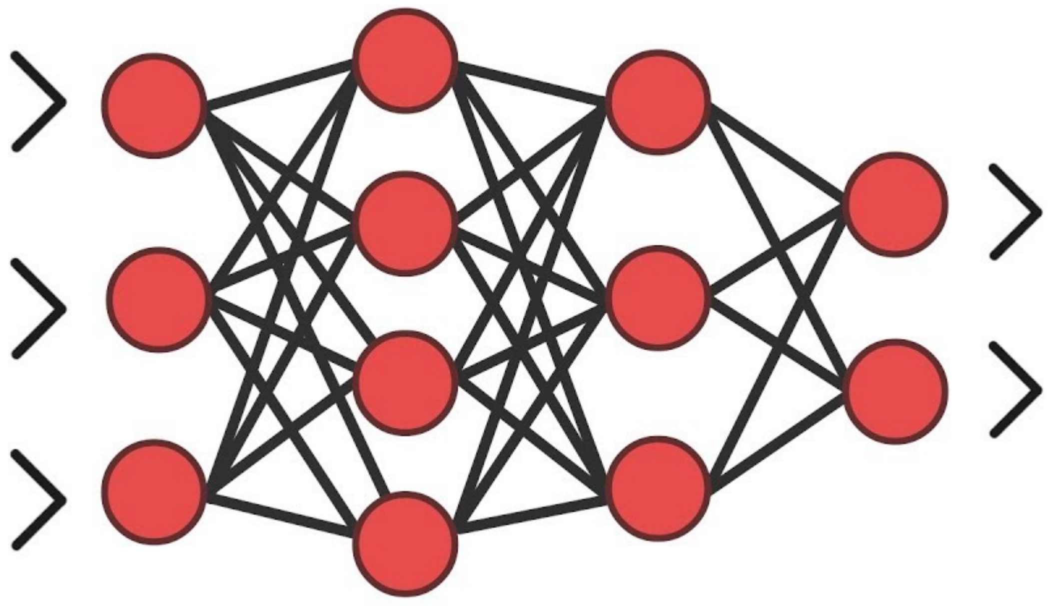
denied

- Age: 54
- Amount: £14K
- Duration: 24M

accepted

Deep neural network

(using your favourite algorithm)



Predicted class:
denied

New instance

Supervised learning

Training set

- Age: 25
- Amount: £40K
- Duration: 36M

denied

- Age: 32
- Amount: £20K
- Duration: 24M

accepted

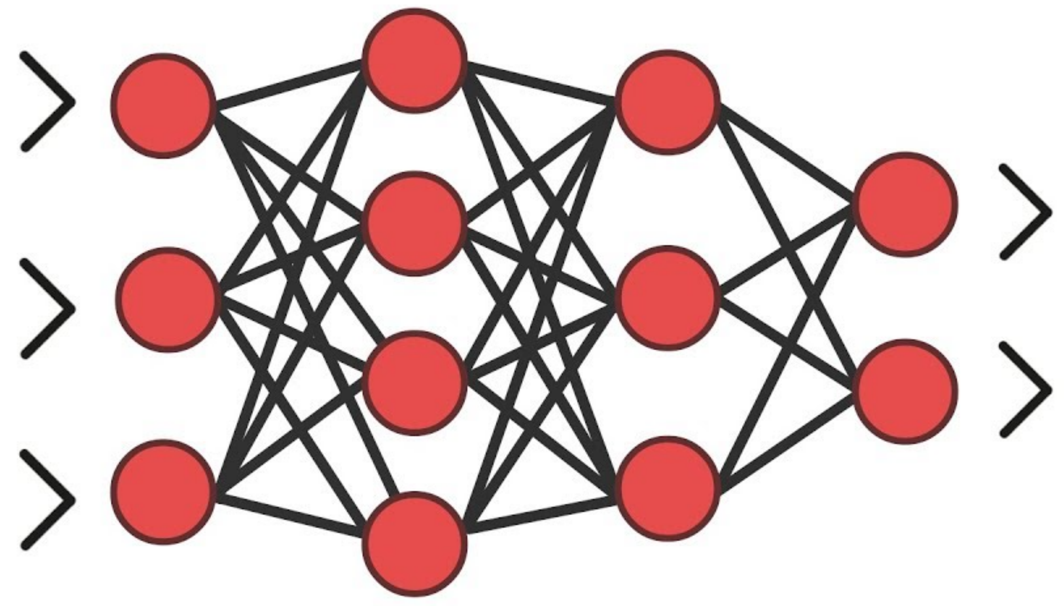
- Age: 82
- Amount: £26K
- Duration: 34M

denied

- Age: 54
- Amount: £14K
- Duration: 24M

accepted

Focus: explaining model predictions

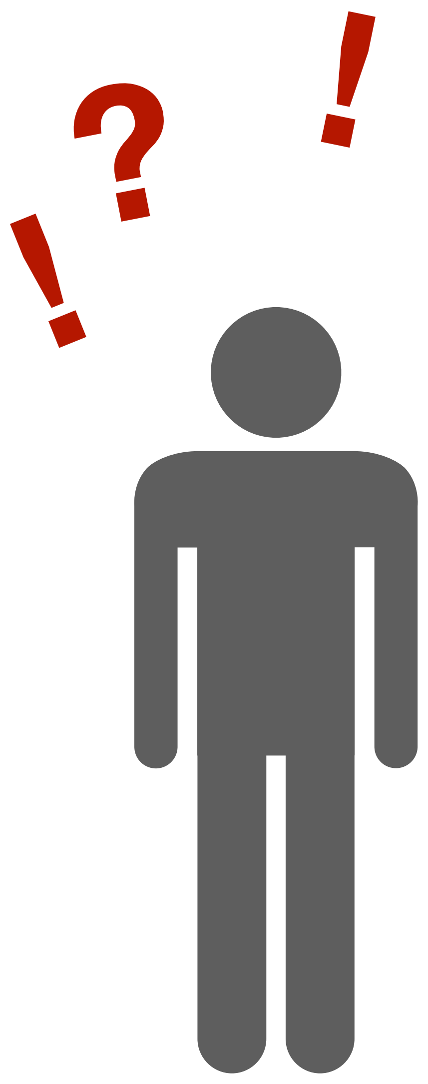


Predicted class: **denied**

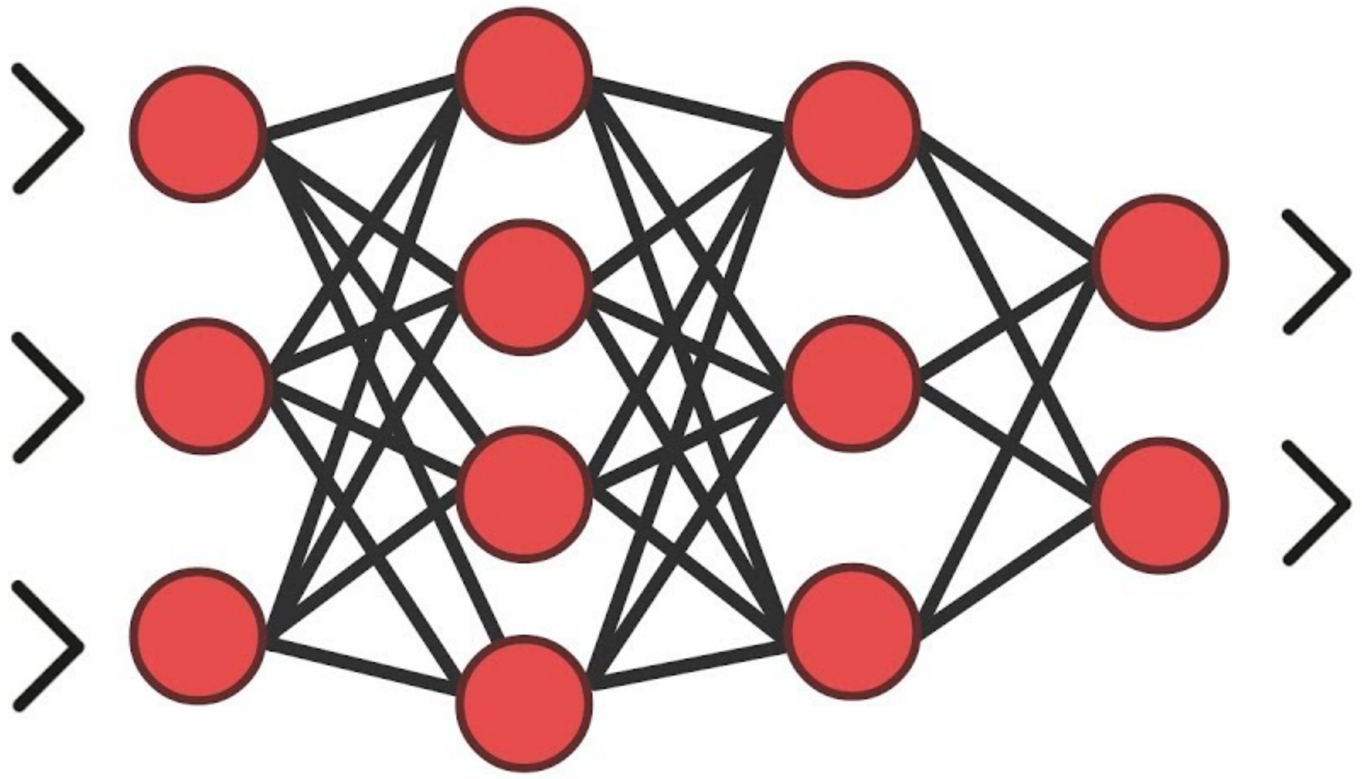
New instance

- Why is it denied?
- Why not accepted?
- How do I get accepted?
- And many more questions…

Challenge



- Age: 30
- Amount: £15K
- Duration: 24M



Loan denied

DNNs are black boxes!

Why is it a problem?

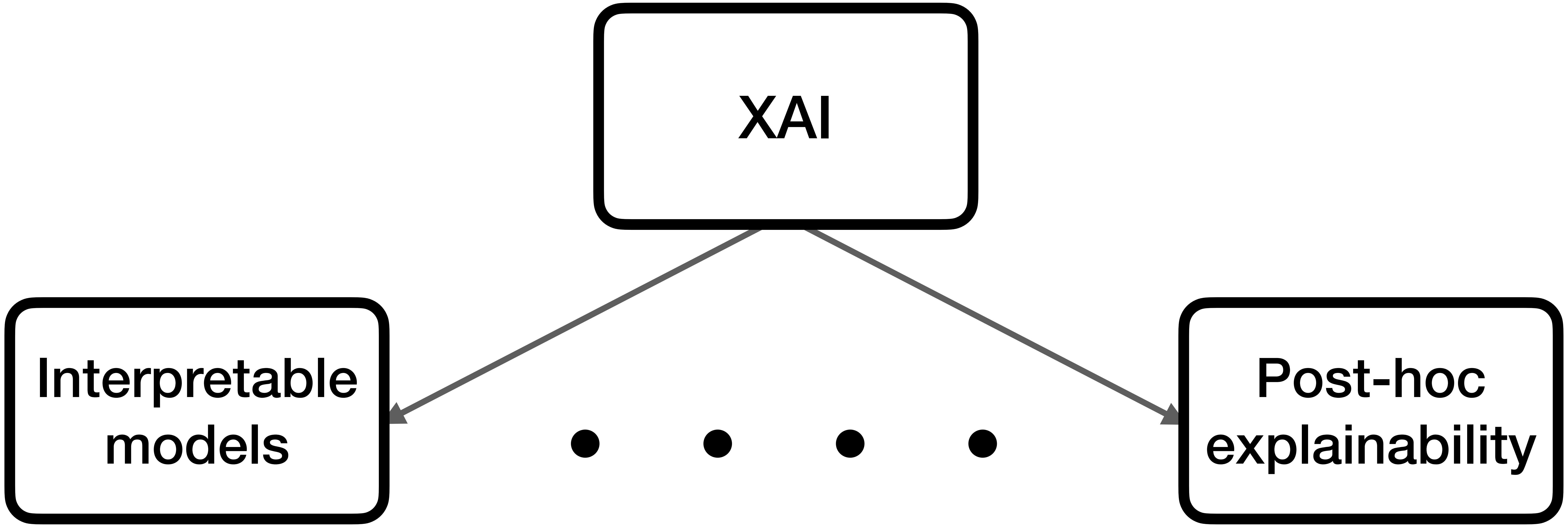
Why is it a problem?

The **General Data Protection Regulation** (GDPR)

- Art 22: **limits to decision-making based solely on automated processing**

- Art 13, 2f: right to be provided with **meaningful information about the logic involved in the decision-making**

How to achieve XAI?



- Linear models
- Decision trees
- Rule-based models

- Deep networks
- Ensemble models

How to achieve XAI?

XAI

Interpretable models

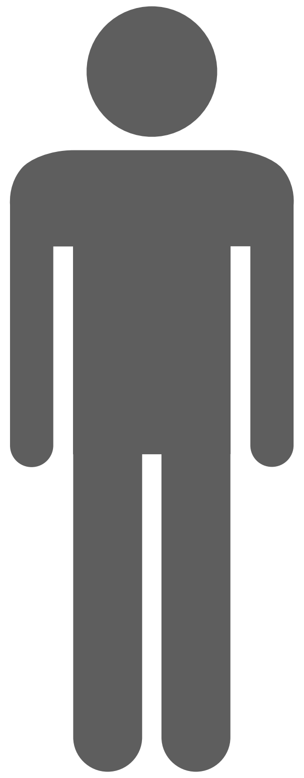
- Linear models
- Decision trees
- Rule-based models

Post-hoc explainability

- Dee
- Ensemble mo

Counterfactual explanations (CXs)

Original instance

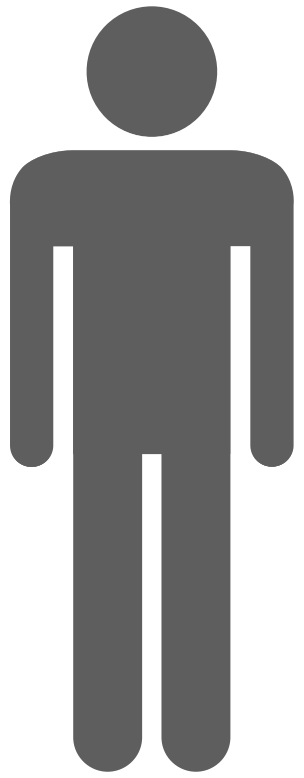


- Age: 30
- Amount: £15K
- Duration: 24M

Loan denied

Counterfactual explanations (CXs)

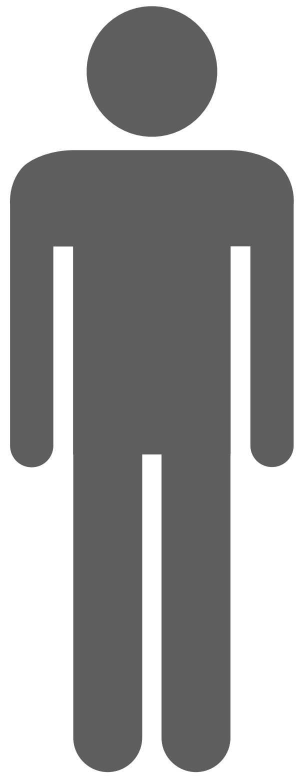
Original instance



- Age: 30
- Amount: £15K
- Duration: 24M

Loan denied

Counterfactual explanation

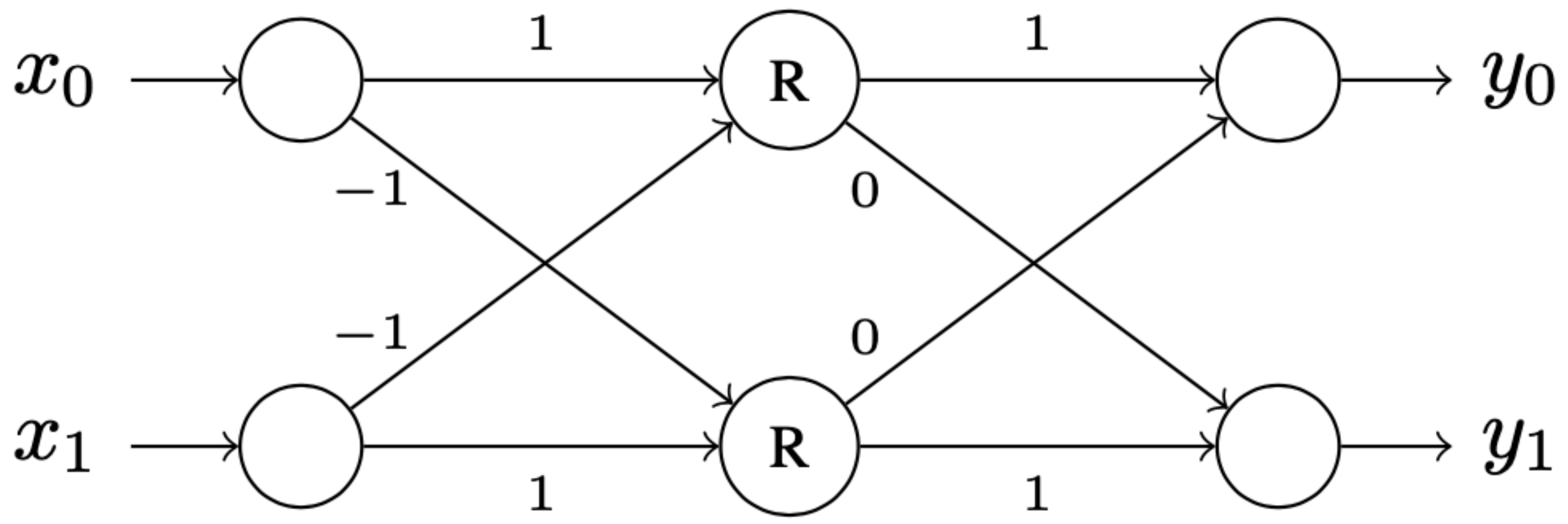


- Age: 30
- Amount: **£10K**
- Duration: 24M

The application would have been accepted **had you asked for £10K instead of £15K**

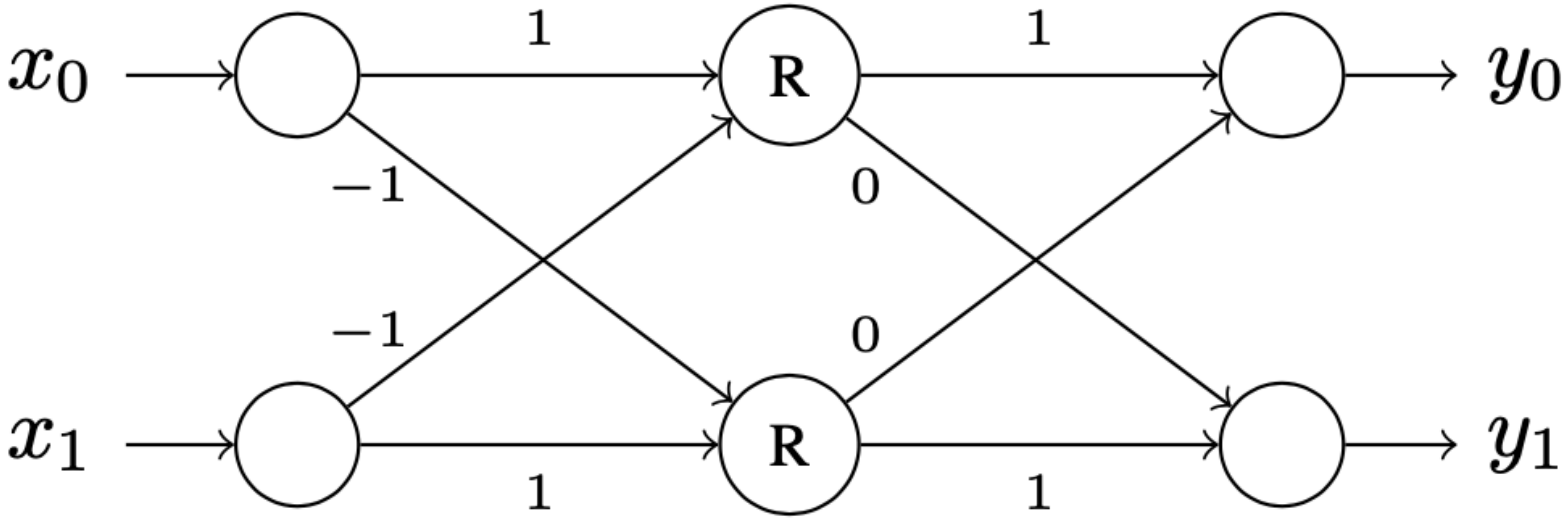
CX example

Consider the neural network $\mathscr{M}$ below:



CX example

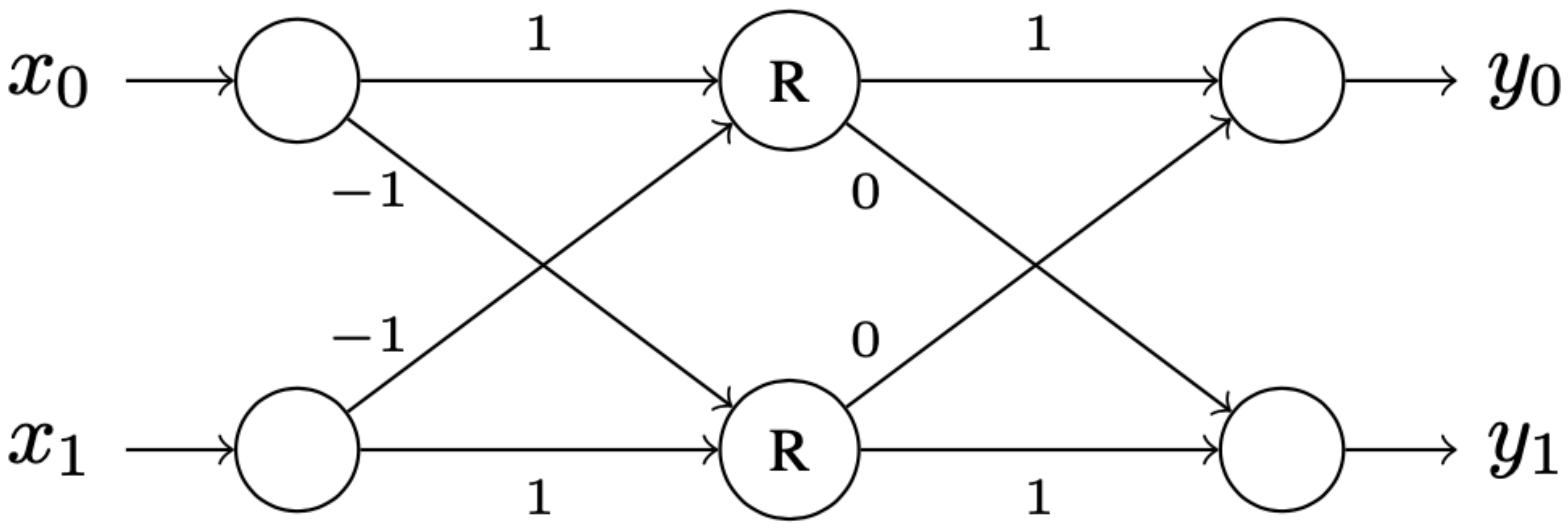
Consider the neural network $\mathscr{M}$ below:



- Given input $x_F = [1,2]$, $\mathscr{M}$ predicts class 1 ($y_1 > y_0$)

CX example

Consider the neural network $\mathscr{M}$ below:



- Given input $x_F = [1,2]$, $\mathscr{M}$ predicts class 1 ($y_1 > y_0$)
- A possible CX may be $x = [2.1,2]$, for which $\mathscr{M}$ predicts class 0

Computing a CX

- Given an input x_F and a binary classifier $\mathscr{M}$ such that $\mathscr{M}(x_F) = c$
- A distance function d

Computing a CX

- Given an input x_F and a binary classifier $\mathscr{M}$ such that $\mathscr{M}(x_F) = c$
- A distance function d

A **counterfactual explanation** x is computed as:

$$\arg\min_{x} d(x_F, x)$$

$$\text{subject to } \mathscr{M}(x) = 1 - c$$

Computing a CX

Most approaches solve relaxation defined as:

$$\arg\min_{x} \ell(\mathcal{M}(x), 1 - c) + \lambda \cdot d(x_F, x)$$

Counterfactual explanations without opening the black box: automated decisions and the GDPR. Wachter et al, Harvard Journal of Law & Technology 2018.

Computing a CX

Most approaches solve relaxation defined as:

$$\arg\min_{x} \boxed{\ell(\mathscr{M}(x), 1 - c)} + \lambda \cdot d(x_F, x)$$

where:

- ℓ is a differentiable loss function which minimises the gap between current and desired prediction

Counterfactual explanations without opening the black box: automated decisions and the GDPR. Wachter et al, Harvard Journal of Law & Technology 2018.

Computing a CX

Most approaches solve relaxation defined as:

$$\arg\min_{x} \ell(\mathscr{M}(x), 1 - c) + \boxed{\lambda \cdot d(x_F, x)}$$

where:

- ℓ is a differentiable loss function which minimises the gap between current and desired prediction

- λ controls distance trade-off

Counterfactual explanations without opening the black box: automated decisions and the GDPR. Wachter et al, Harvard Journal of Law & Technology 2018.

Tool support

AI Explainability 360 (v0.3.0)

Build passing | docs passing | pypi package 0.2.1

The AI Explainability 360 toolkit is an open-source library that supports interpretability and explainability of datasets and machine learning models. The AI Explainability 360 Python package includes a comprehensive set of algorithms that cover different dimensions of explanations along with proxy explainability metrics. The AI Explainability 360 toolkit supports tabular, text, images, and time series data.

The AI Explainability 360 interactive experience provides a gentle introduction to the concepts and capabilities by walking through an example use case for different consumer personas. The tutorials and example notebooks offer a deeper, data scientist-oriented introduction. The complete API is also available.

There is no single approach to explainability that works best. There are many ways to explain: data vs. model, directly interpretable vs. post hoc explanation, local vs. global, etc. It may therefore be confusing to figure out which algorithms are most appropriate for a given use case. To help, we have created some guidance material and a taxonomy tree that can be consulted.

https://github.com/Trusted-AI/AIX360

ALIBI EXPLAIN

CI passing | docs passing | codecov 85% | python 3.8 | 3.9 | 3.10 | 3.11 | pypi v0.9.4 | conda-forge v0.7.0 | license Apache-2.0 | chat on slack

Alibi is an open source Python library aimed at machine learning model inspection and interpretation. The focus of the library is to provide high-quality implementations of black-box, white-box, local and global explanation methods for classification and regression models.

https://github.com/SeldonIO/alibi

Captum

license BSD-3-Clause | pytorch v0.6.0 | pypi v0.6.0 | circleci failing | platform noarch | conda-forge v0.6.0 | recipe captum | docs captum

Captum is a model interpretability and understanding library for PyTorch. Captum means comprehension in Latin and contains general purpose implementations of integrated gradients, saliency maps, smoothgrad, vargrad and others for PyTorch models. It has quick integration for models built with domain-specific libraries such as torchvision, torchtext, and others.

Captum is currently in beta and under active development!

https://github.com/pytorch/captum

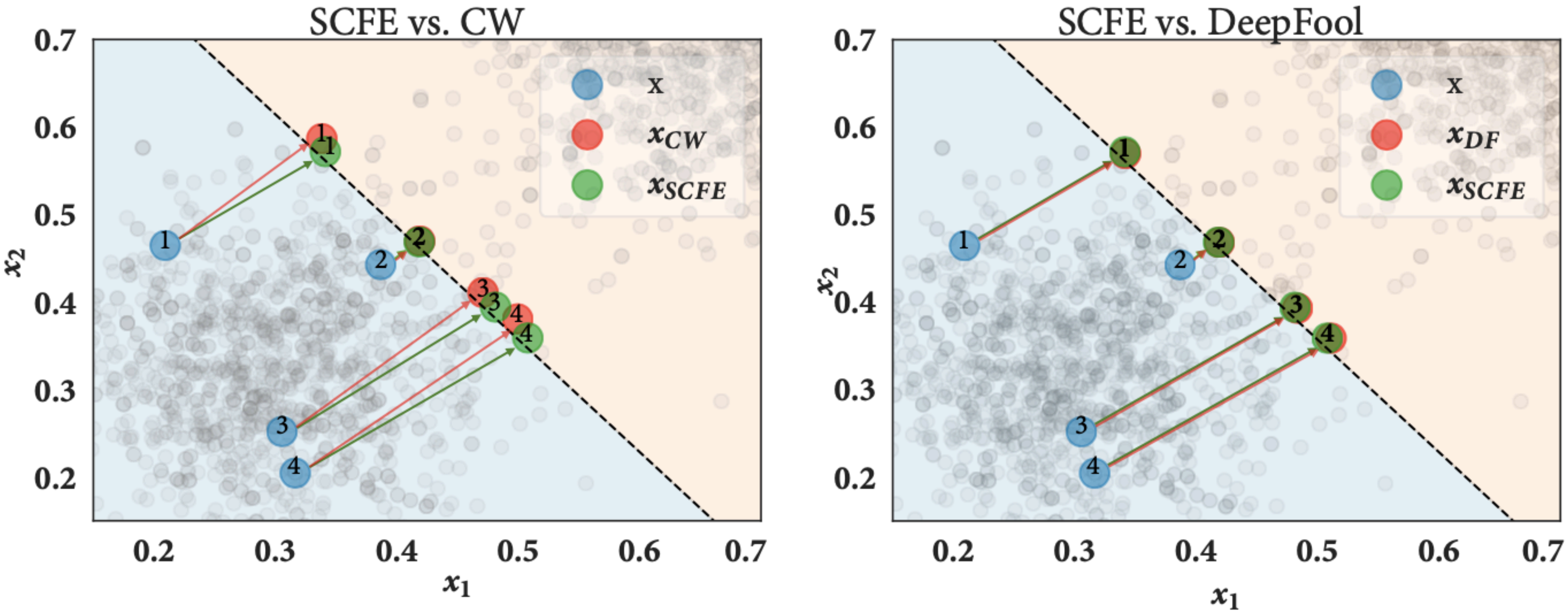
CARLA - Counterfactual And Recourse Library

CARLA is a python library to benchmark counterfactual explanation and recourse models. It comes out-of-the box with commonly used datasets and various machine learning models. Designed with extensibility in mind: Easily include your own counterfactual methods, new machine learning models or other datasets. Find extensive documentation here! Our arXiv paper can be found here.

What is algorithmic recourse? As machine learning (ML) models are increasingly being deployed in high-stakes applications, there has been growing interest in providing recourse to individuals adversely impacted by model predictions (e.g., below we depict the canonical recourse example for an applicant whose loan has been denied). This library provides a starting point for researchers and practitioners alike, who wish to understand the inner workings of various counterfactual explanation and recourse methods and their underlying assumptions that went into the design of these methods.

https://github.com/carla-recourse/CARLA

Is minimising distance always good?



CXs are often **indistinguishable** from **adversarial examples**!

Exploring Counterfactual Explanations Through the Lens of Adversarial Examples: A Theoretical and Empirical Analysis. Pawelczyk et al, AISTATS 2022.

Brittle explanations ahead!

Threats

1. Input perturbations
2. Model perturbations
3. Model multiplicity
4. Noisy execution

Robust XAI

Threats

1. Input perturbations
2. Model perturbations
3. Model multiplicity
4. Noisy execution

List of references is partial - too much to cover in 90 minutes!

Robust XAI

Threats

1. Input perturbations
2. Model perturbations
3. Model multiplicity
4. Noisy execution

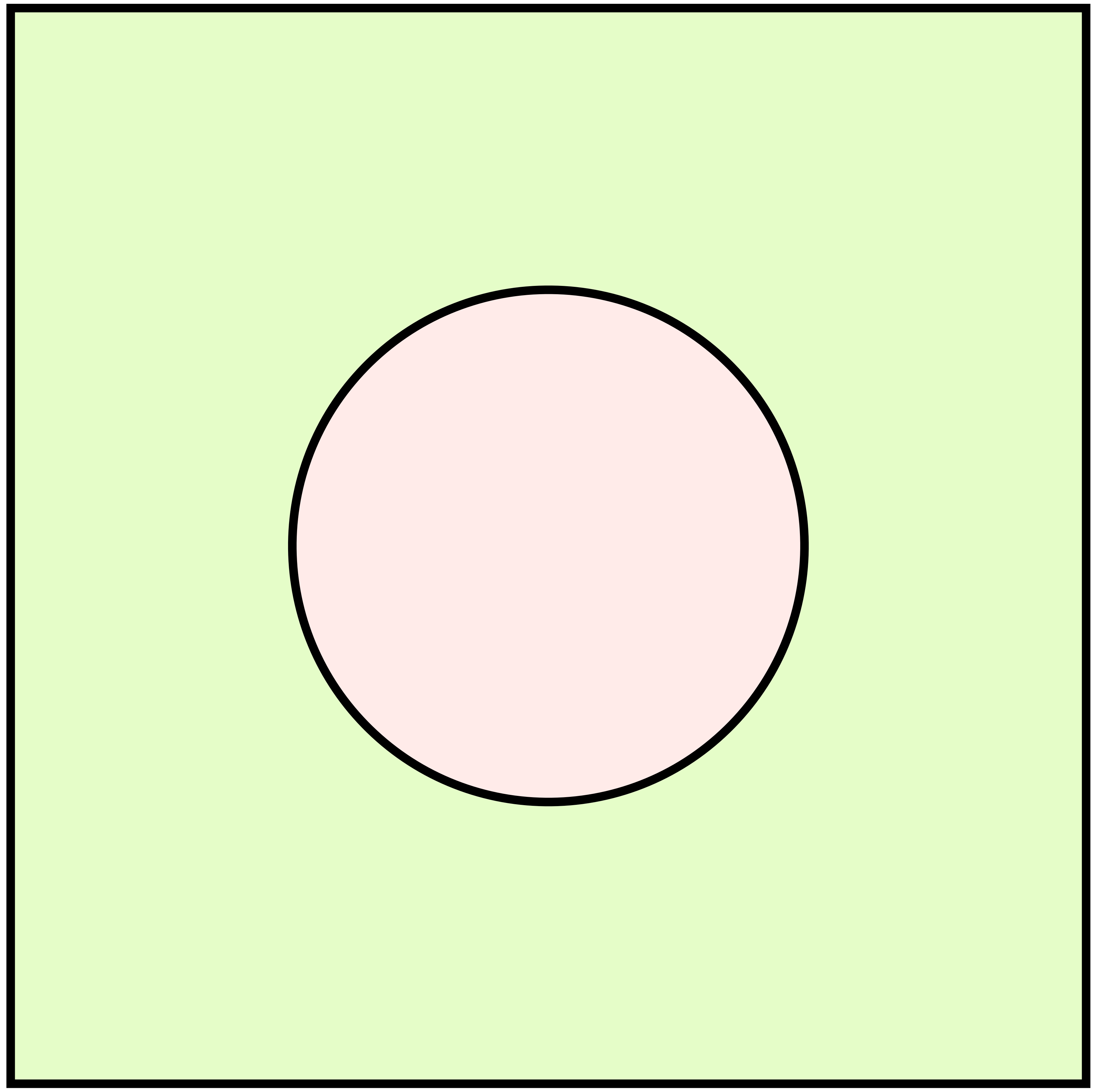
Heuristic vs **Exhaustive** robustness guarantees

Brittle explanations ahead!

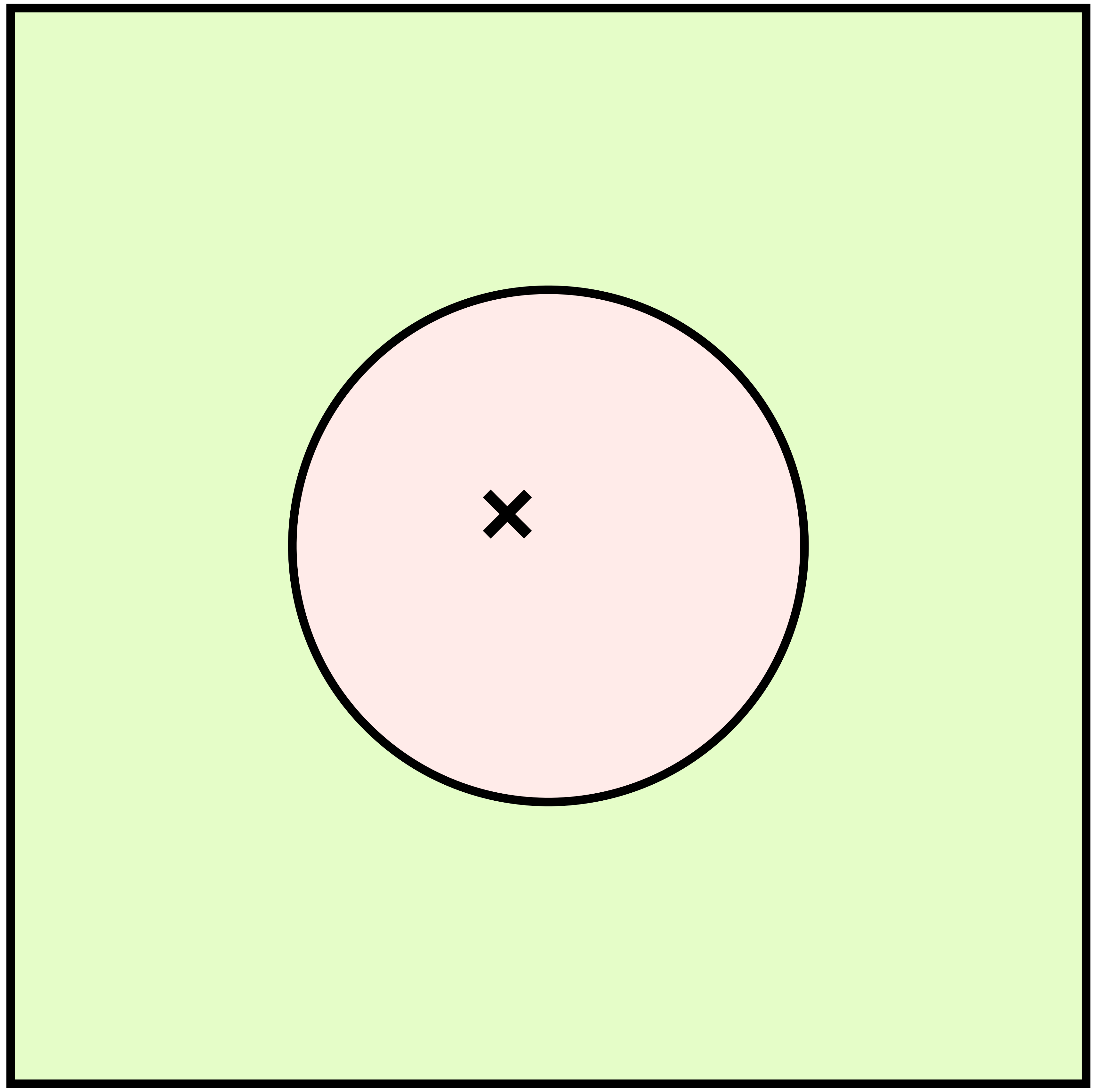
Threats

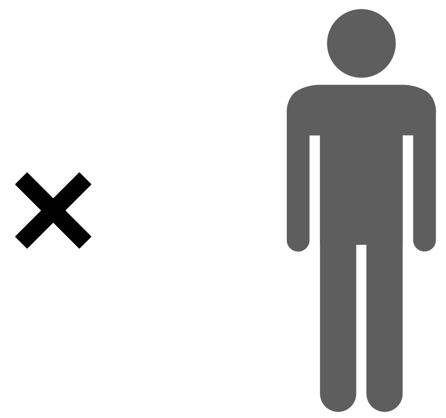
1. **Input perturbations**
2. Model perturbations
3. Model multiplicity
4. Noisy execution

Input perturbations



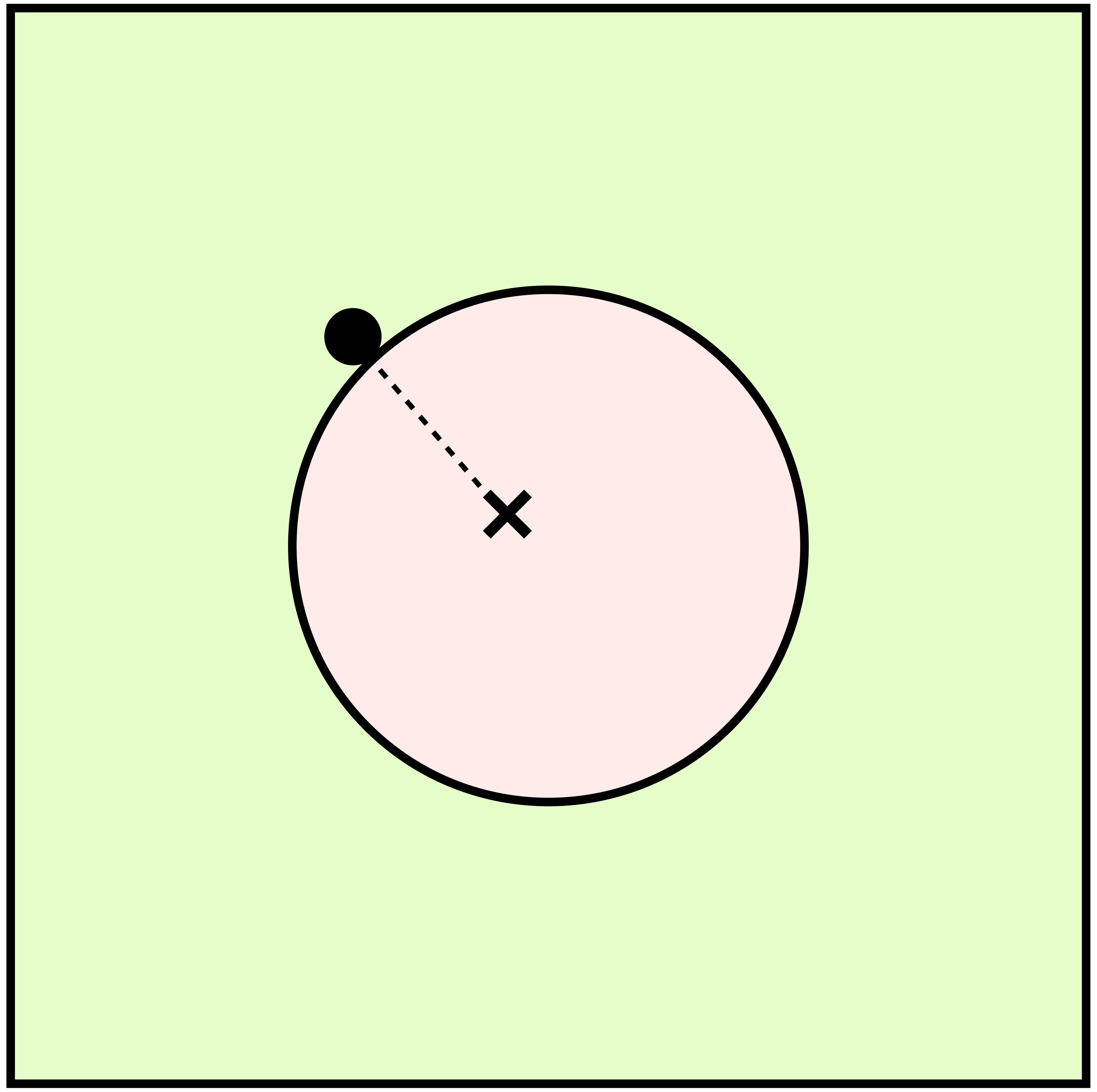
Input perturbations





- Age: 30
- Amount: £15K
- Duration: 24M

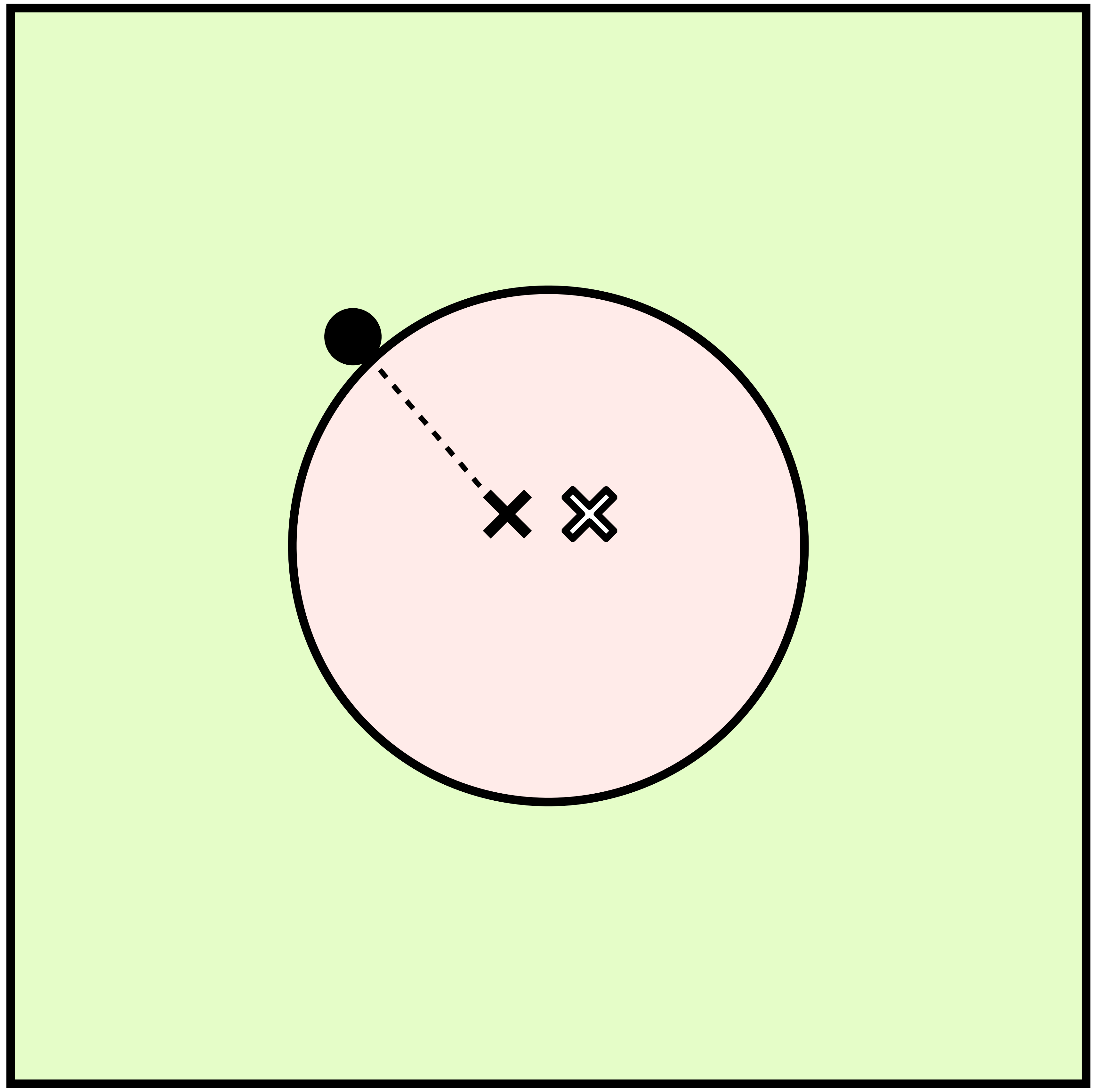
Input perturbations



- Age: 30
- Amount: £15K
- Duration: 24M

- Age: 30
- Amount: **£13K**
- Duration: **24M**

Input perturbations

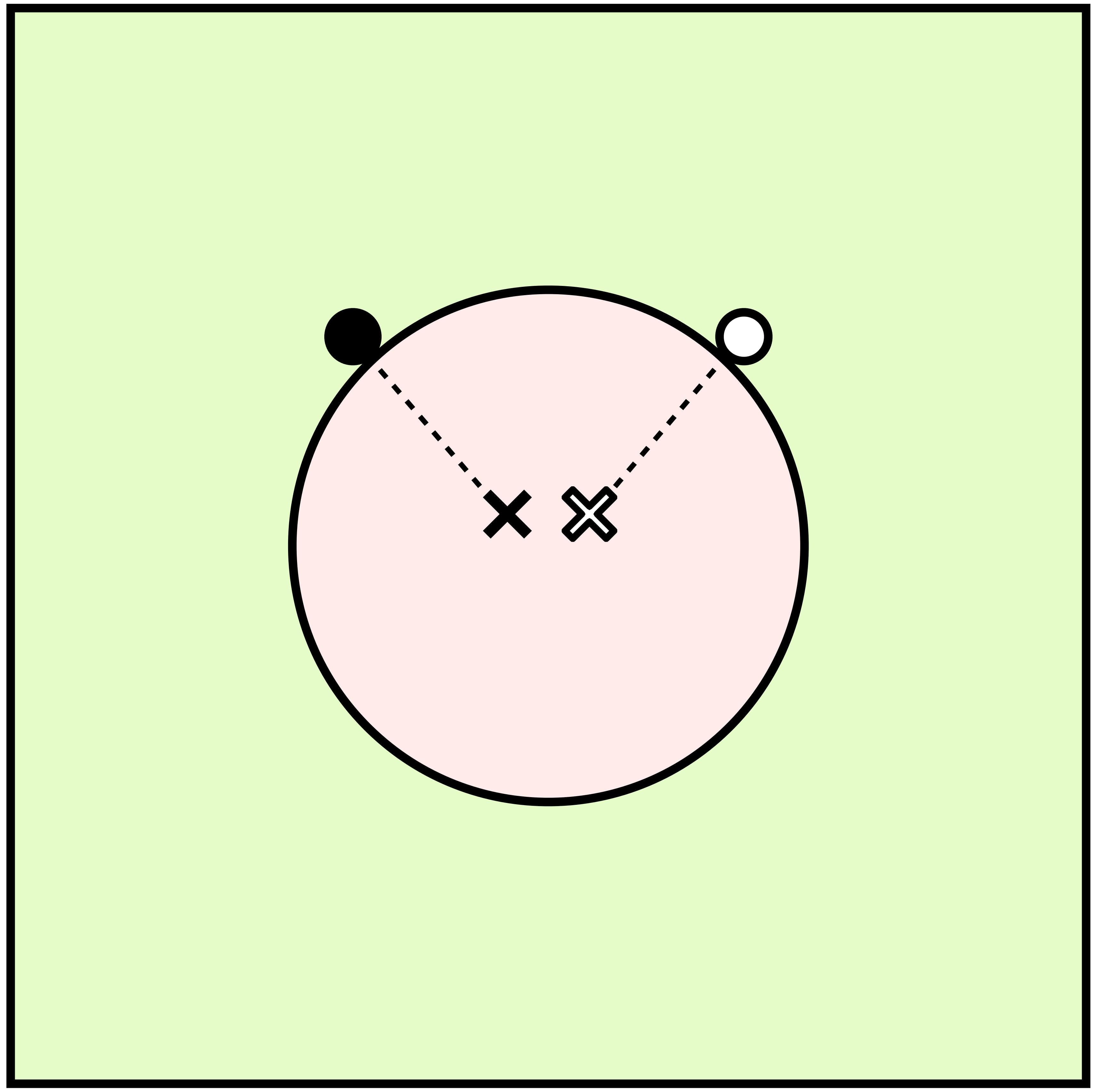


- Age: 30
- Amount: £15K
- Duration: 24M

- Age: 30
- Amount: **£13K**
- Duration: **24M**

- Age: 30
- Amount: £16K
- Duration: 24M

Input perturbations



×
- Age: 30
- Amount: £15K
- Duration: 24M

●
- Age: 30
- Amount: **£13K**
- Duration: **24M**

⊗
- Age: 30
- Amount: £16K
- Duration: 24M

? ? ?

○
- Age: 30
- Amount: **£10K**
- Duration: **12M**

Implications

Lack of robustness to input changes poses a number of problems!

we expect phenomena in the world that are similar to have similar explanations

Robustness in Machine Learning Explanations: Does it Matter? Hancox-Li, FAT* 2020.

Implications

Lack of robustness to input changes poses a number of problems!

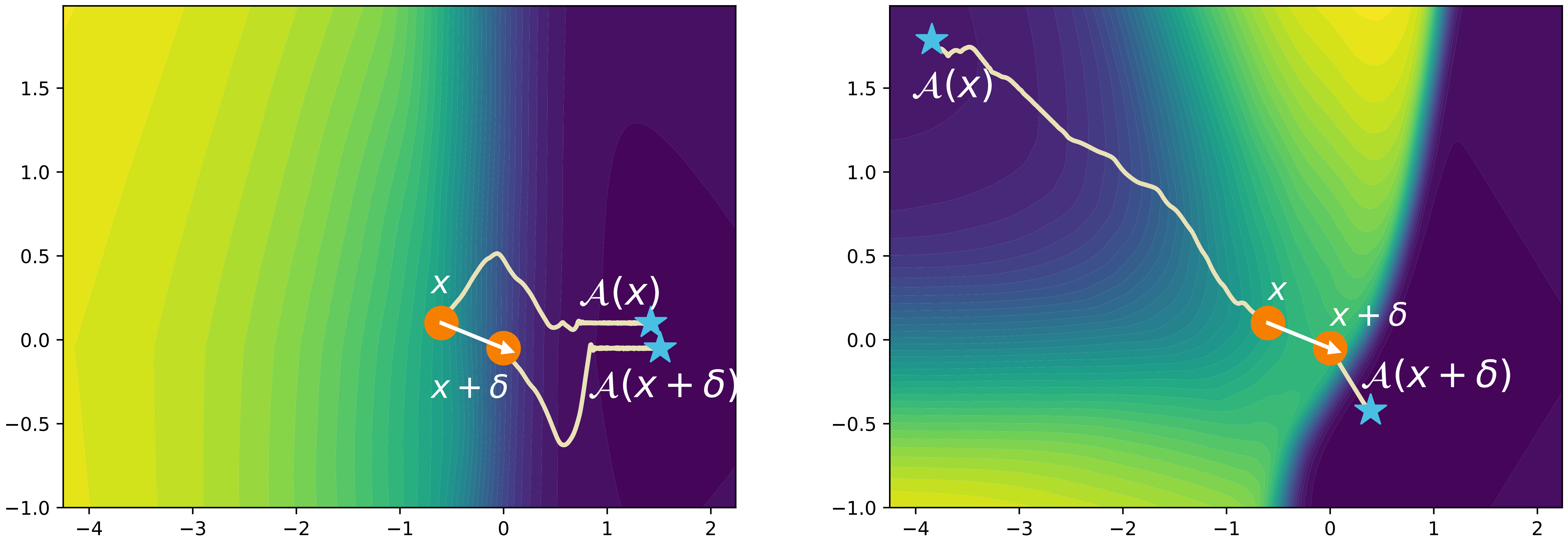
we expect phenomena in the world that are similar to have similar explanations

- is the explanation really capturing how the black-box works?
- we would expect **neighbouring inputs** to be **processed in similar ways**
- uncertainty in how data is collected may have huge impact on explanation

Robustness in Machine Learning Explanations: Does it Matter? Hancox-Li, FAT* 2020.

Implications

Can be exploited to train adversarial models that generate unfair explanations!



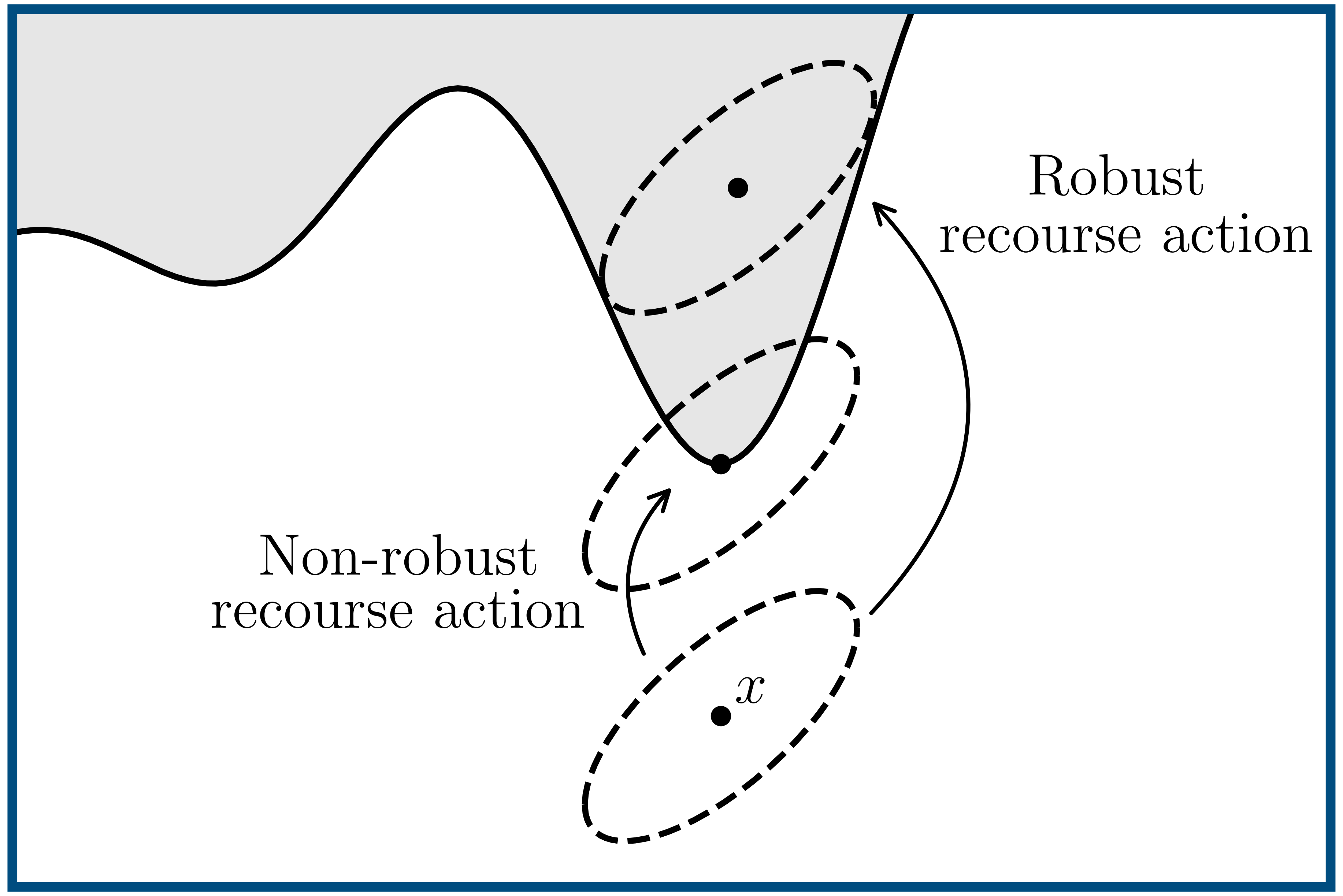
(a) **Training with BCE Objective** (b) **Training Adversarial Model**

Counterfactual Explanations Can Be Manipulated. Slack et al, NeurIPS 2021.

Solutions

Input perturbations may invalidate CXs!

- Dominguez-Olmedo propose a method to preserve validity (minmax formulation)

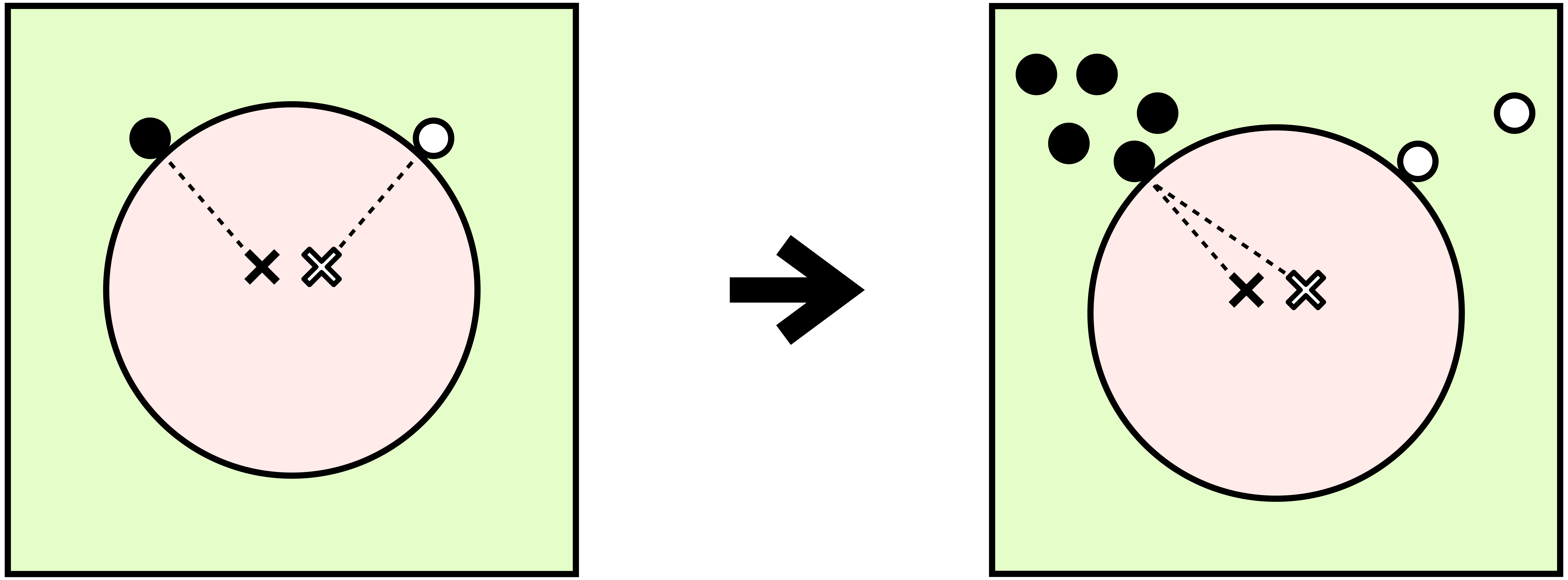


On the Adversarial Robustness of Causal Algorithmic Recourse. Dominguez-Olmedo et al, ICML 2022.

Solutions

Zhang et al propose to use **density** to guide CX search

- Similar inputs should “gravitate” towards similar CXs



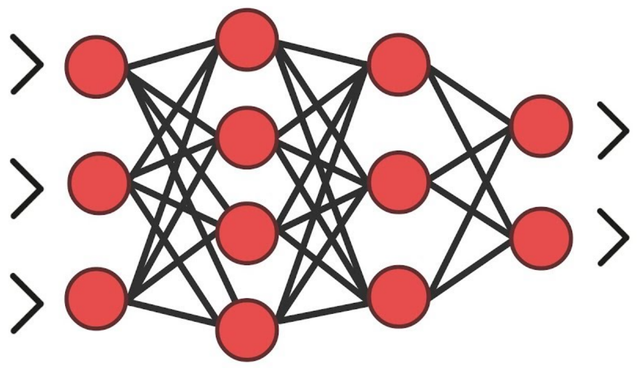
Density-based Realiable and Robust Explainer for Counterfactual Explanations. Zhang et al, Expert Systems with Applications, 2023.

Brittle explanations ahead!

Threats

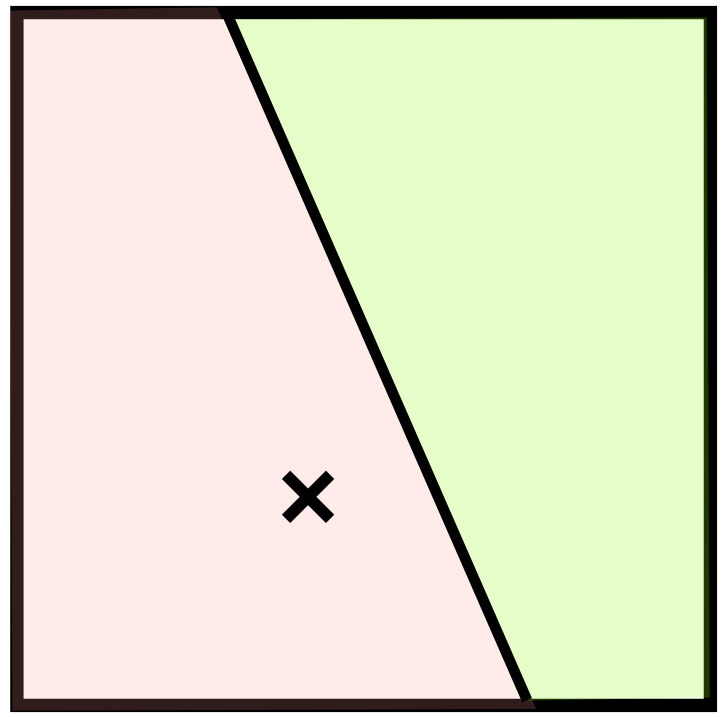
1. Input perturbations
2. **Model perturbations**
3. Model multiplicity
4. Noisy execution

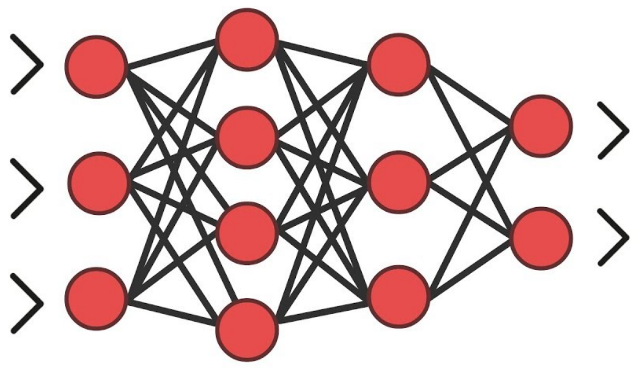
Model perturbations

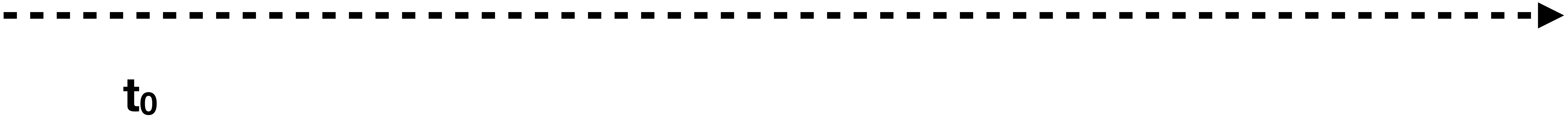


t_0

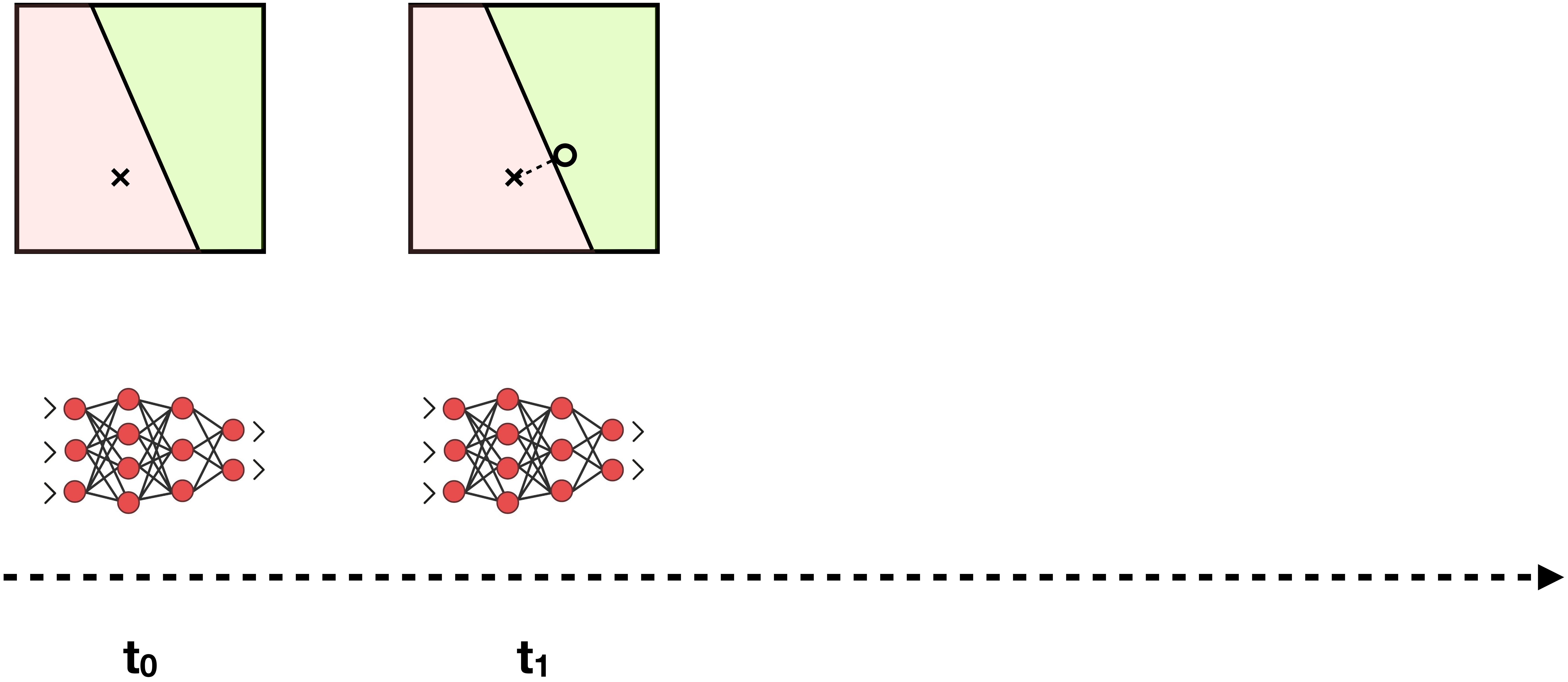
Model perturbations



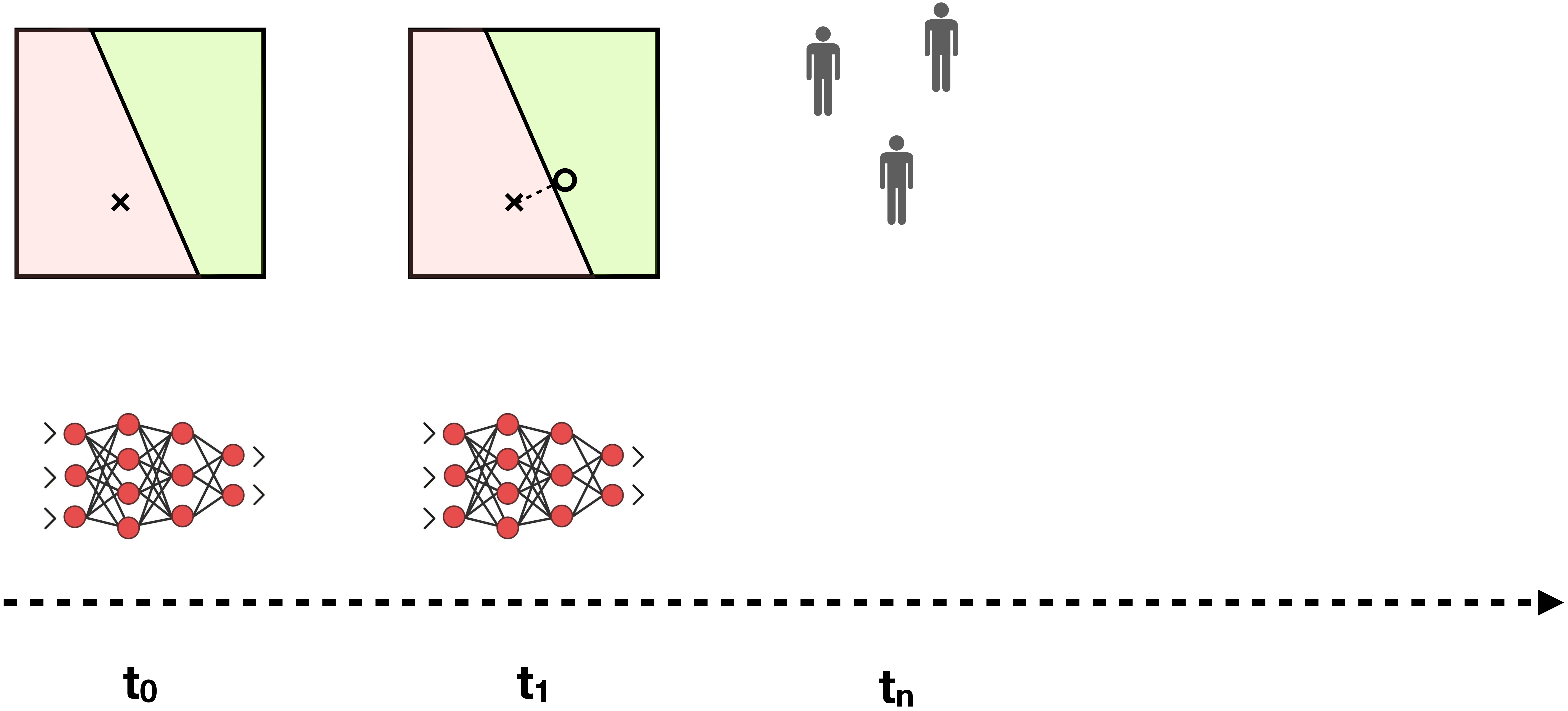




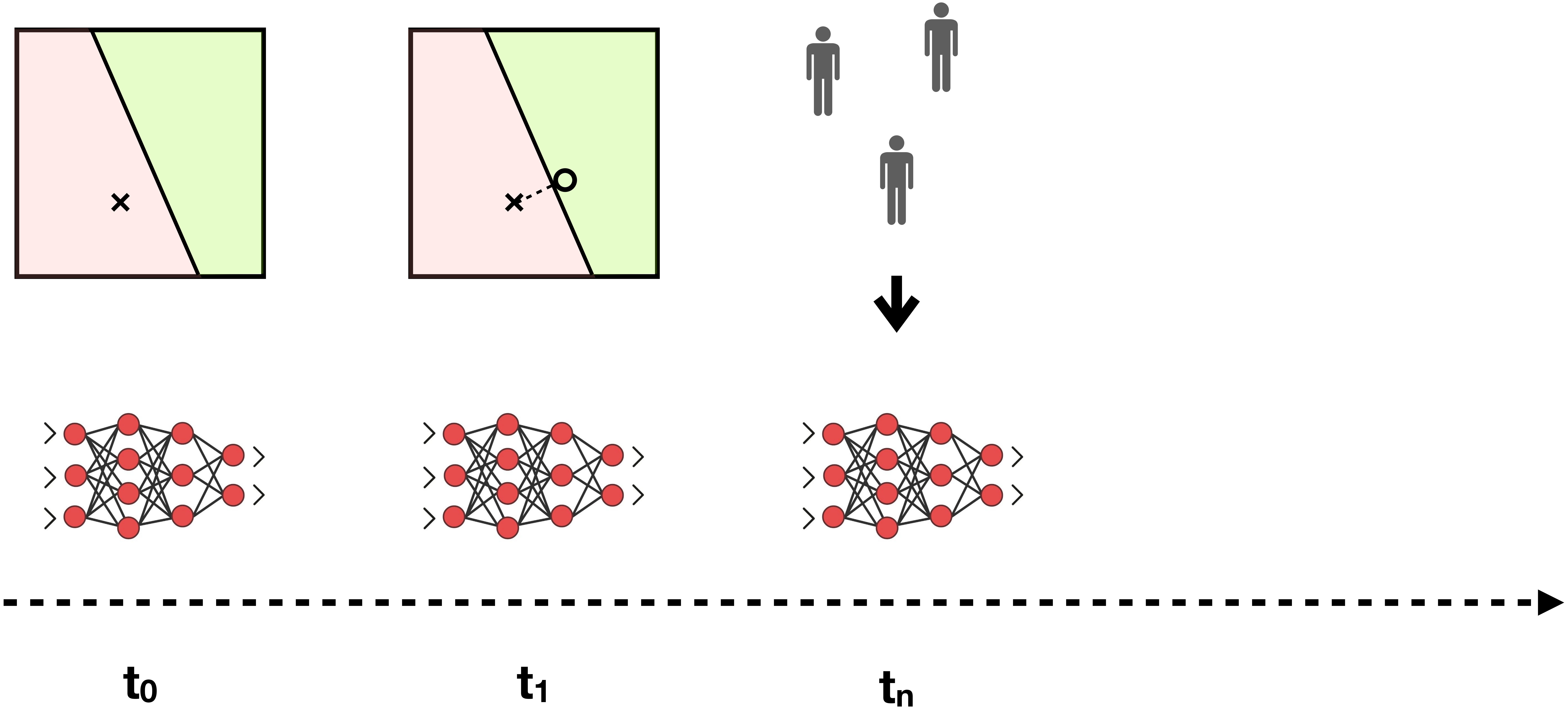
Model perturbations



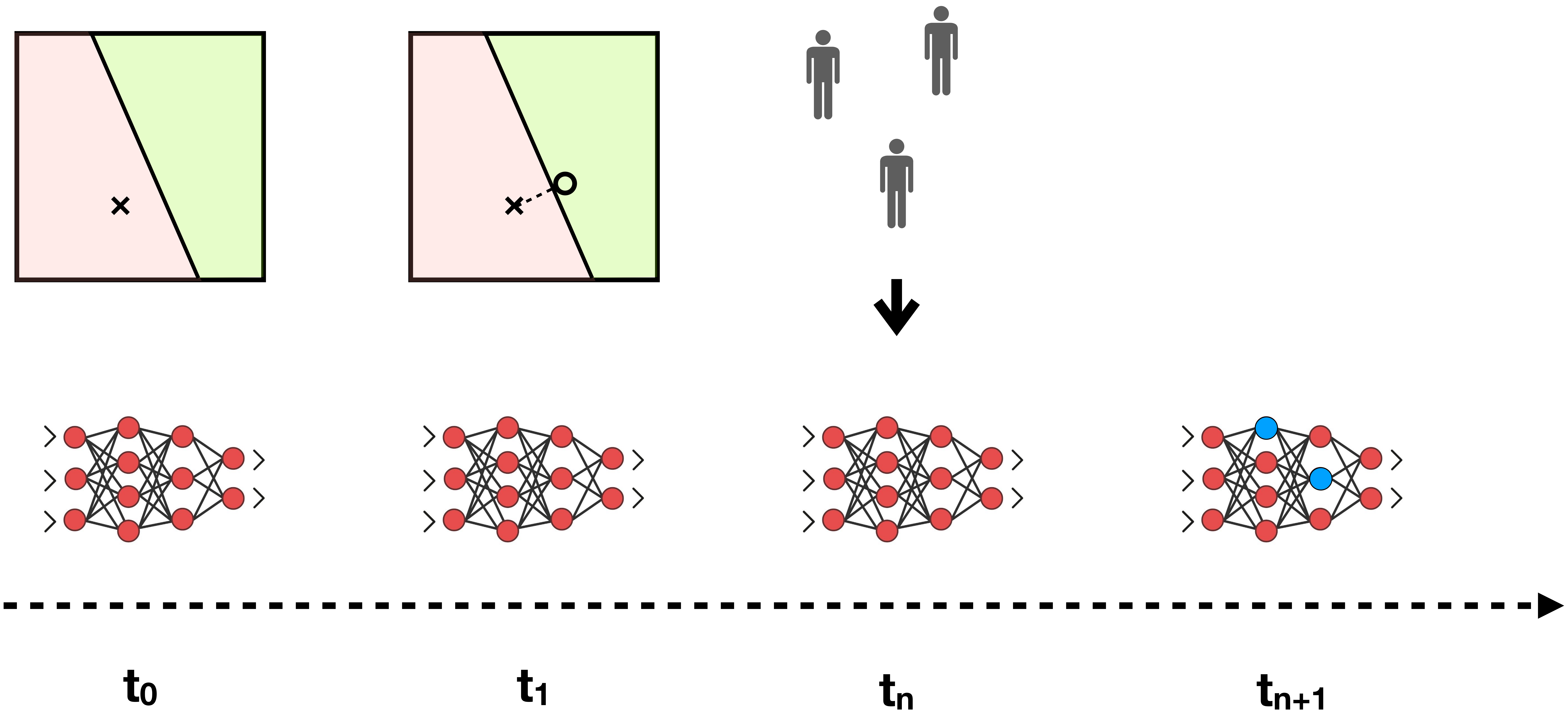
Model perturbations



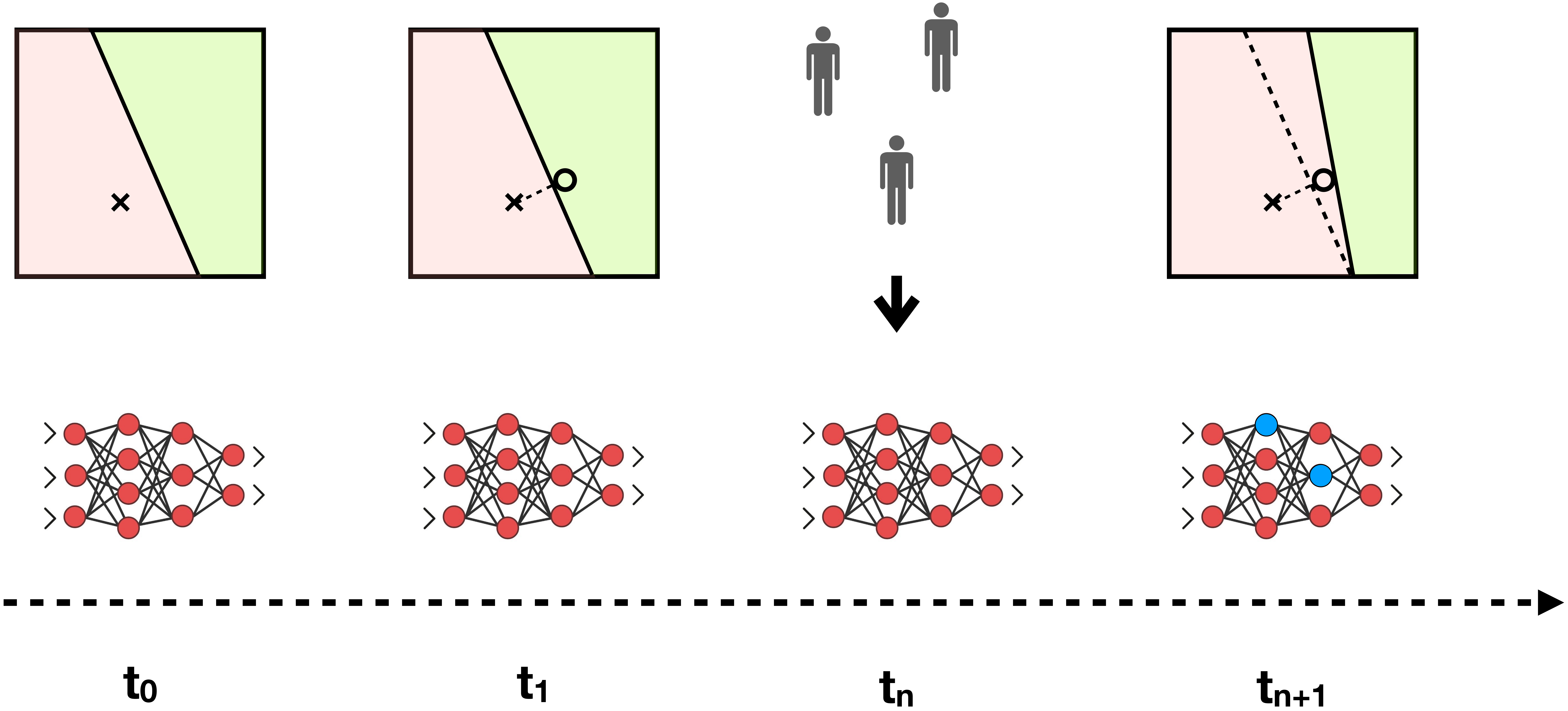
Model perturbations



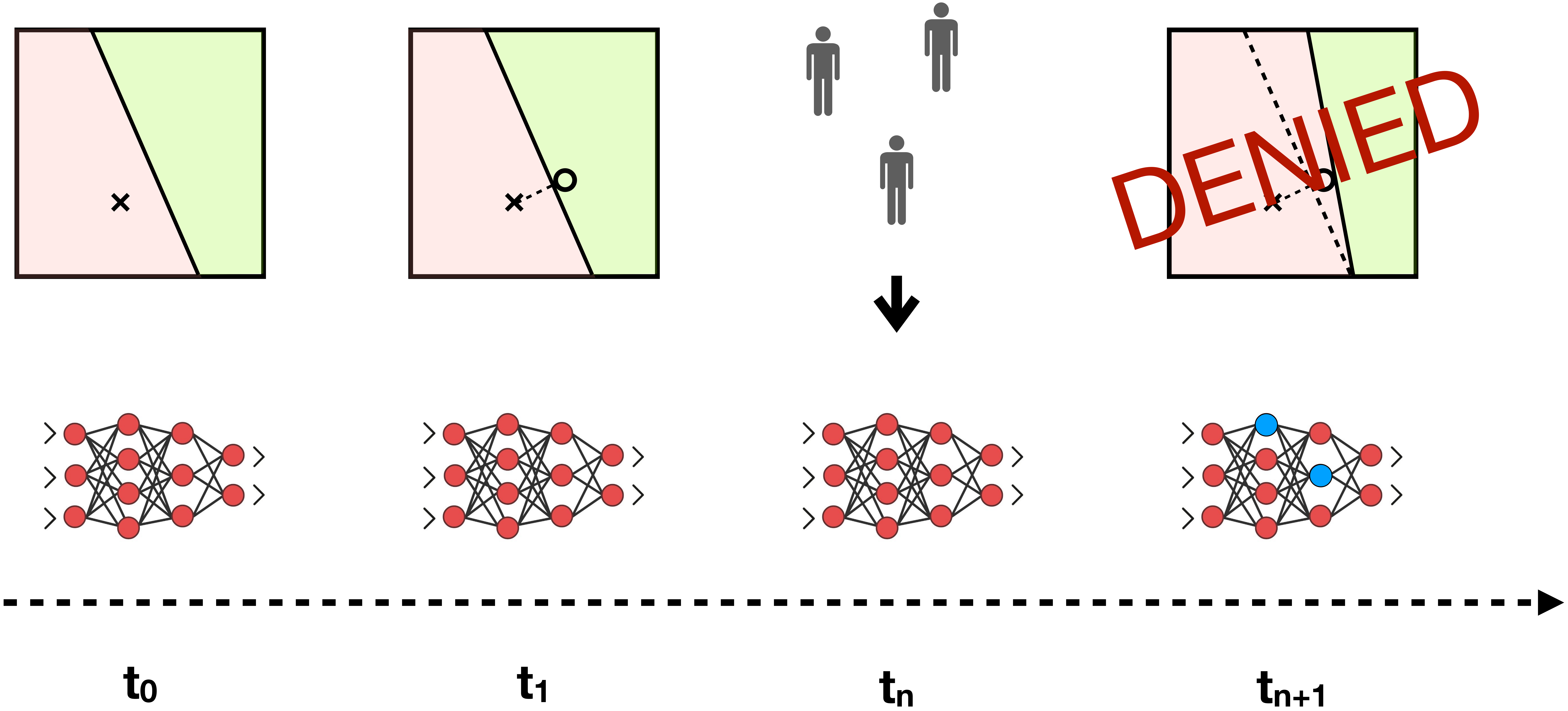
Model perturbations



Model perturbations



Model perturbations



Implications

Model shifts may occur as a result of data shifts

Implications

Model shifts may occur as a result of data shifts

Dilemma

Implications

Model shifts may occur as a result of data shifts

Dilemma

- **Trust** the old CX, although possibly contradicted by new data

Implications

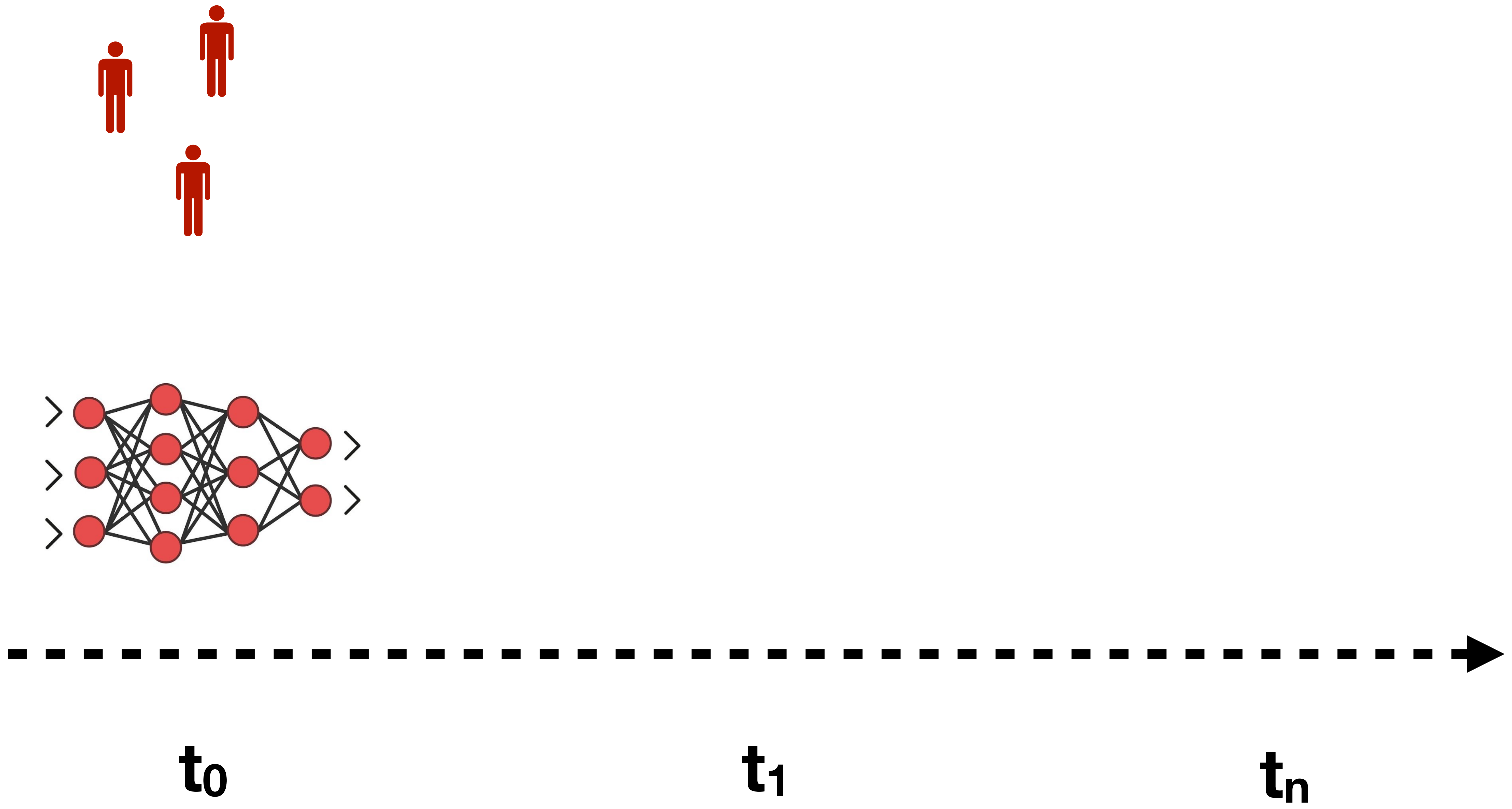
Model shifts may occur as a result of data shifts

Dilemma

- **Trust** the old CX, although possibly contradicted by new data
- **Trash** the old CX, possibly upsetting end users

Solutions

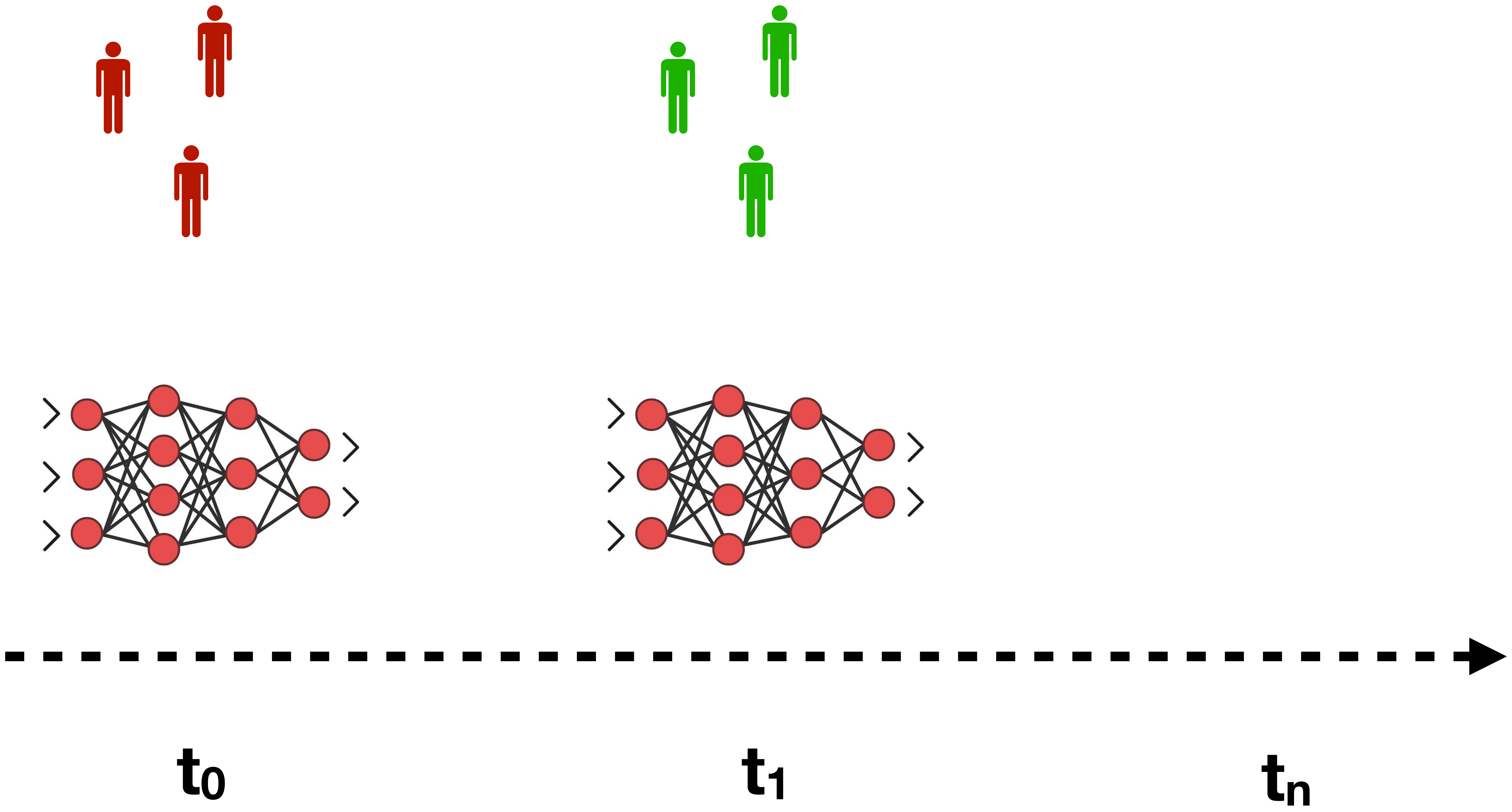
Ferrario and Loi proposed an augmentation technique to mitigate the issue



The Robustness of Counterfactual Explanations Over Time. Ferrario and Loi, IEEE Access, 2022.

Solutions

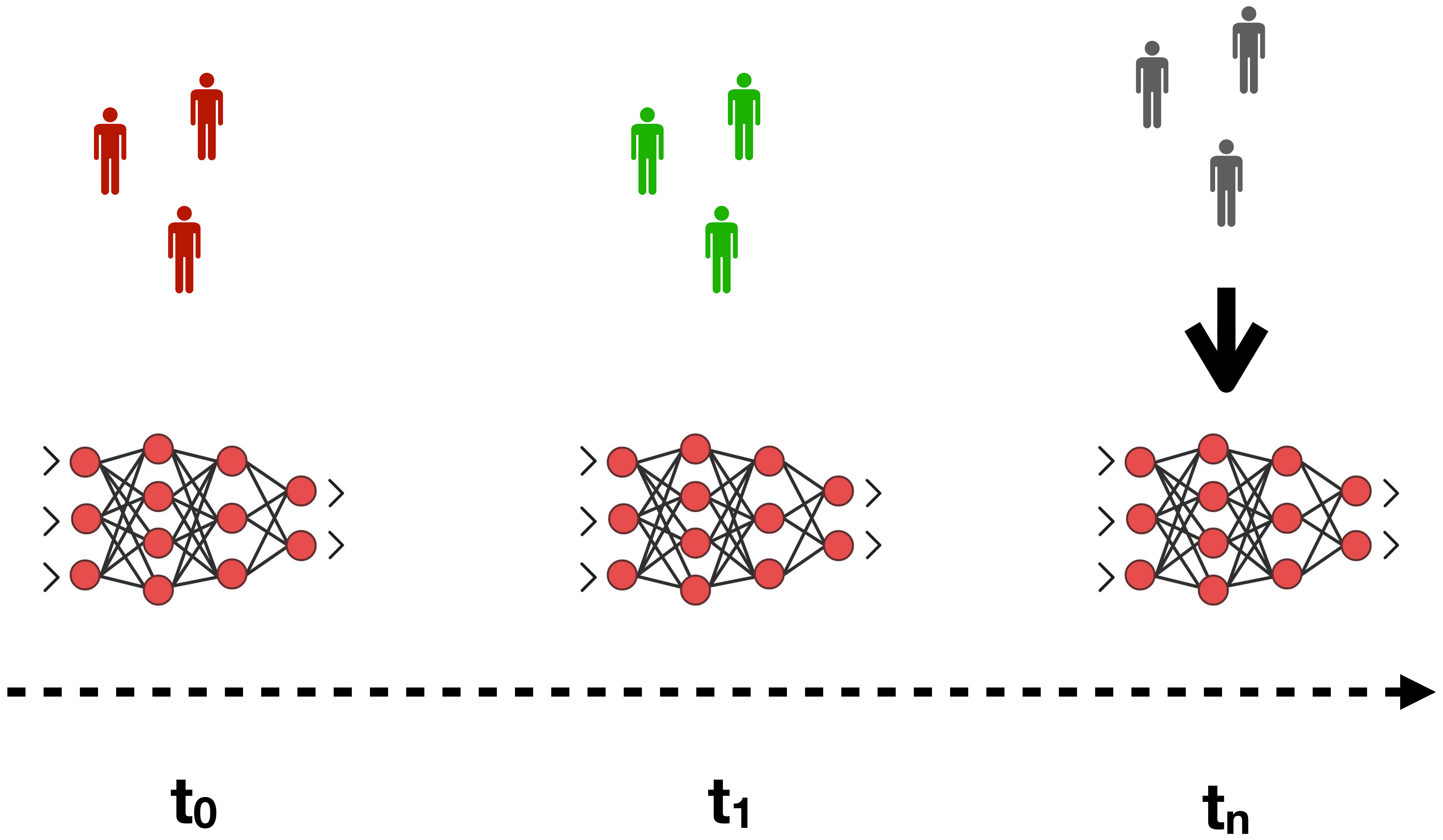
Ferrario and Loi proposed an augmentation technique to mitigate the issue



The Robustness of Counterfactual Explanations Over Time. Ferrario and Loi, IEEE Access, 2022.

Solutions

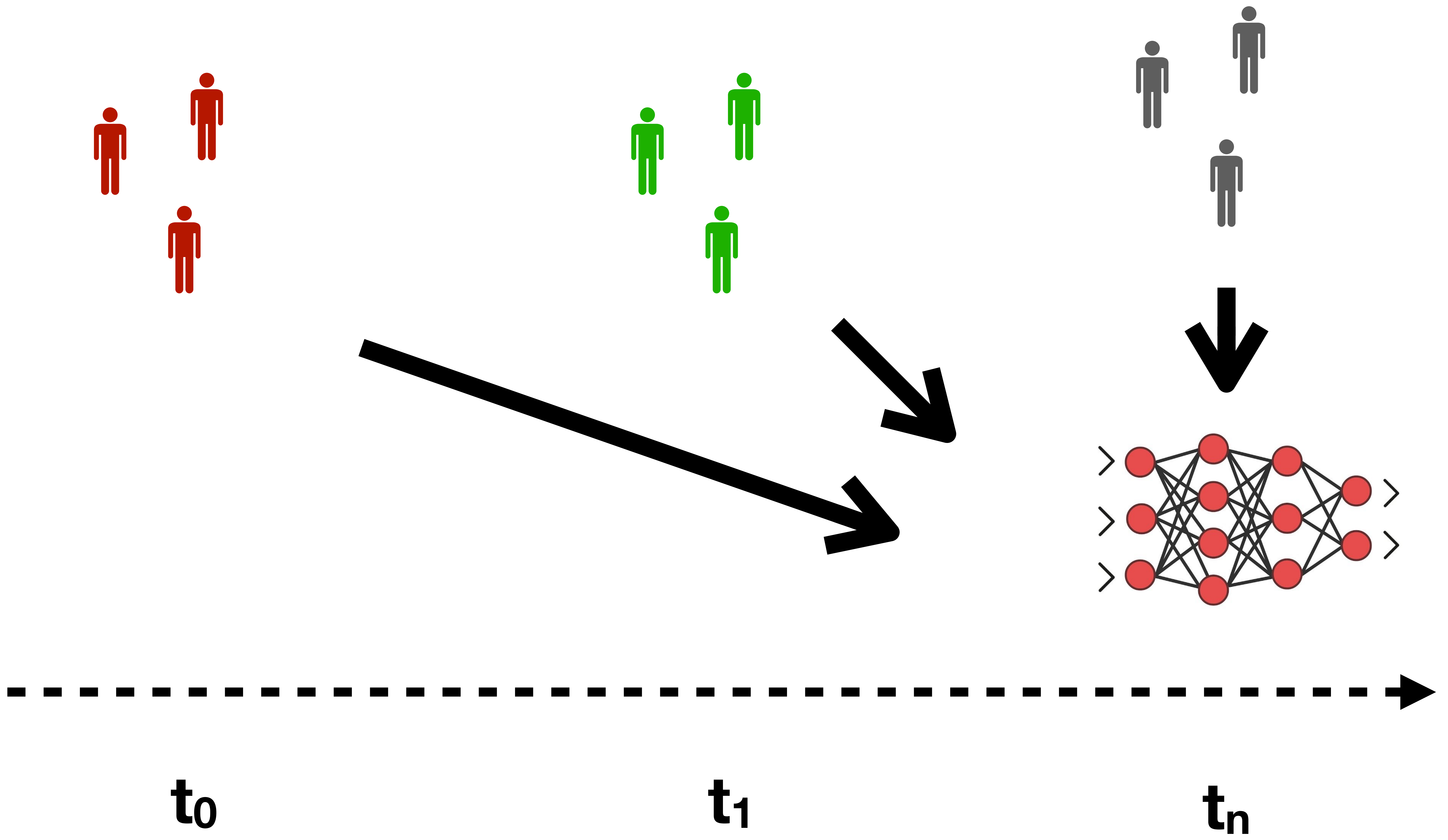
Ferrario and Loi proposed an augmentation technique to mitigate the issue



The Robustness of Counterfactual Explanations Over Time. Ferrario and Loi, IEEE Access, 2022.

Solutions

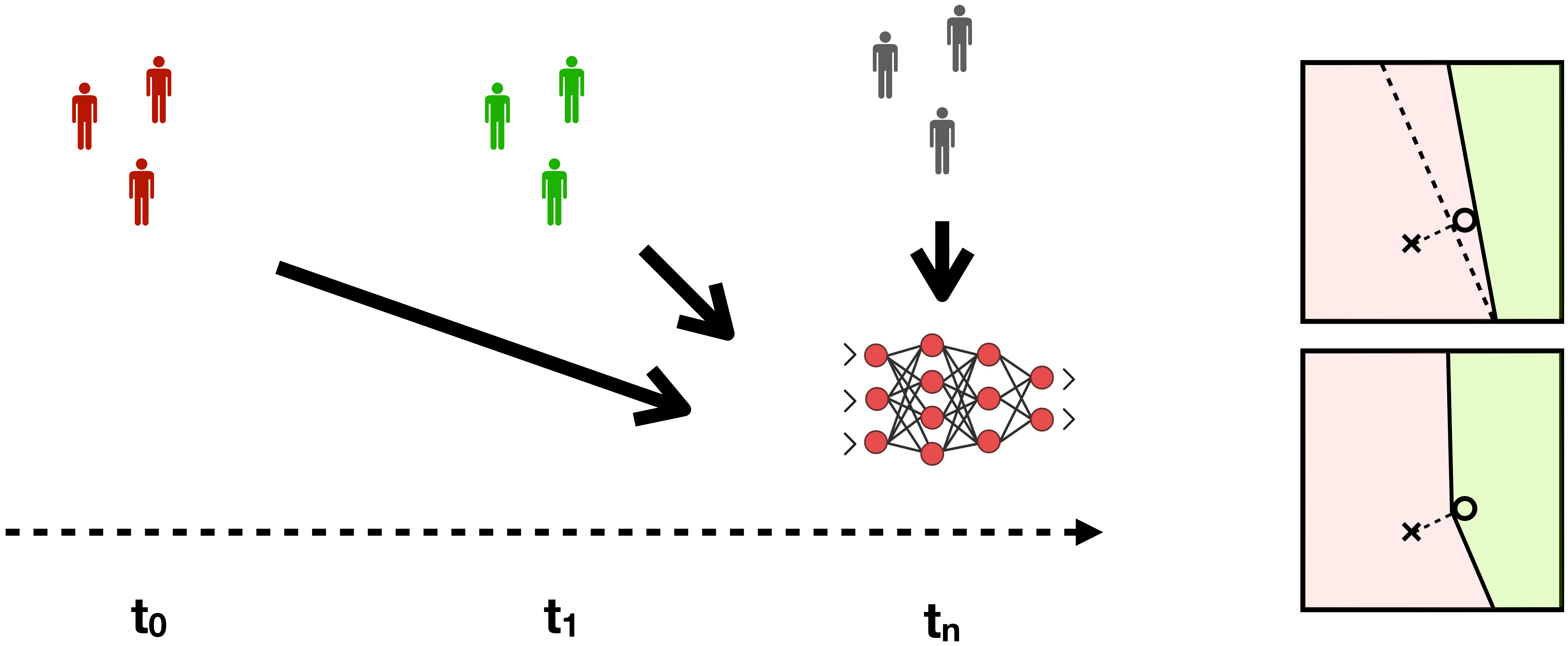
Ferrario and Loi proposed an augmentation technique to mitigate the issue



The Robustness of Counterfactual Explanations Over Time. Ferrario and Loi, IEEE Access, 2022.

Solutions

Ferrario and Loi proposed an augmentation technique to mitigate the issue



The Robustness of Counterfactual Explanations Over Time. Ferrario and Loi, IEEE Access, 2022.

Solutions

Upadhyay et al use a minmax formulation to inject model robustness

Towards Robust and Reliable Algorithmic Recourse. Upadhyay et al, NeurIPS, 2021.

Solutions

Upadhyay et al use a minmax formulation to inject model robustness

- Assume the existence of a **set of plausible model shifts Δ**

Towards Robust and Reliable Algorithmic Recourse. Upadhyay et al, NeurIPS, 2021.

Solutions

Upadhyay et al use a minmax formulation to inject model robustness

- Assume the existence of a **set of plausible model shifts Δ**
- Use $\mathcal{M}_\delta$ to denote perturbed version of $\mathcal{M}$ under $\delta \in \Delta$

$$\arg\min_{x} \arg\max_{\delta \in \Delta} \ell(\mathcal{M}_\delta(x), 1 - c) + \lambda \cdot d(x_F, x)$$

Towards Robust and Reliable Algorithmic Recourse. Upadhyay et al, NeurIPS, 2021.

Solutions

Upadhyay et al use a minmax formulation to inject model robustness

- Assume the existence of a **set of plausible model shifts Δ**
- Use $\mathcal{M}_\delta$ to denote perturbed version of $\mathcal{M}$ under $\delta \in \Delta$

$$\arg\min_{x} \boxed{\arg\max_{\delta \in \Delta} \ell(\mathcal{M}_\delta(x), 1 - c)} + \lambda \cdot d(x_F, x)$$

Towards Robust and Reliable Algorithmic Recourse. Upadhyay et al, NeurIPS, 2021.

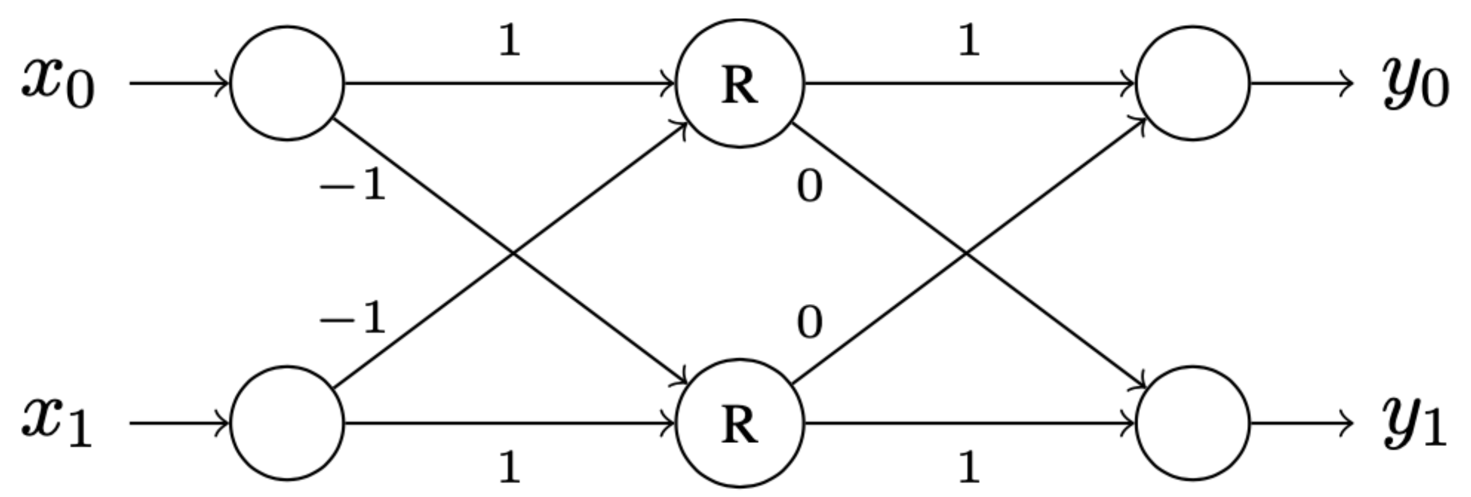
Solutions

Jiang et al use interval abstractions to obtain formal robustness guarantees

Formalising the Robustness of Counterfactual Explanations for Neural Networks. Jiang et al, AAAI 2023.
Provably Robust and Plausible Counterfactual Explanations for Neural Networks via Robust Optimisation. Jiang et al, ACML 2023.

Solutions

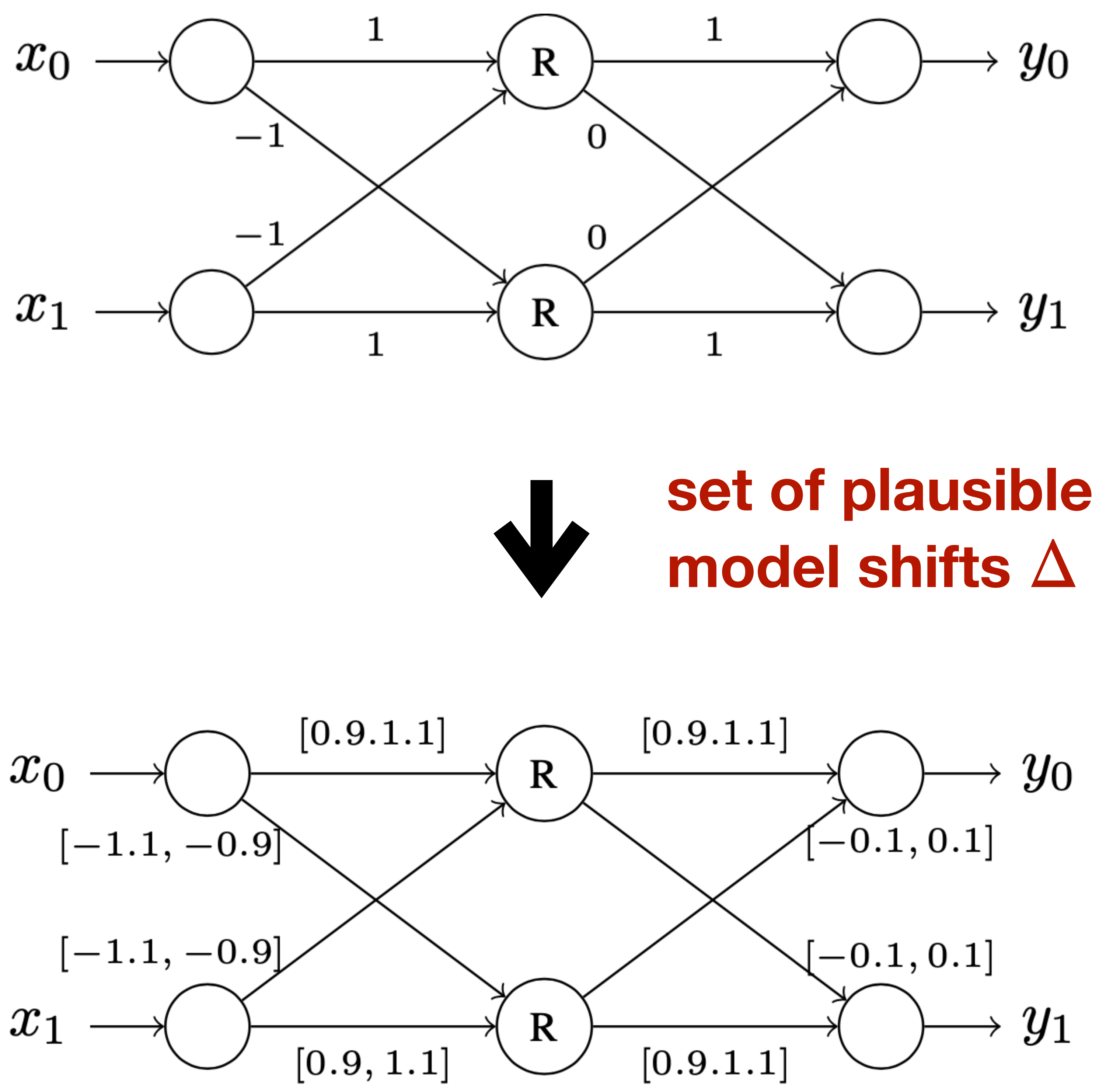
Jiang et al use interval abstractions to obtain formal robustness guarantees



Formalising the Robustness of Counterfactual Explanations for Neural Networks. Jiang et al, AAAI 2023.
Provably Robust and Plausible Counterfactual Explanations for Neural Networks via Robust Optimisation. Jiang et al, ACML 2023.

Solutions

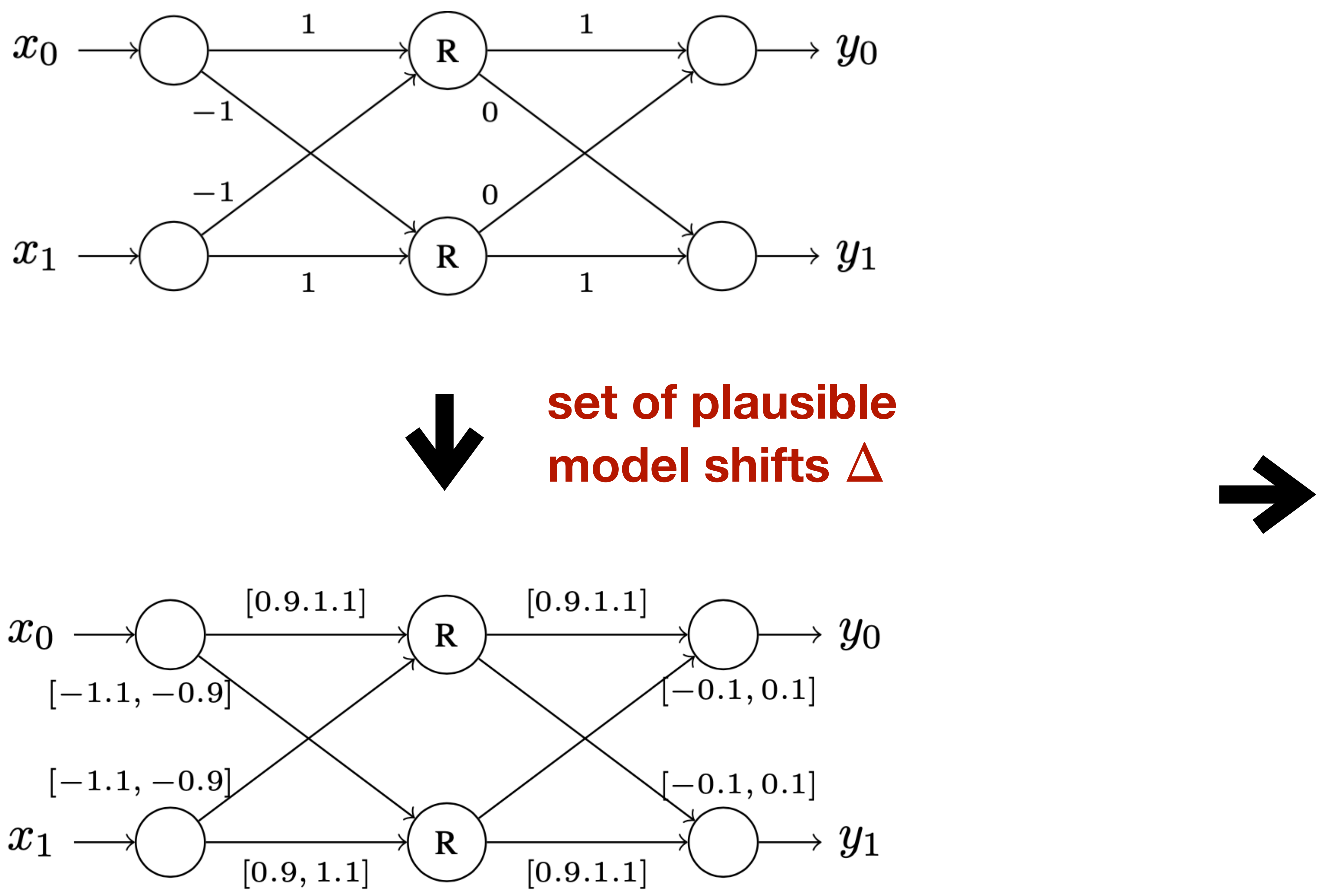
Jiang et al use interval abstractions to obtain formal robustness guarantees

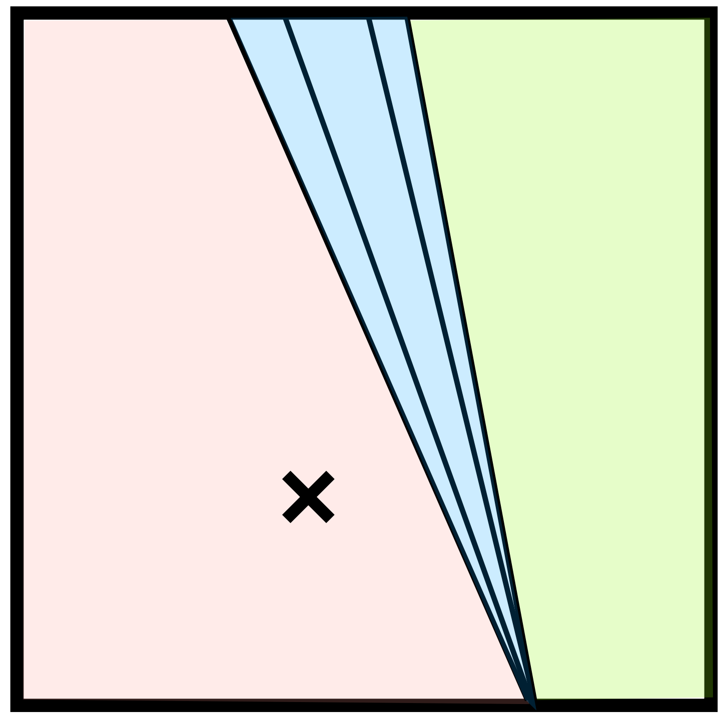


Formalising the Robustness of Counterfactual Explanations for Neural Networks. Jiang et al, AAAI 2023.
Provably Robust and Plausible Counterfactual Explanations for Neural Networks via Robust Optimisation. Jiang et al, ACML 2023.

Solutions

Jiang et al use interval abstractions to obtain formal robustness guarantees

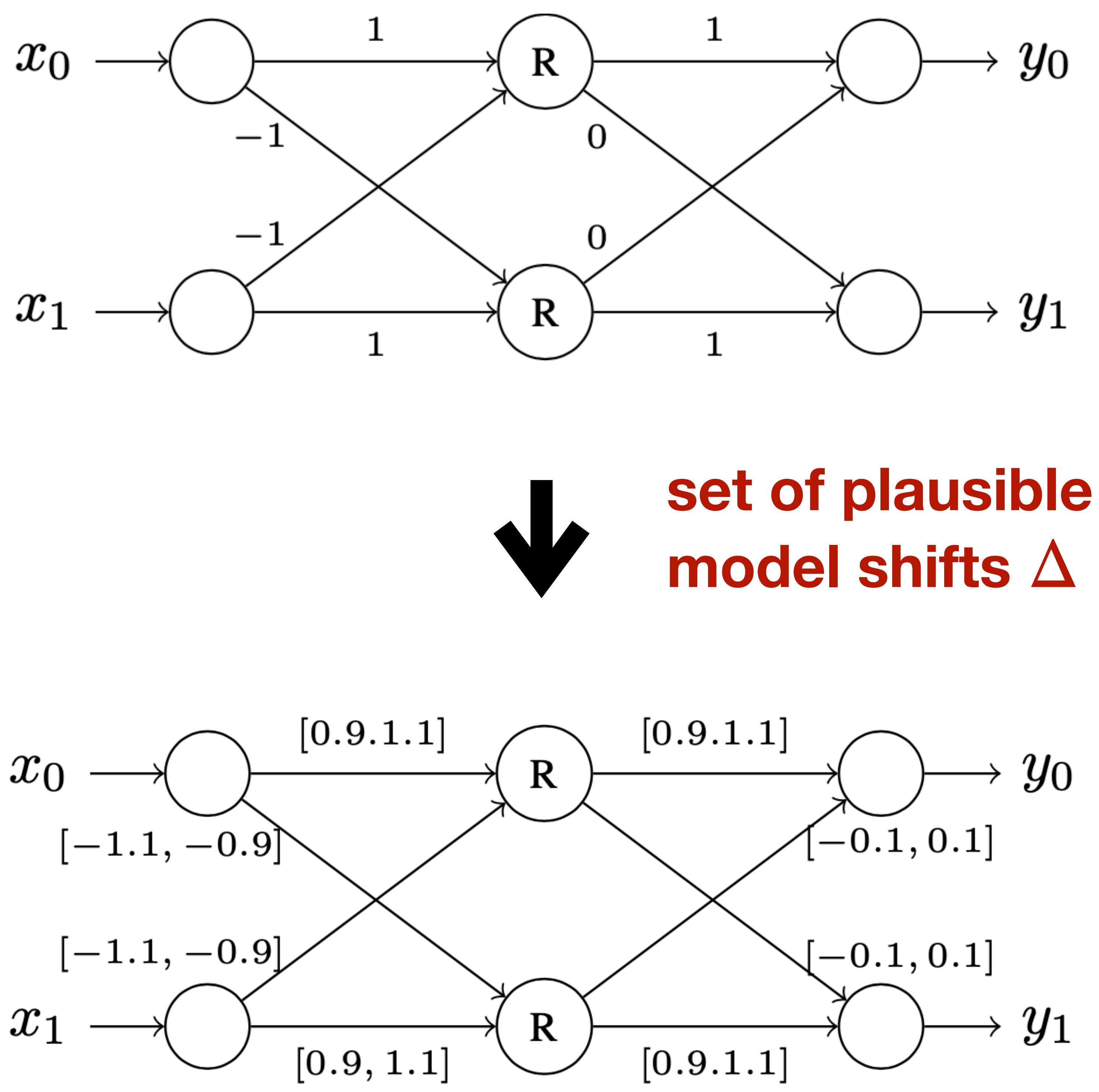


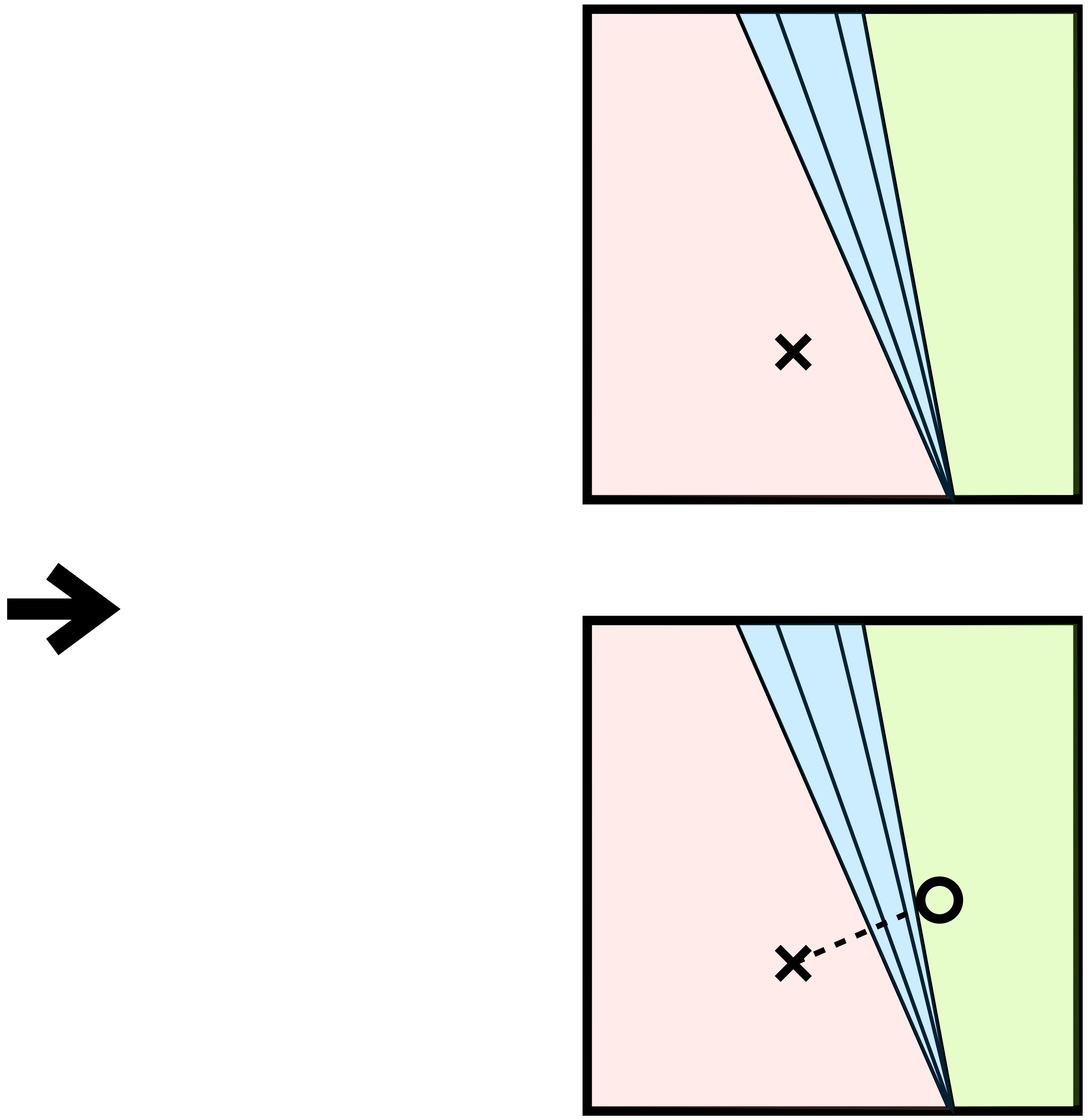


Formalising the Robustness of Counterfactual Explanations for Neural Networks. Jiang et al, AAAI 2023.
Provably Robust and Plausible Counterfactual Explanations for Neural Networks via Robust Optimisation. Jiang et al, ACML 2023.

Solutions

Jiang et al use interval abstractions to obtain formal robustness guarantees





Formalising the Robustness of Counterfactual Explanations for Neural Networks. Jiang et al, AAAI 2023.
Provably Robust and Plausible Counterfactual Explanations for Neural Networks via Robust Optimisation. Jiang et al, ACML 2023.

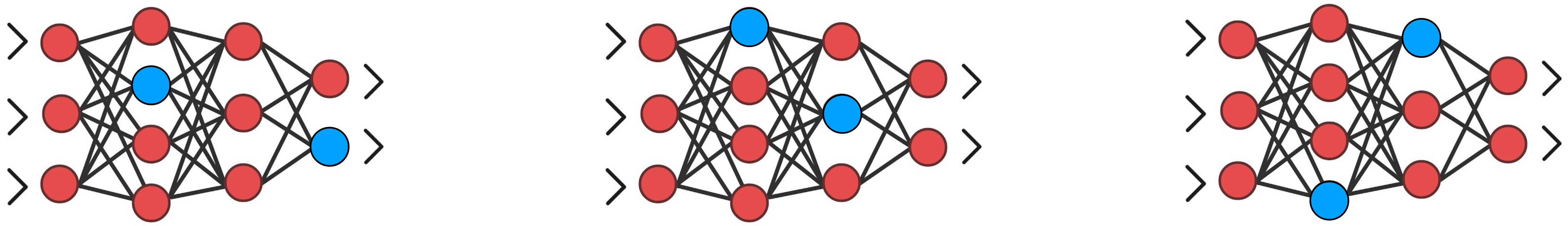
Brittle explanations ahead!

Threats

1. Input perturbations
2. Model perturbations
3. **Model multiplicity**
4. Noisy execution

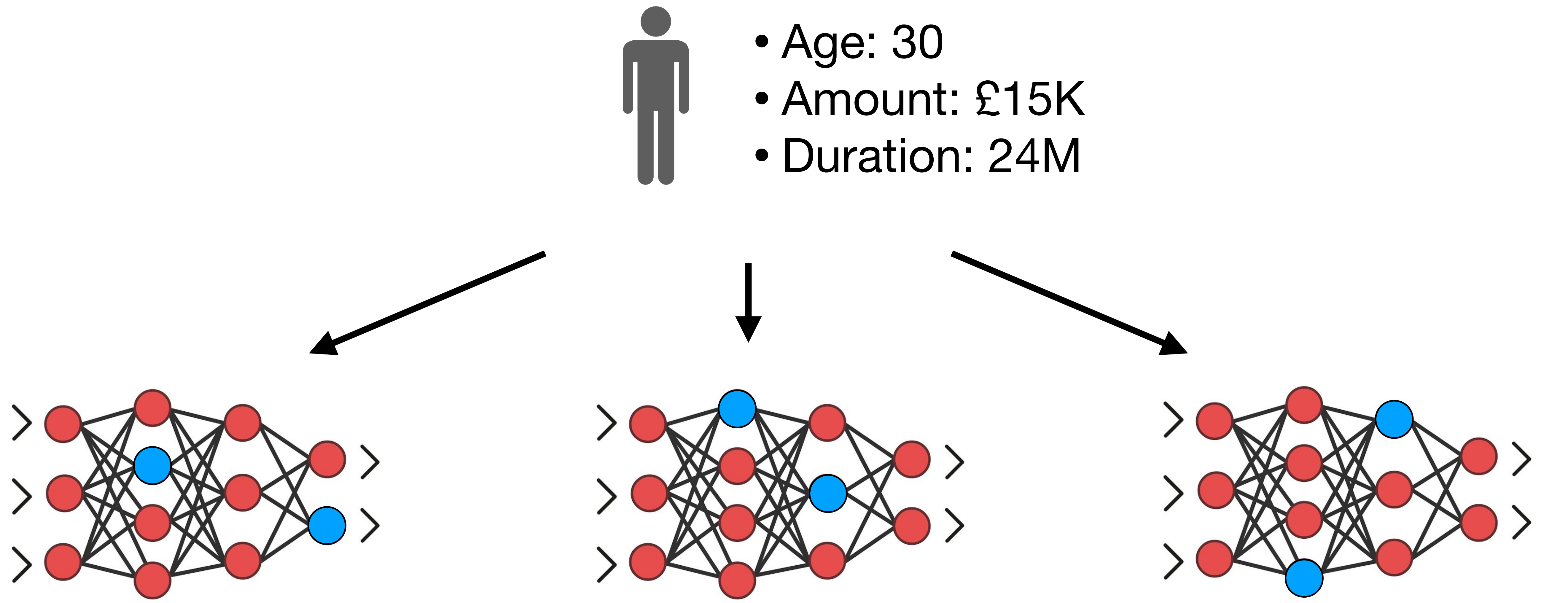
Model multiplicity

Situation where models of equal accuracy differ in the process by which they reach a given prediction



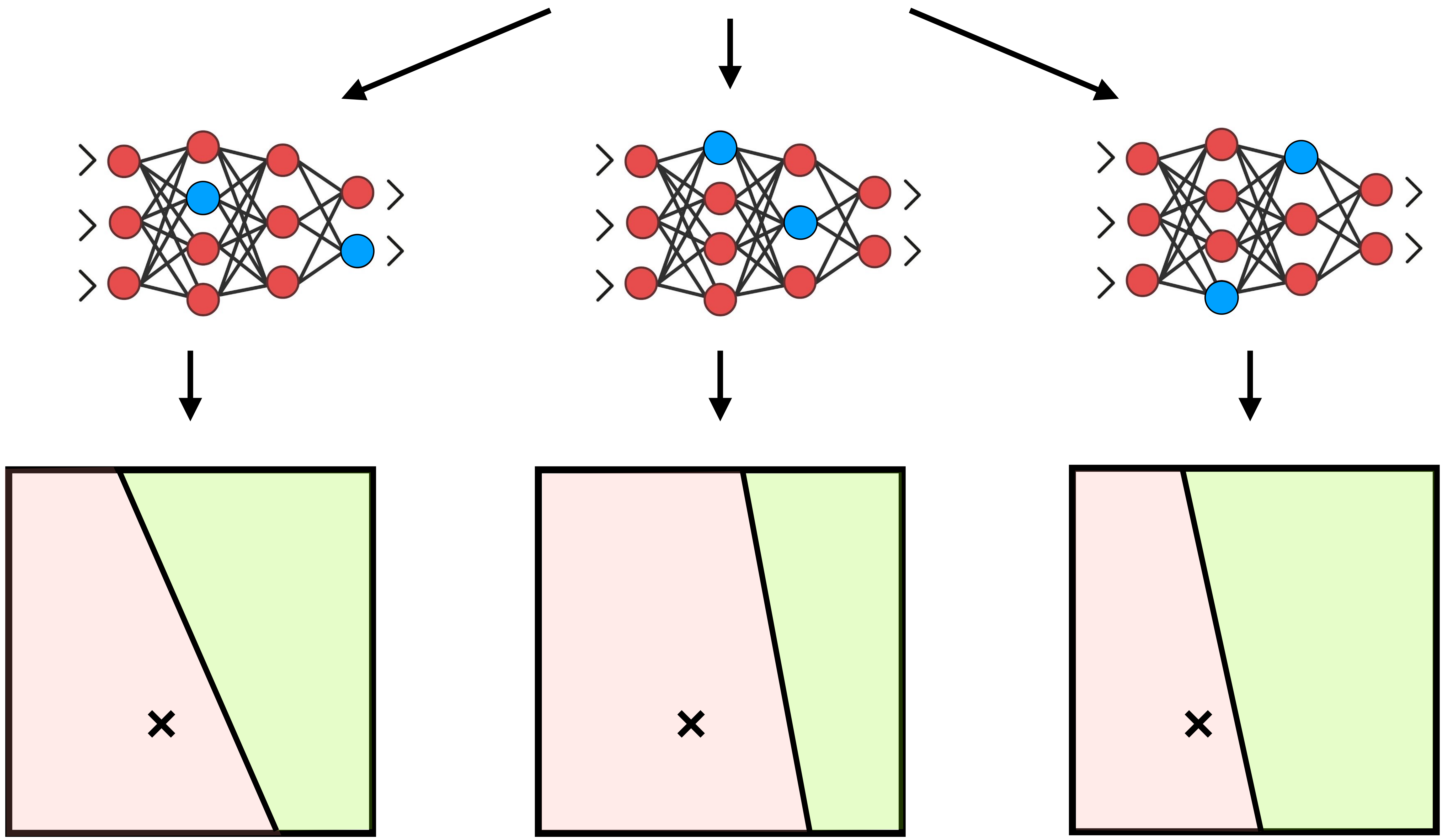
Model Multiplicity: Opportunities, Concerns, and Solutions. Black et al, ACM FAccT’22.

Model multiplicity

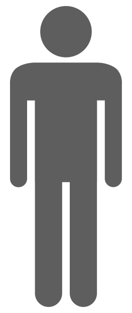


Model multiplicity

- Age: 30
- Amount: £15K
- Duration: 24M

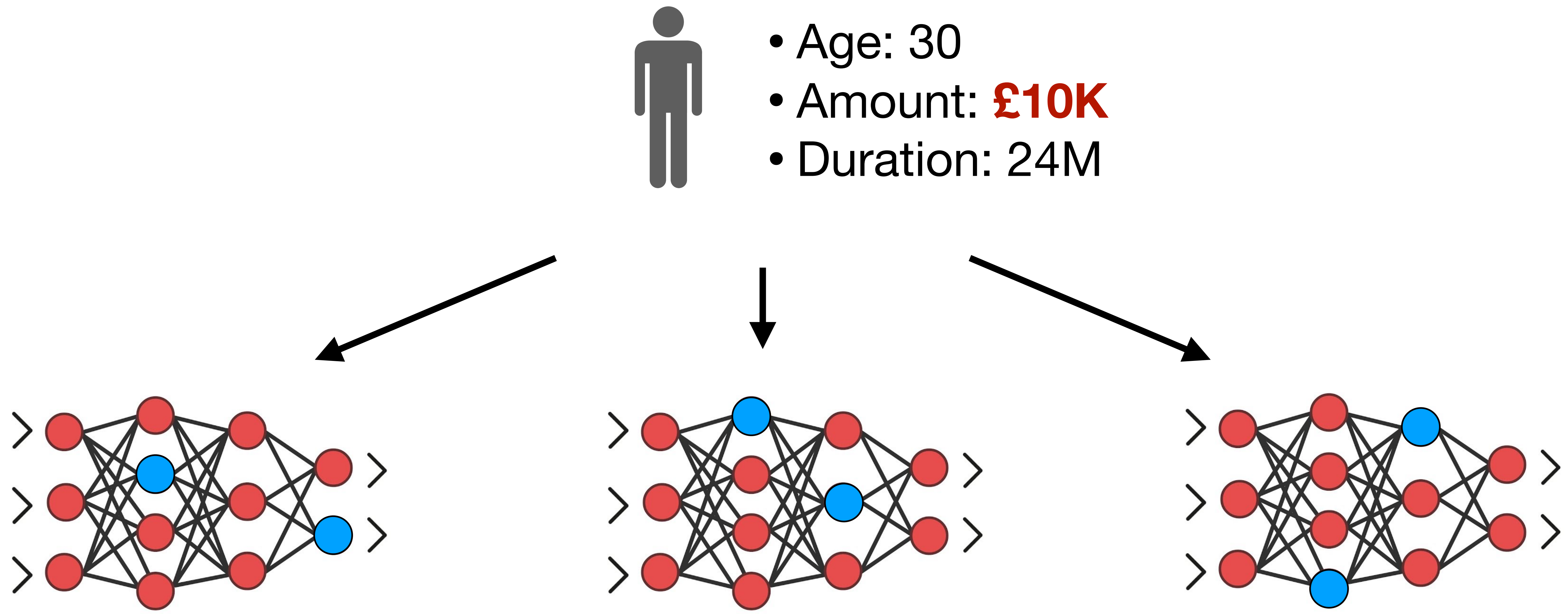


Model multiplicity



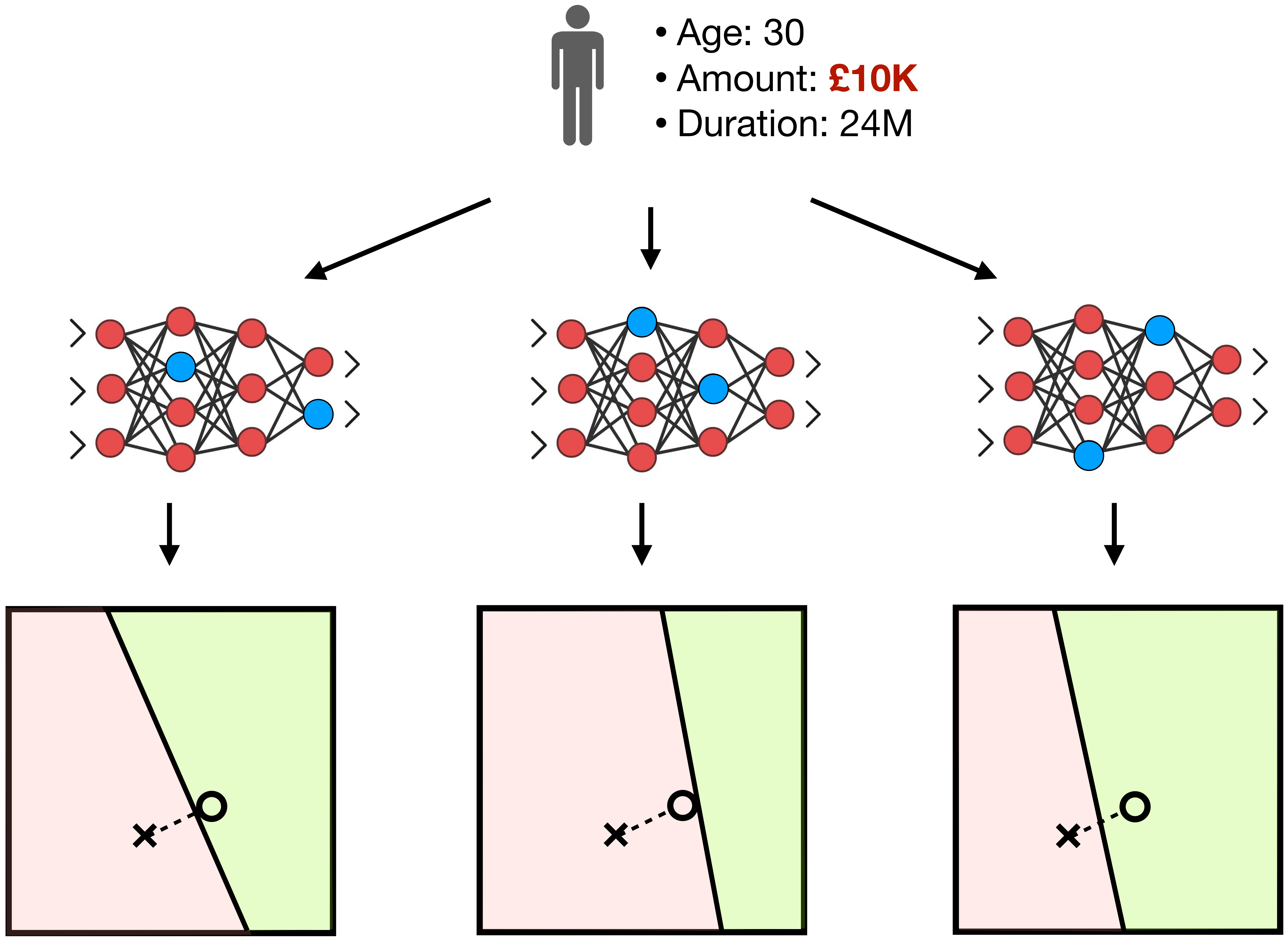
- Age: 30
- Amount: **£10K**
- Duration: 24M

Model multiplicity

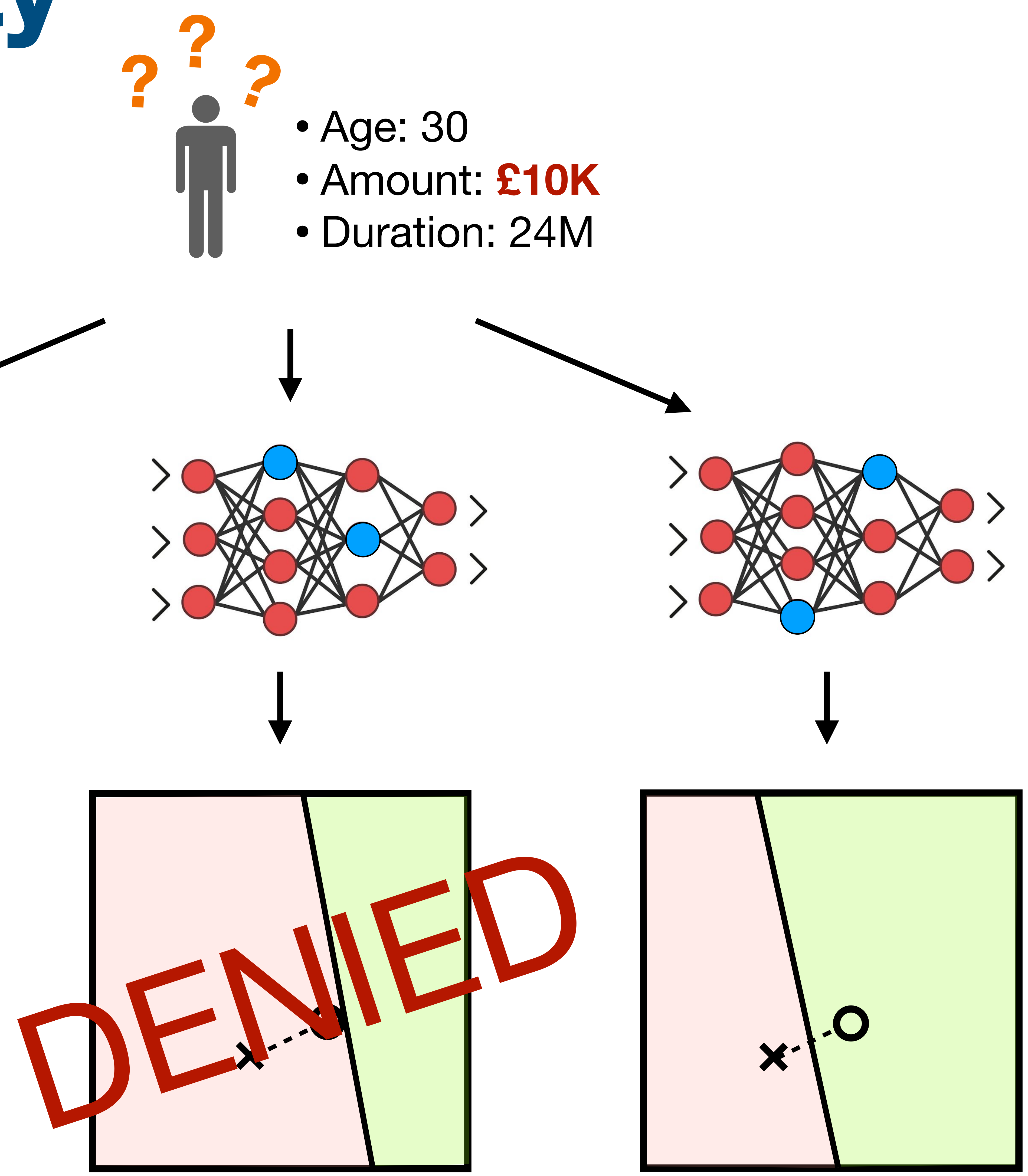


Model multiplicity

- Age: 30
- Amount: **£10K**
- Duration: 24M

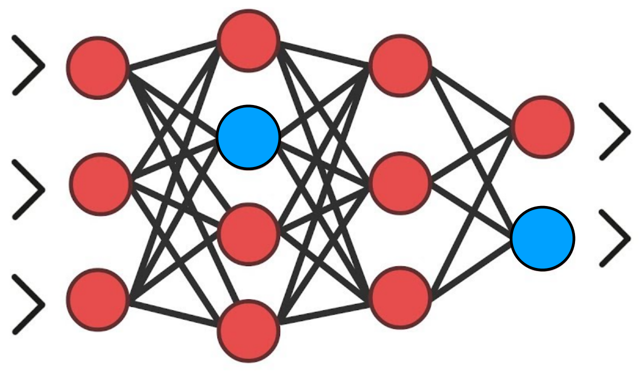


Model multiplicity

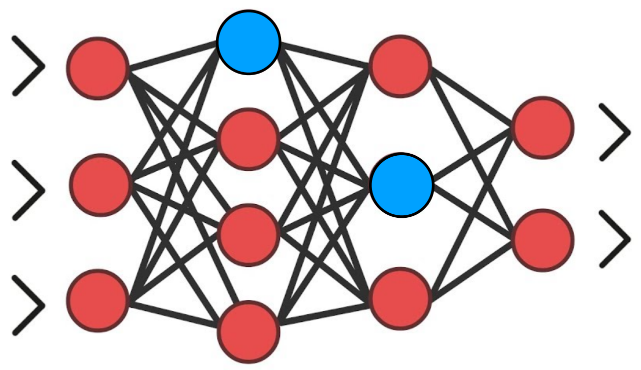


Implications

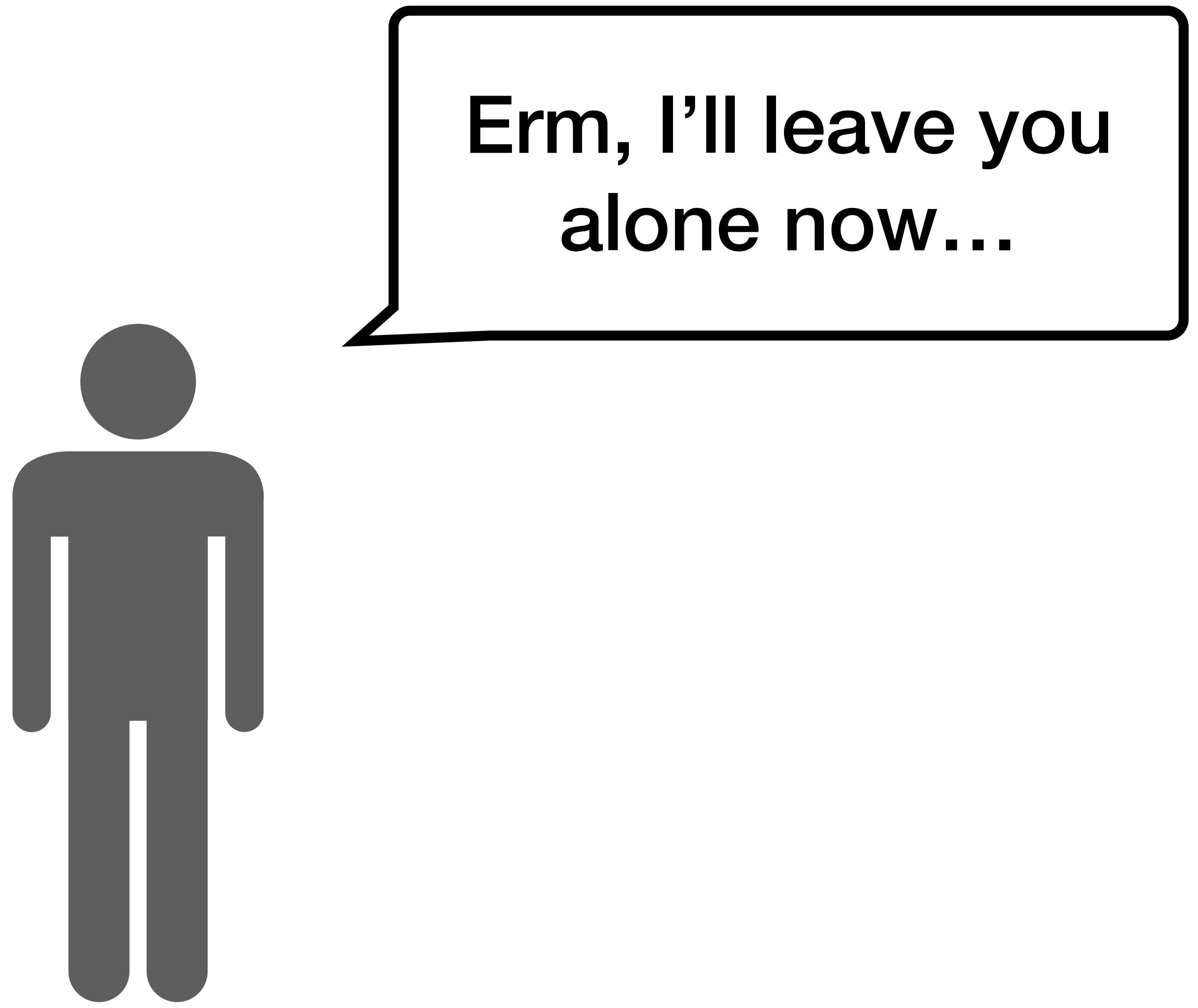
- Disagreeing models might raise concerns about the **justifiability** of CXs
- Different models might offer **better/worse recourse** options



Increase by £50



That’s not enough!



Solutions

Black et al present an extensive discussion on model multiplicity

- Not targeting CXs specifically but also applicable to XAI

Model Multiplicity: Opportunities, Concerns, and Solutions. Black et al, ACM FAccT'22.

Solutions

Black et al present an extensive discussion on model multiplicity

- Not targeting CXs specifically but also applicable to XAI

They propose some approaches to deal with multiplicity:

Model Multiplicity: Opportunities, Concerns, and Solutions. Black et al, ACM FAccT'22.

Solutions

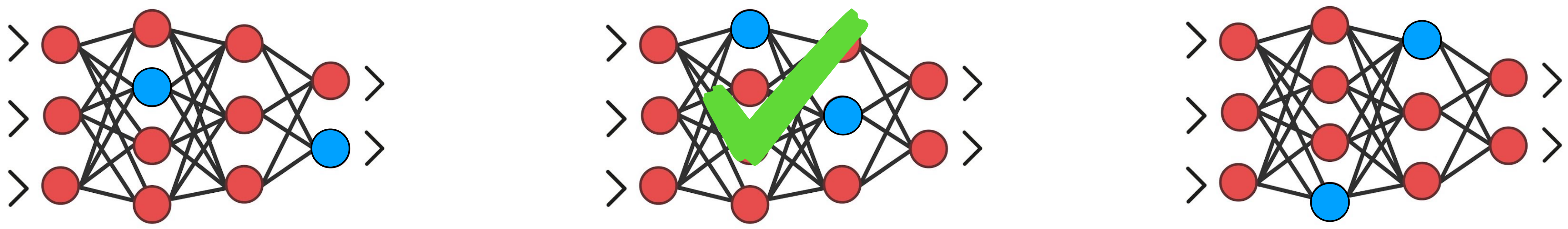
Black et al present an extensive discussion on model multiplicity

- Not targeting CXs specifically but also applicable to XAI

They propose some approaches to deal with multiplicity:

- **Meta-rules**

“Always choose the model that has at least 95% accuracy”



Model Multiplicity: Opportunities, Concerns, and Solutions. Black et al, ACM FAccT’22.

Solutions

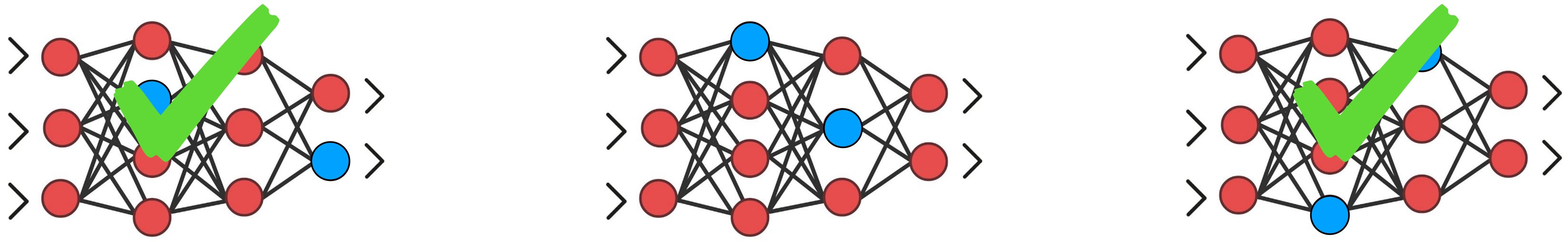
Black et al present an extensive discussion on model multiplicity

- Not targeting CXs specifically but also applicable to XAI

They propose some approaches to deal with multiplicity:

- Meta-rules
- **Majority voting**

"Two out of three agree, they must be correct"



Model Multiplicity: Opportunities, Concerns, and Solutions. Black et al, ACM FAccT'22.

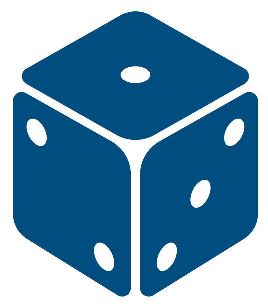
Solutions

Black et al present an extensive discussion on model multiplicity

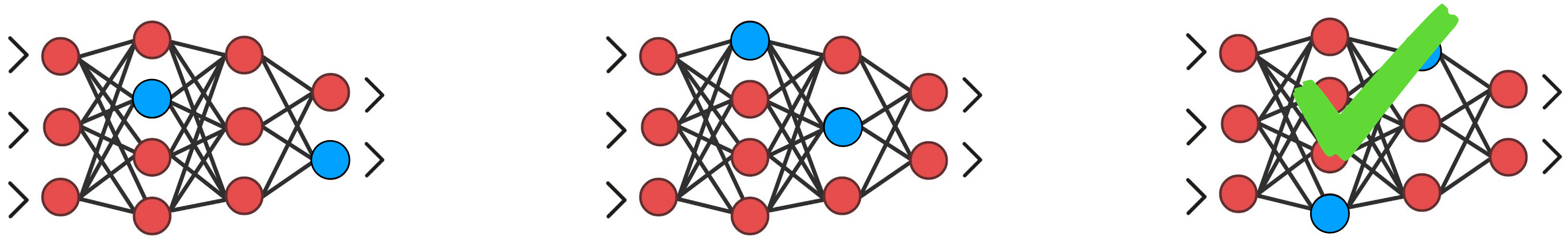
- Not targeting CXs specifically but also applicable to XAI

They propose some approaches to deal with multiplicity:

- Meta-rules
- Majority voting
- **Randomised choice**



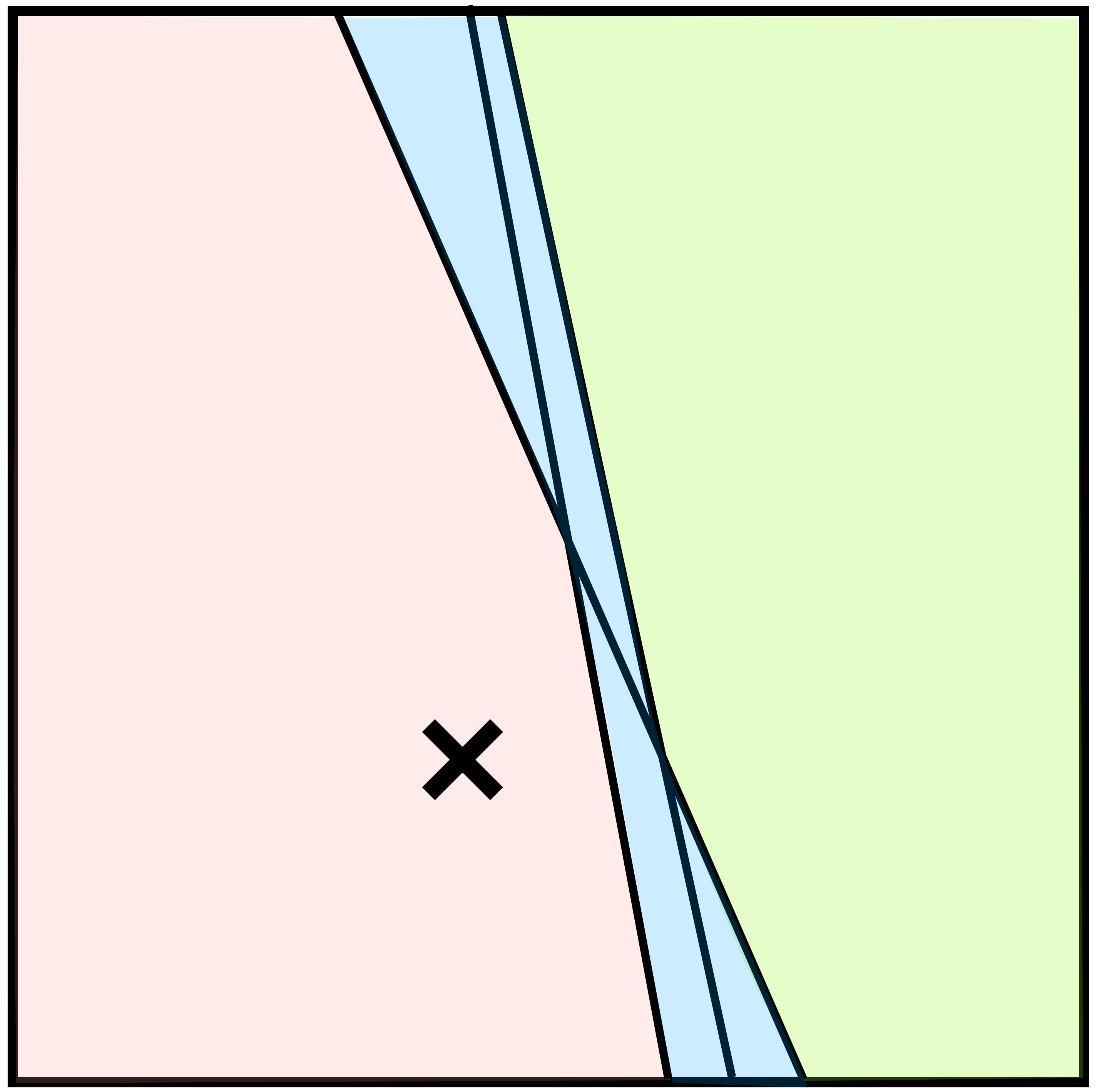
“Sample a model and use it”



Model Multiplicity: Opportunities, Concerns, and Solutions. Black et al, ACM FAccT’22.

Solutions

Pawelczyk et al analyse robustness of CXs under model multiplicity:

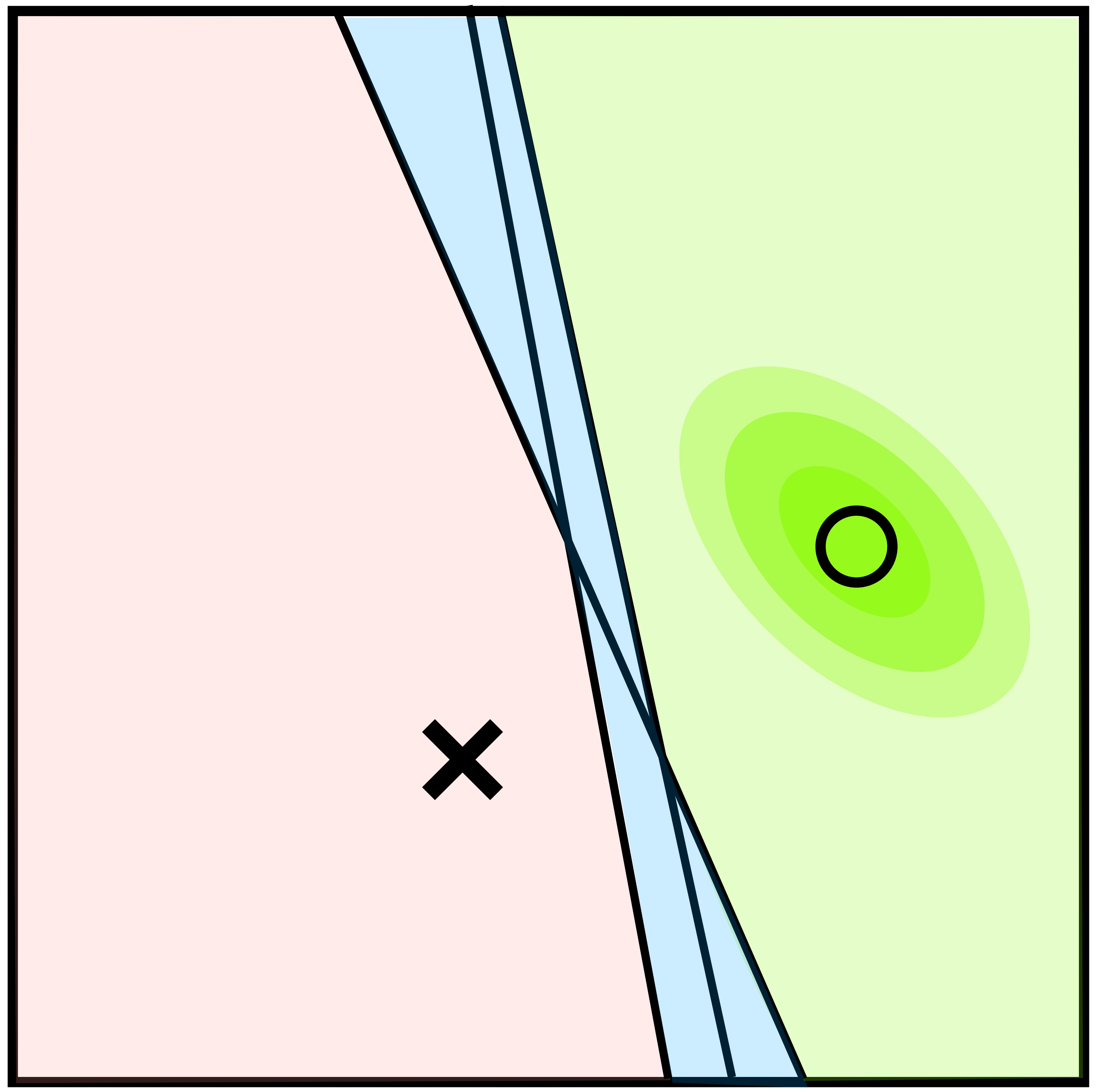


On Counterfactual Explanations under Predictive Multiplicity. Pawelczyk et al, UAI 2020.

Solutions

Pawelczyk et al analyse robustness of CXs under model multiplicity:

- CXs on **data manifold** are more **robust**

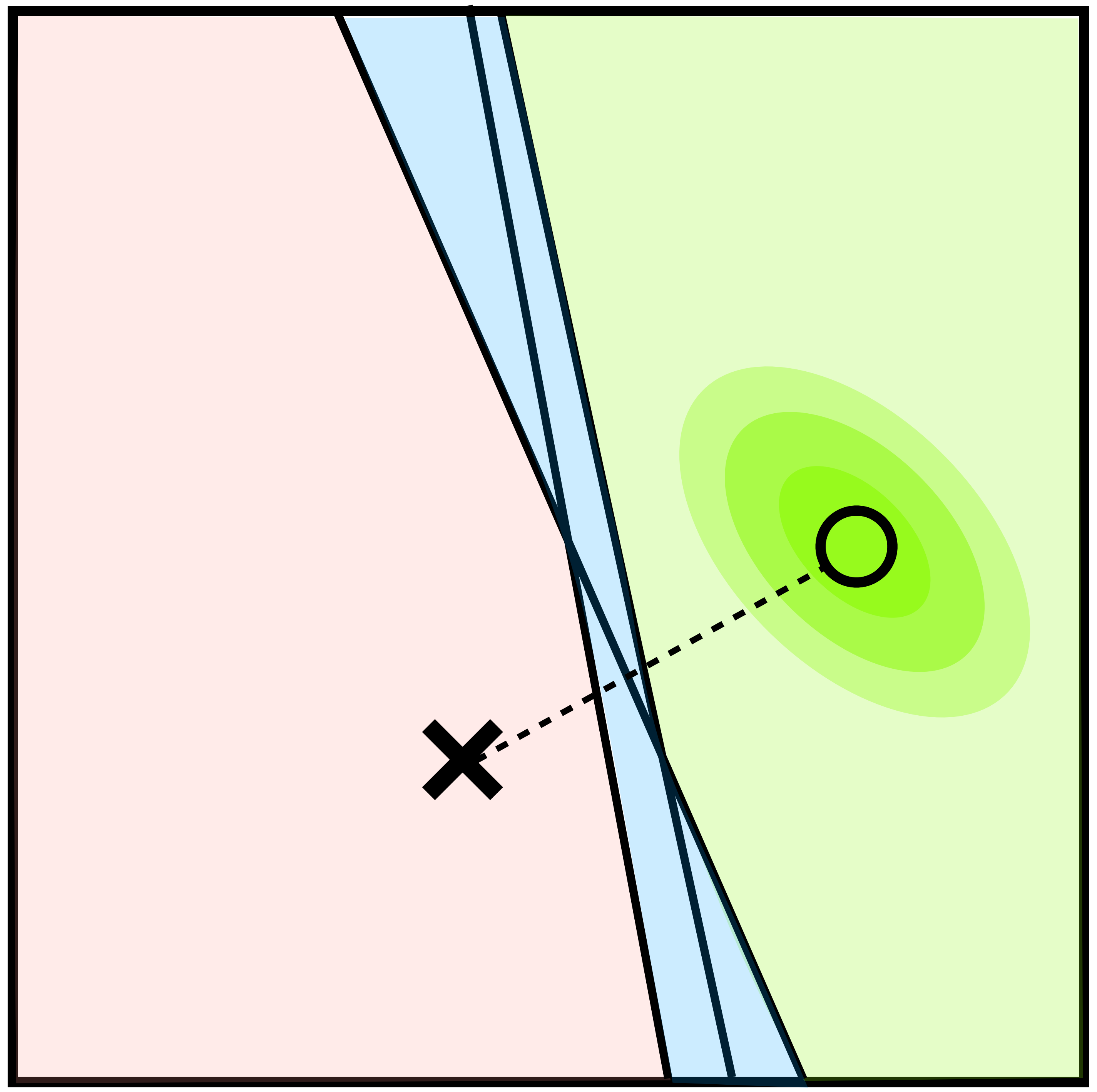


On Counterfactual Explanations under Predictive Multiplicity. Pawelczyk et al, UAI 2020.

Solutions

Pawelczyk et al analyse robustness of CXs under model multiplicity:

- CXs on **data manifold** are more **robust**
- Robust CXs are **more expensive**



On Counterfactual Explanations under Predictive Multiplicity. Pawelczyk et al, UAI 2020.

Solutions

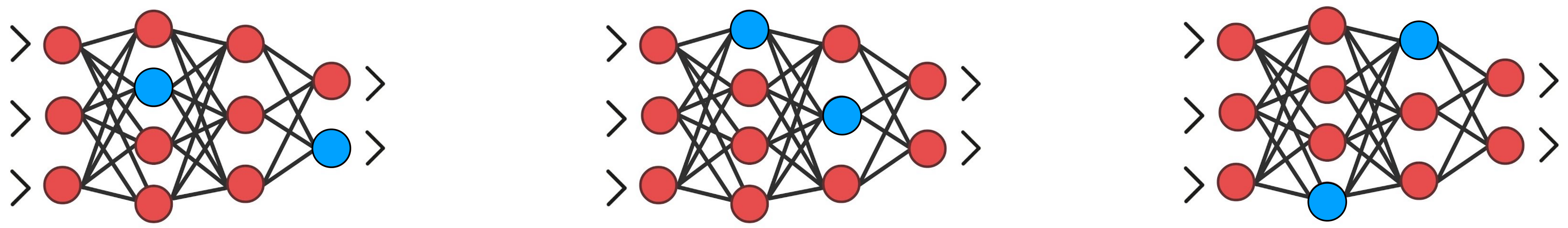
Leofante et al present an approach to **generate robust CXs** under multiplicity

Counterfactual Explanations and Model Multiplicity: a Relational Verification View. Leofante et al, KR 2023.

Solutions

Leofante et al present an approach to **generate robust CXs** under multiplicity

- Assumes **pre-defined set** of models

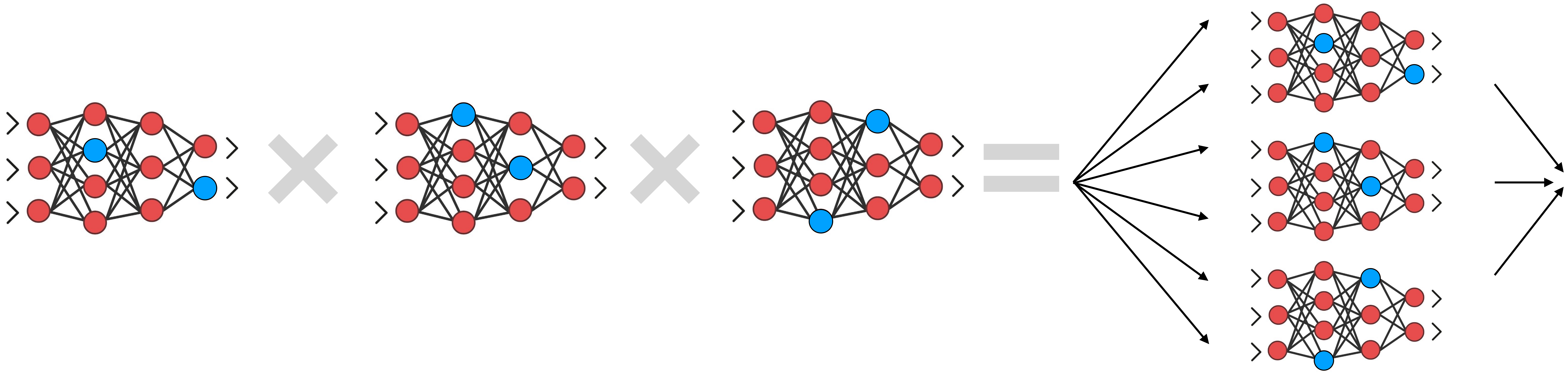


Counterfactual Explanations and Model Multiplicity: a Relational Verification View. Leofante et al, KR 2023.

Solutions

Leofante et al present an approach to **generate robust CXs** under multiplicity

- Assumes **pre-defined set** of models
- Builds **product network** to reason under multiplicity in one go



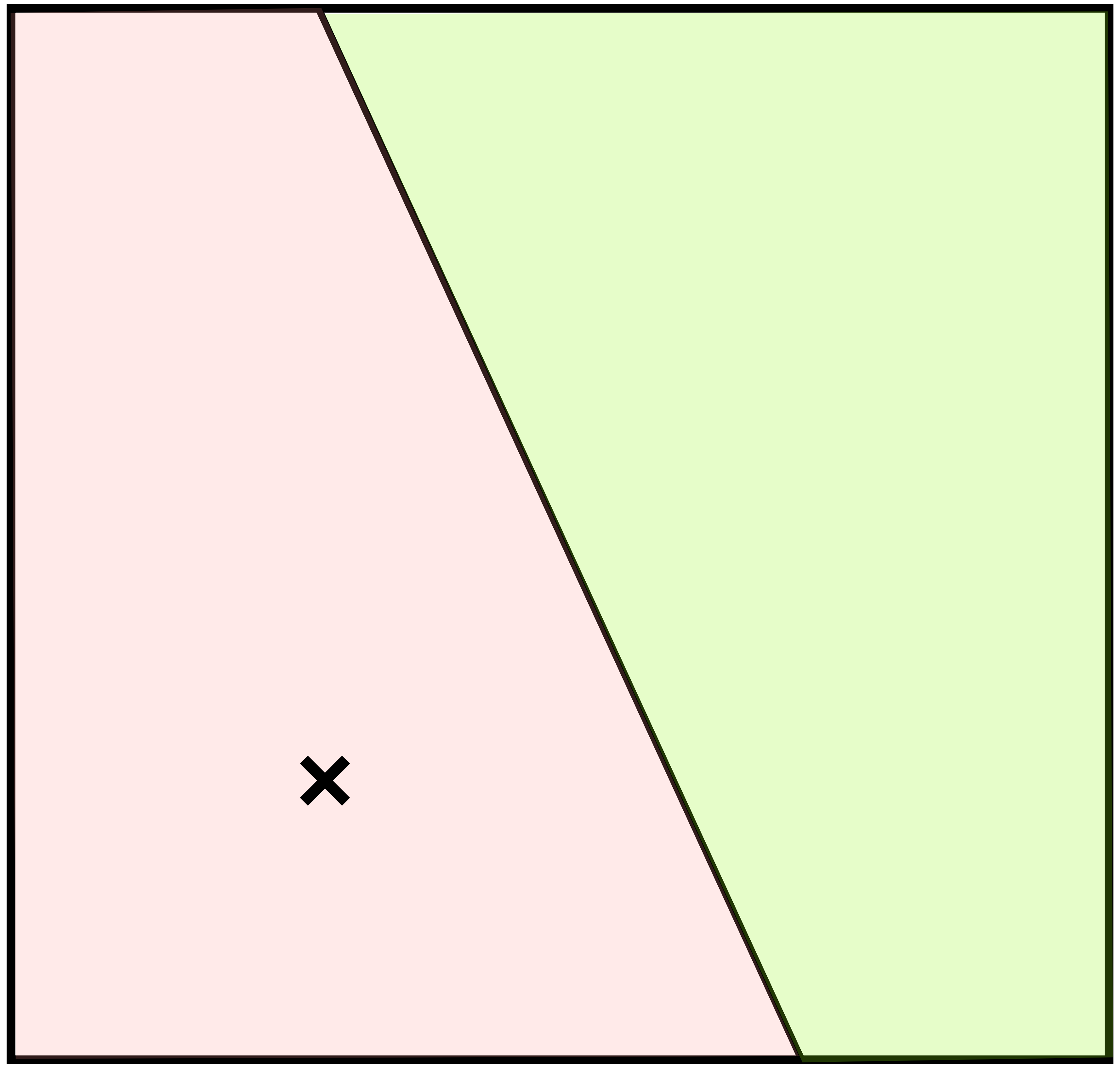
Counterfactual Explanations and Model Multiplicity: a Relational Verification View. Leofante et al, KR 2023.

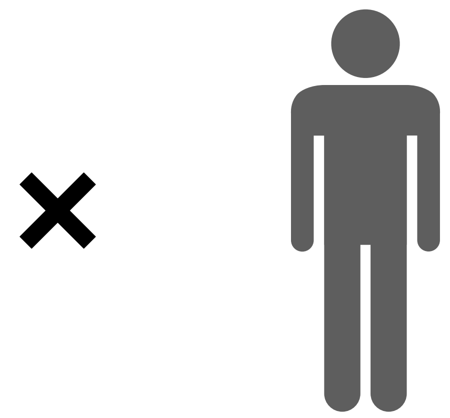
Brittle explanations ahead!

Threats

1. Input perturbations
2. Model perturbations
3. Model multiplicity
4. **Noisy execution**

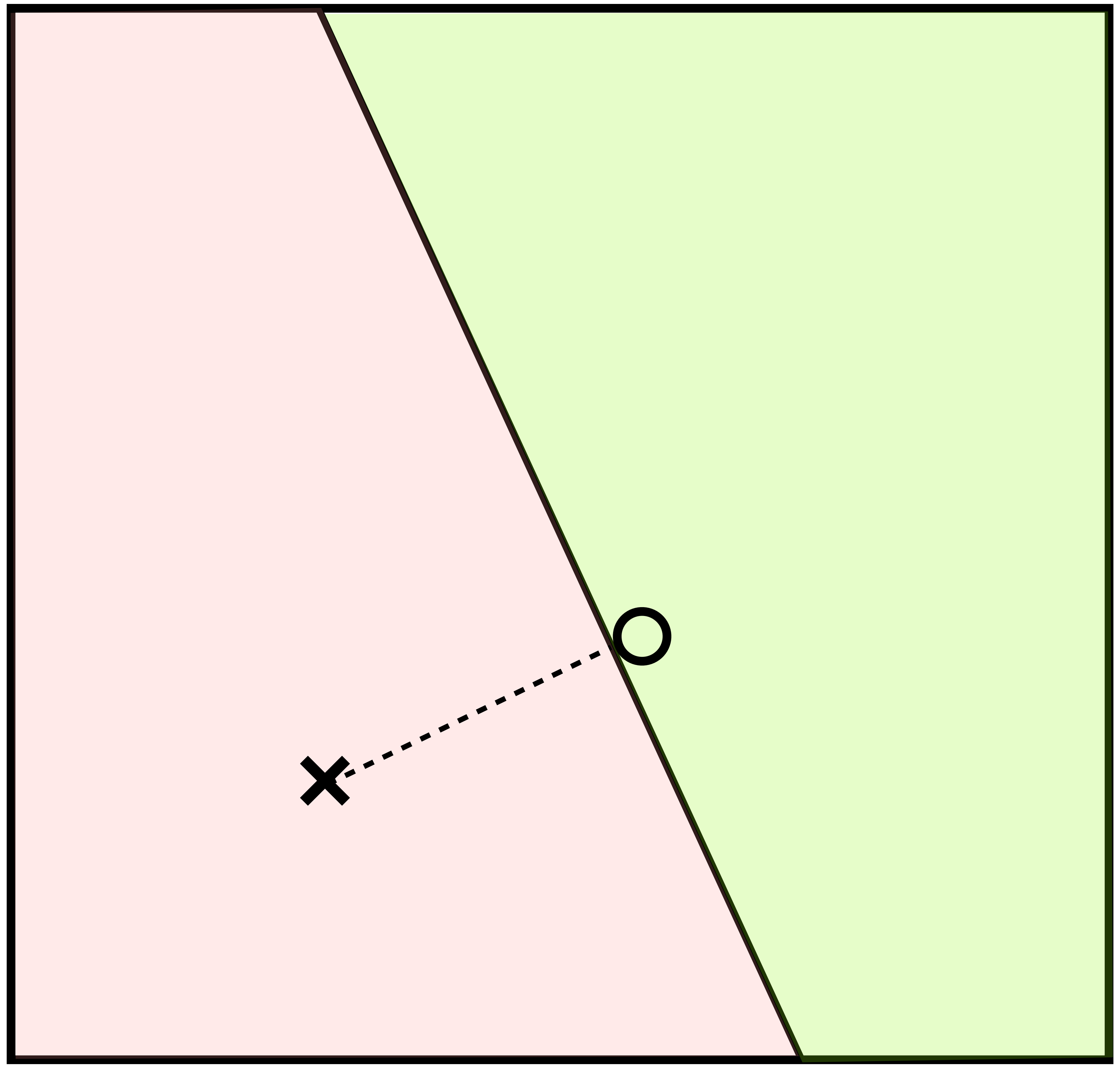
Noisy execution





- Age: 30
- Amount: **£15K**
- Duration: 24M

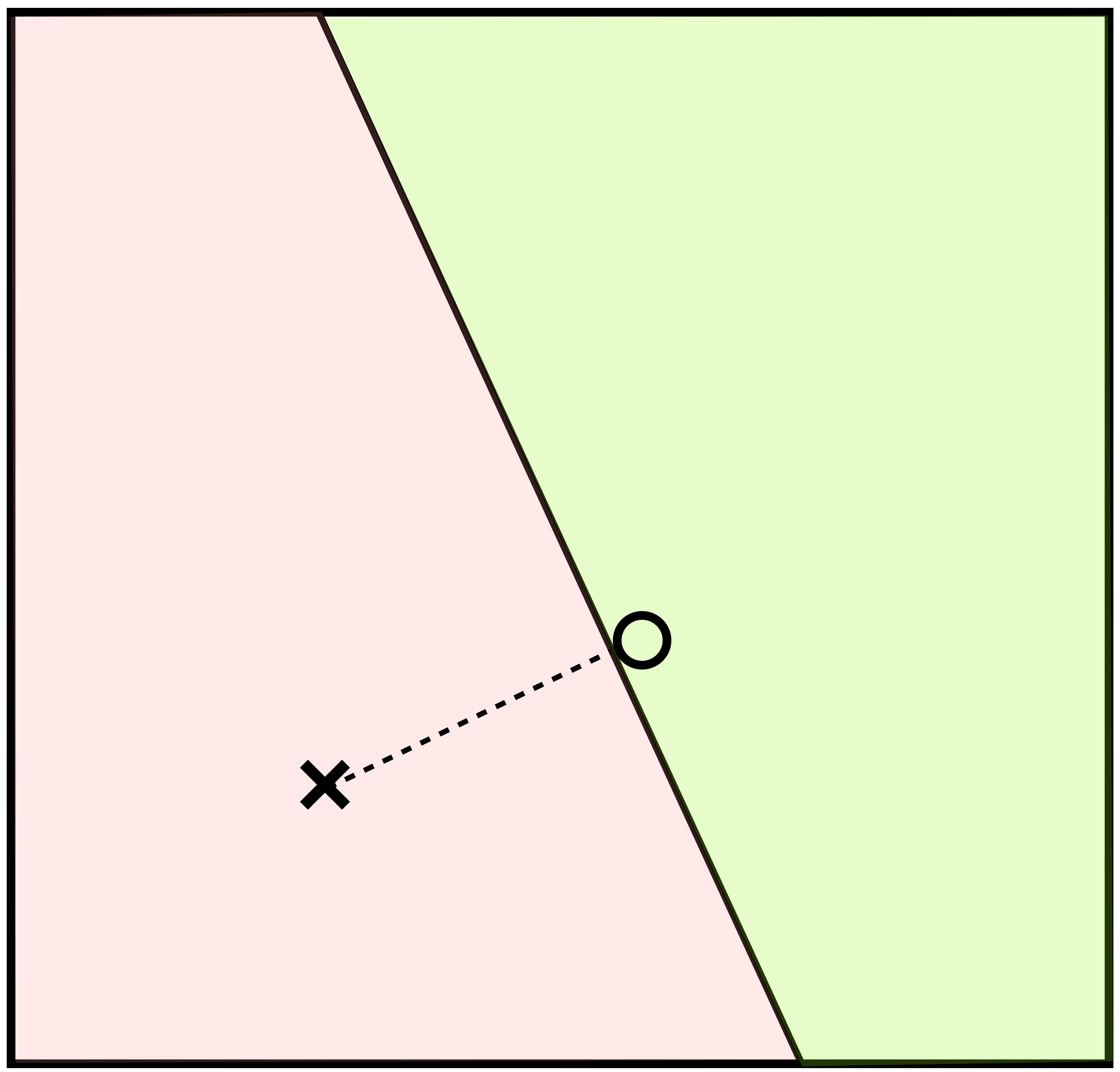
Noisy execution



- Age: 30
- Amount: **£15K**
- Duration: 24M

- Age: 30
- Amount: **£10K**
- Duration: 24M

Noisy execution



×

- Age: 30
- Amount: **£15K**
- Duration: 24M

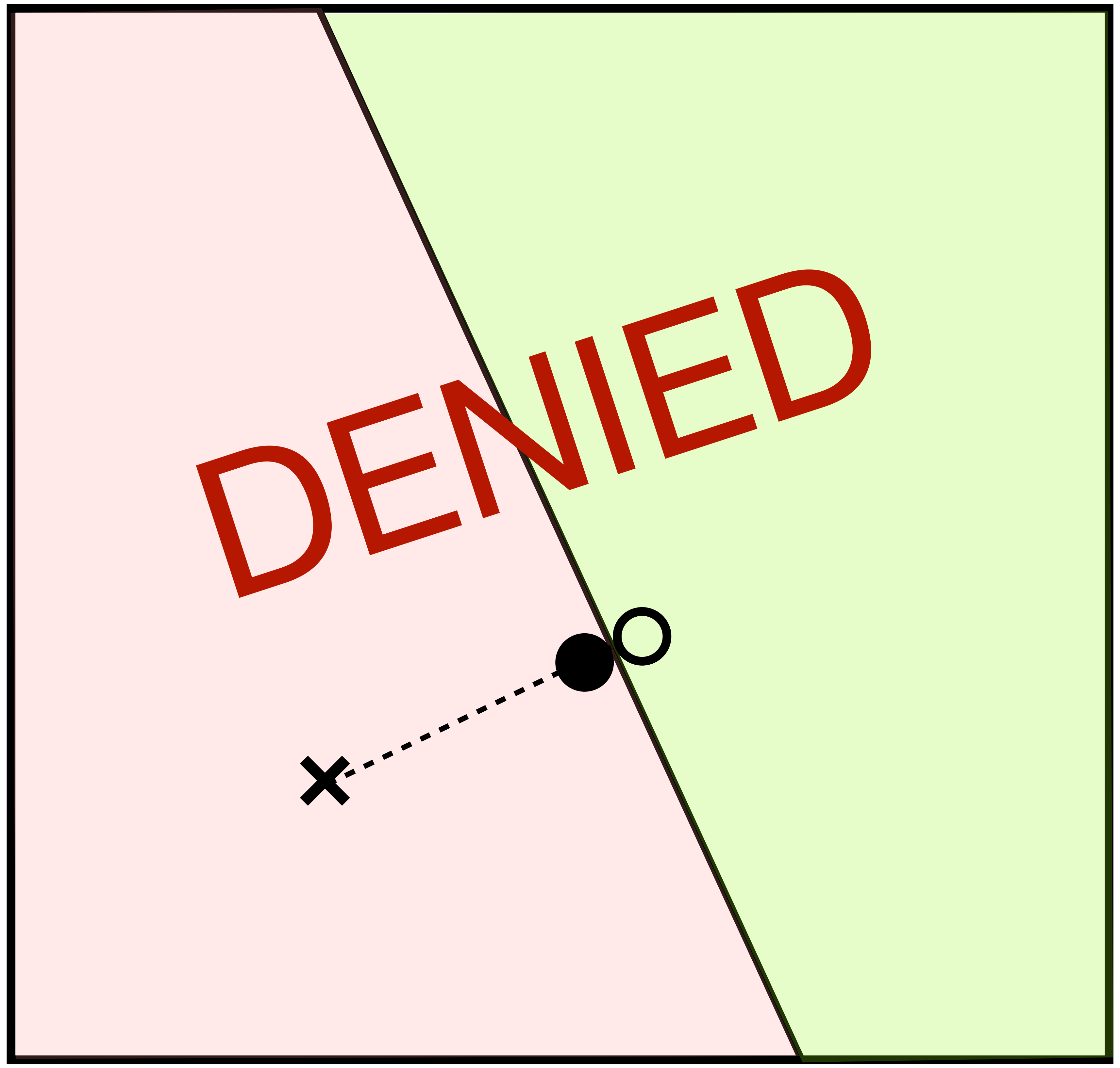
○

- Age: 30
- Amount: **£10K**
- Duration: 24M

●

- Age: 30
- Amount: **£9.9K**
- Duration: 24M

Noisy execution



- Age: 30
- Amount: **£15K**
- Duration: 24M

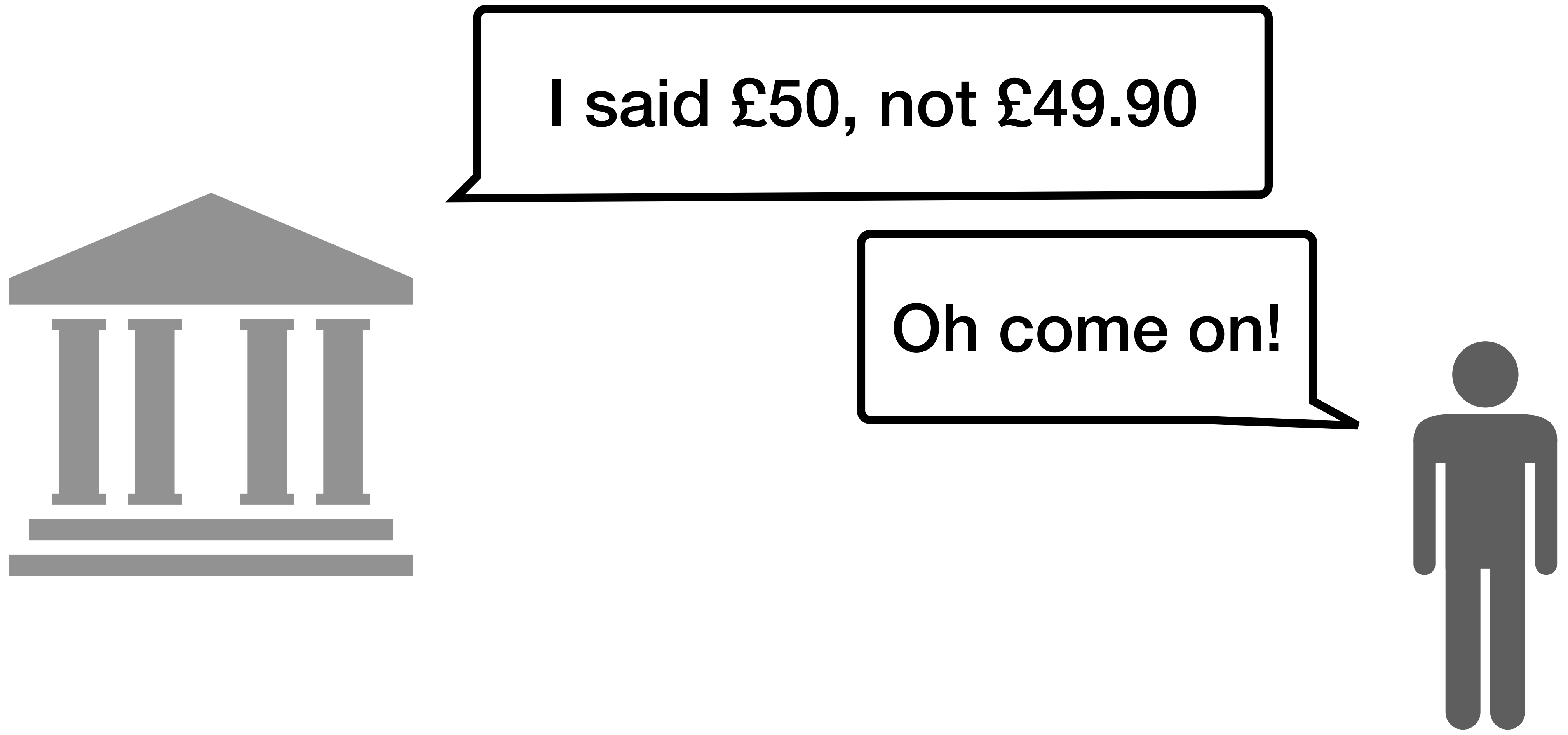
- Age: 30
- Amount: **£10K**
- Duration: 24M

- Age: 30
- Amount: **£9.9K**
- Duration: 24M

Implications

Recourses are often noisily implemented in real-world settings

- Noise may **invalidate** CX
- **Jeopardise** explanatory function
- **Reduce** trust



Manipulation-Proof Machine Learning. Björkegren et al, arxiv preprint https://arxiv.org/abs/2004.03865, 2020.

Solutions

Pawelczyk et al propose to account for noisy execution during CX generation

Probabilistically Robust Recourse: Navigating the Trade-offs between Costs and Robustness in Algorithmic Recourse. Pawelczyk et al, ICLR 2023.

Solutions

Pawelczyk et al propose to account for noisy execution during CX generation

- Given input x_F, CX x and model $\mathscr{M}$

Solutions

Pawelczyk et al propose to account for noisy execution during CX generation

- Given input x_F, CX x and model $\mathcal{M}$
- Define **invalidation rate** $\Delta(x) = \mathbb{E}_{\epsilon}[\mathcal{M}(x) - \mathcal{M}(x + \epsilon)]$

Probabilistically Robust Recourse: Navigating the Trade-offs between Costs and Robustness in Algorithmic Recourse. Pawelczyk et al, ICLR 2023.

Solutions

Pawelczyk et al propose to account for noisy execution during CX generation

- Given input x_F, CX x and model $\mathscr{M}$
- Define **invalidation rate** $\Delta(x) = \mathbb{E}_\epsilon[\mathscr{M}(x) - \mathscr{M}(x+\epsilon)]$
- Define noise-aware loss $\mathscr{L}$ as

$$\lambda_1 \cdot \ell_1(\Delta(x), \rho) + \lambda_2 \cdot \ell_2(\mathscr{M}(x), 1 - c) + \lambda_3 \cdot d(x_F, x)$$

Probabilistically Robust Recourse: Navigating the Trade-offs between Costs and Robustness in Algorithmic Recourse. Pawelczyk et al, ICLR 2023.

Solutions

Pawelczyk et al propose to account for noisy execution during CX generation

- Given input x_F, CX x and model $\mathscr{M}$
- Define **invalidation rate** $\Delta(x) = \mathbb{E}_\epsilon[\mathscr{M}(x) - \mathscr{M}(x+\epsilon)]$
- Define noise-aware loss $\mathscr{L}$ as

$$\boxed{\lambda_1 \cdot \ell_1(\Delta(x), \rho)} + \lambda_2 \cdot \ell_2(\mathscr{M}(x), 1 - c) + \lambda_3 \cdot d(x_F, x)$$

Probabilistically Robust Recourse: Navigating the Trade-offs between Costs and Robustness in Algorithmic Recourse. Pawelczyk et al, ICLR 2023.

Solutions

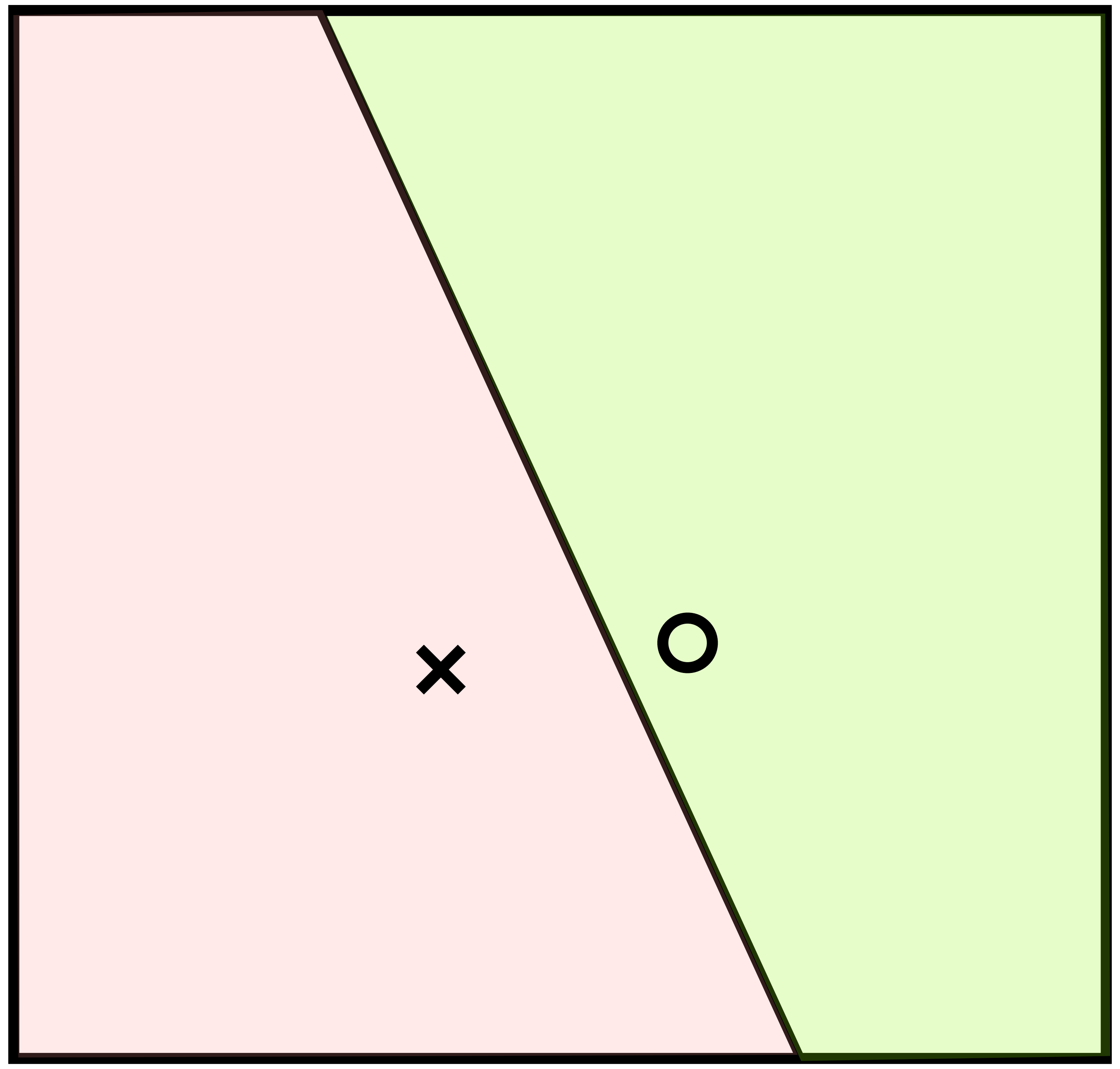
Leofante and Lomuscio use formal verification to identify robust CXs

Towards Robust Contrastive Explanations for Human-Neural Multi-agent Systems. Leofante and Lomuscio, AAMAS 2023.
Robust Explanations for Human-Neural Multi-agent Systems with Formal Verification. Leofante and Lomuscio, EUMAS 2023.

Solutions

Leofante and Lomuscio use formal verification to identify robust CXs

- Given a CX x and model $\mathscr{M}$

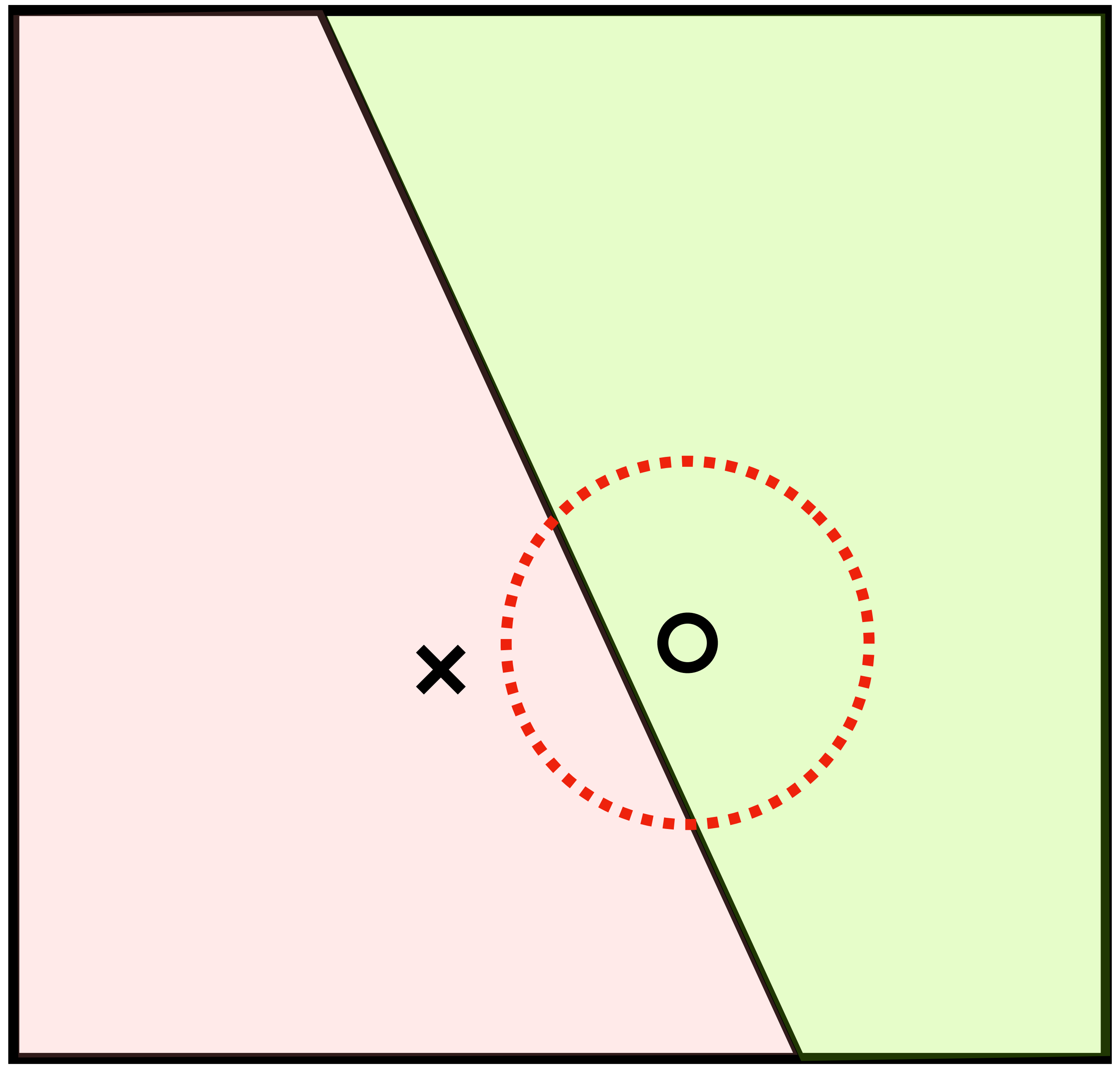


Towards Robust Contrastive Explanations for Human-Neural Multi-agent Systems. Leofante and Lomuscio, AAMAS 2023.
Robust Explanations for Human-Neural Multi-agent Systems with Formal Verification. Leofante and Lomuscio, EUMAS 2023.

Solutions

Leofante and Lomuscio use formal verification to identify robust CXs

- Given a CX x and model $\mathscr{M}$
- Check **local robustness** of $\mathscr{M}$ around x using verifiers

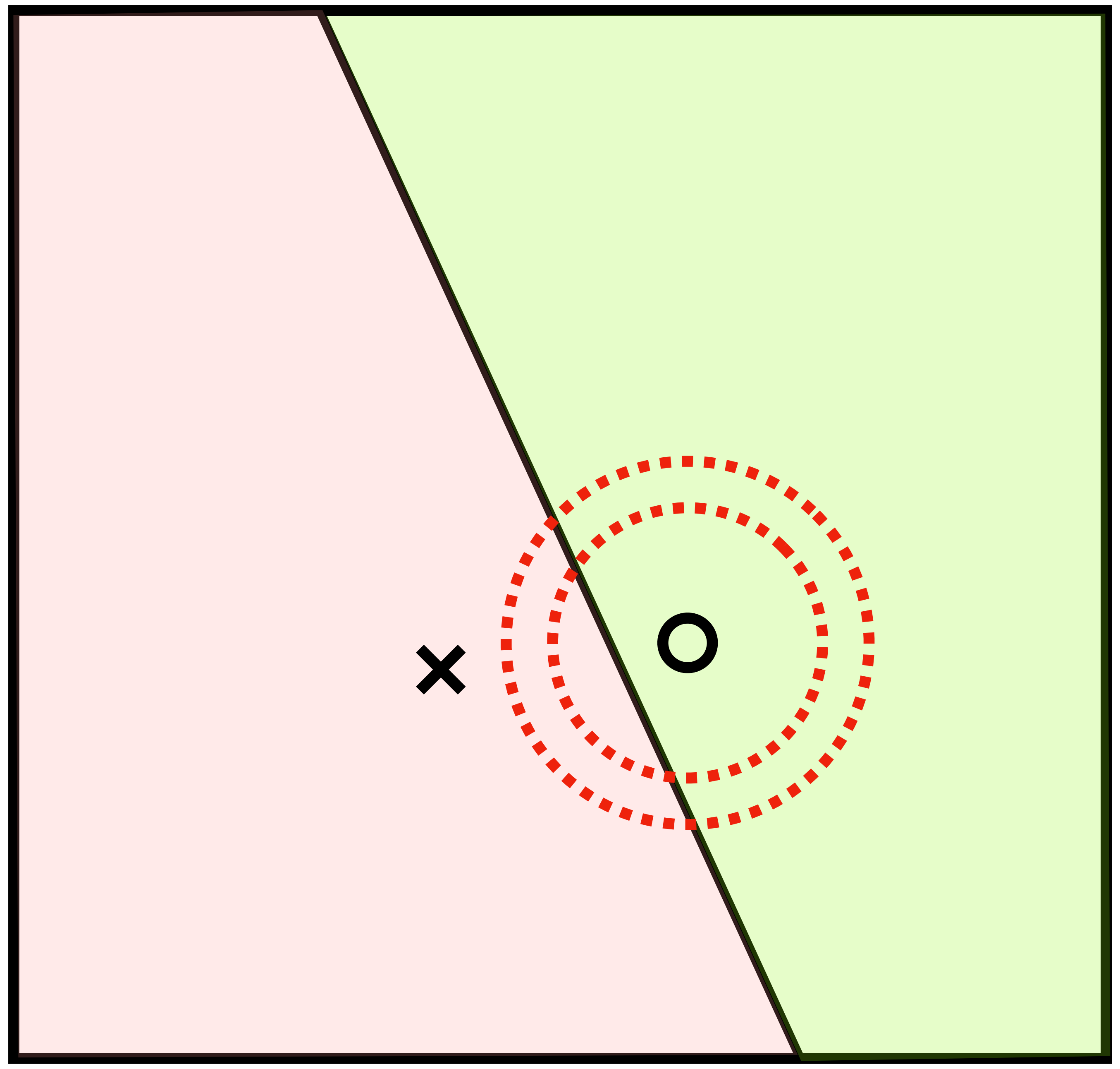


Towards Robust Contrastive Explanations for Human-Neural Multi-agent Systems. Leofante and Lomuscio, AAMAS 2023.
Robust Explanations for Human-Neural Multi-agent Systems with Formal Verification. Leofante and Lomuscio, EUMAS 2023.

Solutions

Leofante and Lomuscio use formal verification to identify robust CXs

- Given a CX x and model $\mathscr{M}$
- Check **local robustness** of $\mathscr{M}$ around x using verifiers

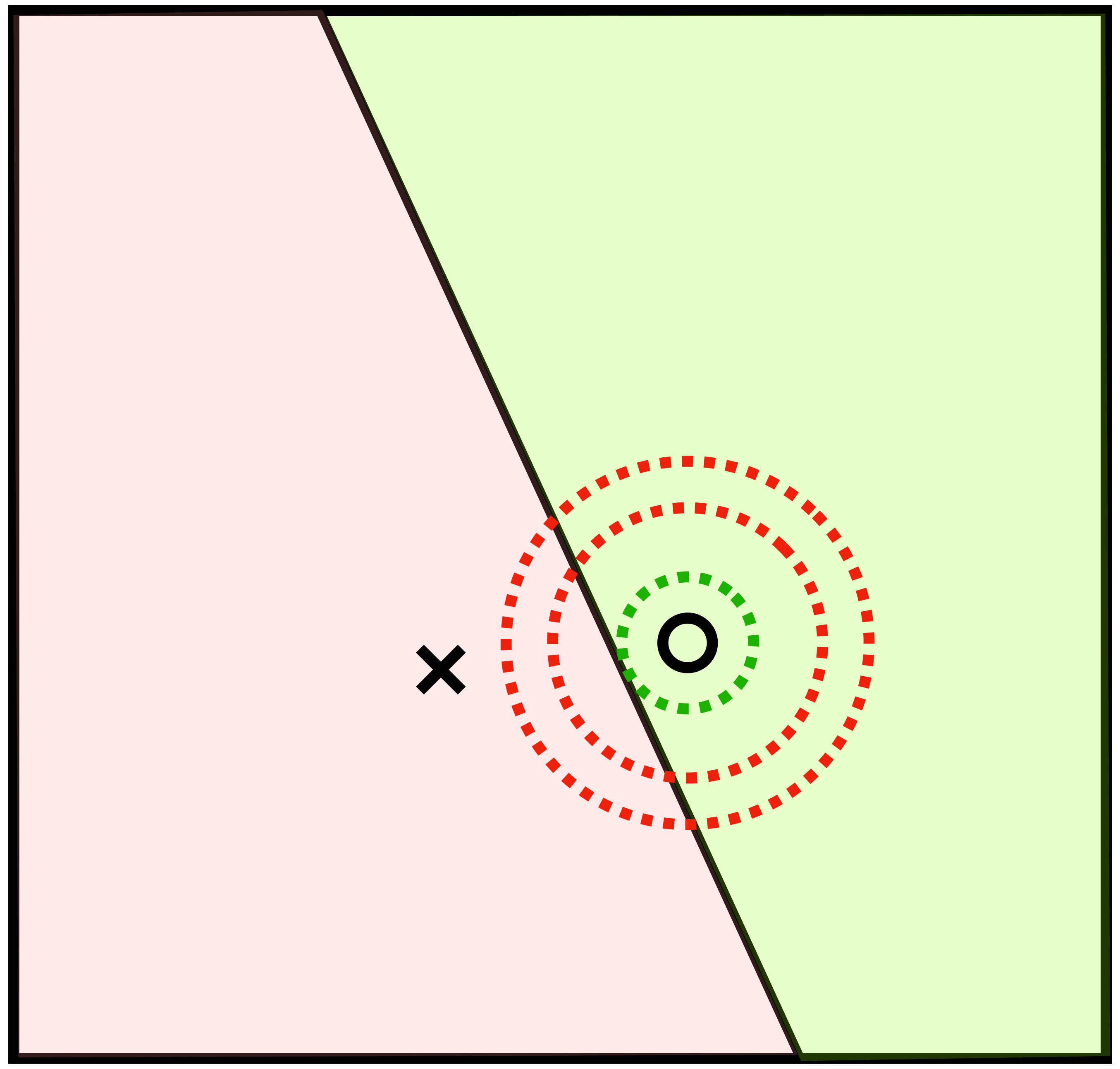


Towards Robust Contrastive Explanations for Human-Neural Multi-agent Systems. Leofante and Lomuscio, AAMAS 2023.
Robust Explanations for Human-Neural Multi-agent Systems with Formal Verification. Leofante and Lomuscio, EUMAS 2023.

Solutions

Leofante and Lomuscio use formal verification to identify robust CXs

- Given a CX x and model $\mathscr{M}$
- Check **local robustness** of $\mathscr{M}$ around x using verifiers
- CX **guaranteed to be robust** when safe radius identified



Towards Robust Contrastive Explanations for Human-Neural Multi-agent Systems. Leofante and Lomuscio, AAMAS 2023.
Robust Explanations for Human-Neural Multi-agent Systems with Formal Verification. Leofante and Lomuscio, EUMAS 2023.

Summing up

- CX generation methods focus on **minimising distance**
- This may result in **brittle explanations**
- We have examined **lack of robustness** in four scenarios:
    - input noise, model shifts, model multiplicity and noisy execution
- Can we borrow ideas from other areas of CS to fix this?

Some interesting (relevant) directions

Robustness and…

- Formal Explainable AI
- Fairness in ML
- Formal verification of neural networks
- Privacy
- Others?

Delivering Trustworthy AI through Formal XAI. Marques-Silva and Ignatiev, AAAI 2022.
Counterfactual Explanations Can Be Manipulated. Slack et al, NeurIPS 2021.
Algorithms for Verifying Deep Neural Networks. Liu et al, Found. Trends Optim. 4(3-4): 244-404, 2021.
On the Privacy Risks of Algorithmic Recourse. Pawelczyk et al, AISTATS 2023.

Thank you!

Contacts:

- f.leofante@imperial.ac.uk
- https://fraleo.github.io/

SCAN ME

References

- Counterfactual explanations without opening the black box: automated decisions and the GDPR. Wachter et al, Harvard Journal of Law & Technology 2018.
- Exploring Counterfactual Explanations Through the Lens of Adversarial Examples: A Theoretical and Empirical Analysis. Pawelczyk et al, AISTATS 2022.
- Robustness in Machine Learning Explanations: Does it Matter? Hancox-Li, FAT* 2020.
- Counterfactual Explanations Can Be Manipulated. Slack et al, NeurIPS 2021.
- On the Adversarial Robustness of Causal Algorithmic Recourse. Dominguez-Olmedo et al, ICML 2022.
- Density-based Realiable and Robust Explainer for Counterfactual Explanations. Zhang et al, Expert Systems with Applications, 2023.
- The Robustness of Counterfactual Explanations Over Time. Ferrario and Loi, IEEE Access, 2022.
- Towards Robust and Reliable Algorithmic Recourse. Upadhyay et al, NeurIPS, 2021.
- Formalising the Robustness of Counterfactual Explanations for Neural Networks. Jiang et al, AAAI 2023.
- Provably Robust and Plausible Counterfactual Explanations for Neural Networks via Robust Optimisation. Jiang et al, ACML 2023 (arxiv preprint at https://arxiv.org/abs/2309.12545)
- Model Multiplicity: Opportunities, Concerns, and Solutions. Black et al, ACM FAccT'22.
- On Counterfactual Explanations under Predictive Multiplicity. Pawelczyk et al, UAI 2020.
- Counterfactual Explanations and Model Multiplicity: a Relational Verification View. Leofante et al, KR, 2023.
- Manipulation-Proof Machine Learning. Björkegren et al, arxiv preprint https://arxiv.org/abs/2004.03865, 2020.
- Probabilistically Robust Recourse: Navigating the Trade-offs between Costs and Robustness in Algorithmic Recourse. Pawelczyk et al, ICLR 2023.
- Towards Robust Contrastive Explanations for Human-Neural Multi-agent Systems. Leofante and Lomuscio, AAMAS 2023.
- Robust Explanations for Human-Neural Multi-agent Systems with Formal Verification. Leofante and Lomuscio, EUMAS 2023.
- Delivering Trustworthy AI through Formal XAI. Marques-Silva and Ignatiev, AAAI 2022.
- Algorithms for Verifying Deep Neural Networks. Liu et al, Found. Trends Optim, 2021.
- On the Privacy Risks of Algorithmic Recourse. Pawelczyk et al, AISTATS 2023.