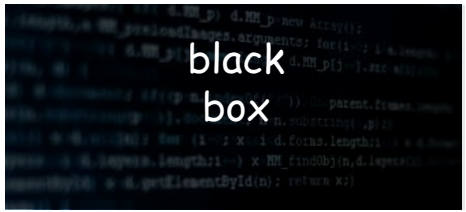
Causality and Explainability for Black Box software

Hana Chockler
causaLens
and
Department of Informatics
King's College, London

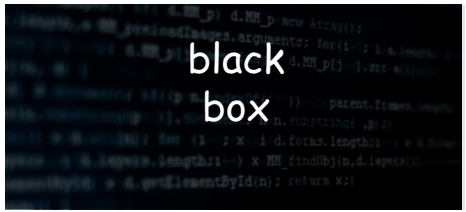
Motivation: Modern computerized systems are huge and difficult to understand



Motivation: Modern computerized systems are huge and difficult to understand

Can we understand and fix errors?

What does the system do?

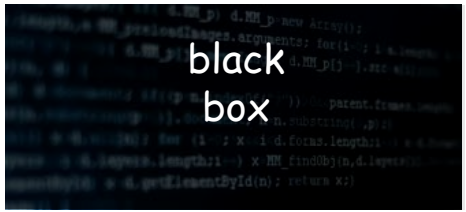


Can we be sure it is correct?

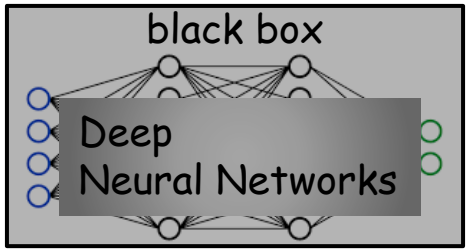
Motivation: Modern computerized systems are huge and difficult or even impossible to understand

Can we understand and fix errors?

What does the system do?



Can we be sure it is correct?



Actual Causality

A theoretical concept from AI

Extends causal counterfactual reasoning

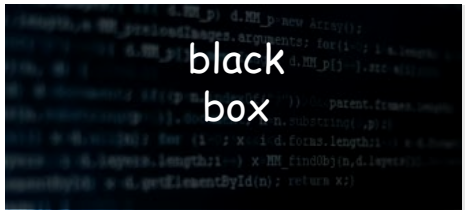
Turns out to be very useful!

Intractable – but there are efficient approximation algorithms and sufficient partial solutions

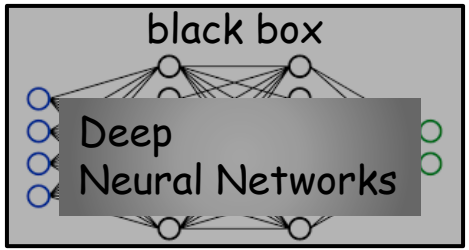
Motivation: Modern computerized systems are huge and difficult or even impossible to understand

Can we understand and fix errors?

What does the system do?



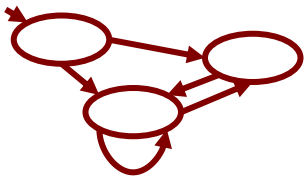
Can we be sure it is correct?



Formal Verification

Is the system correct?

A formal model of the system M:



A correctness specification φ

Does M satisfy φ?

no

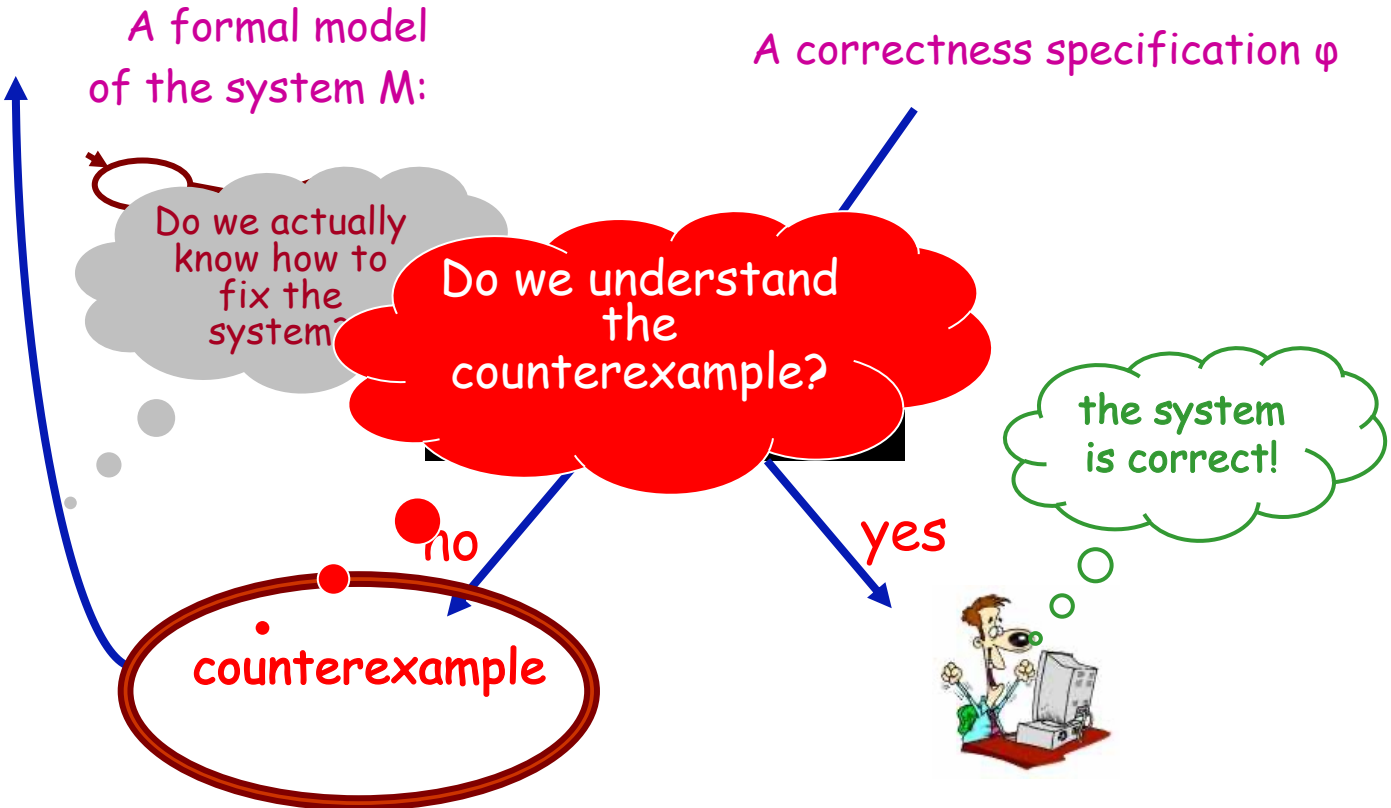
counterexample

yes

the system is correct!

Formal Verification

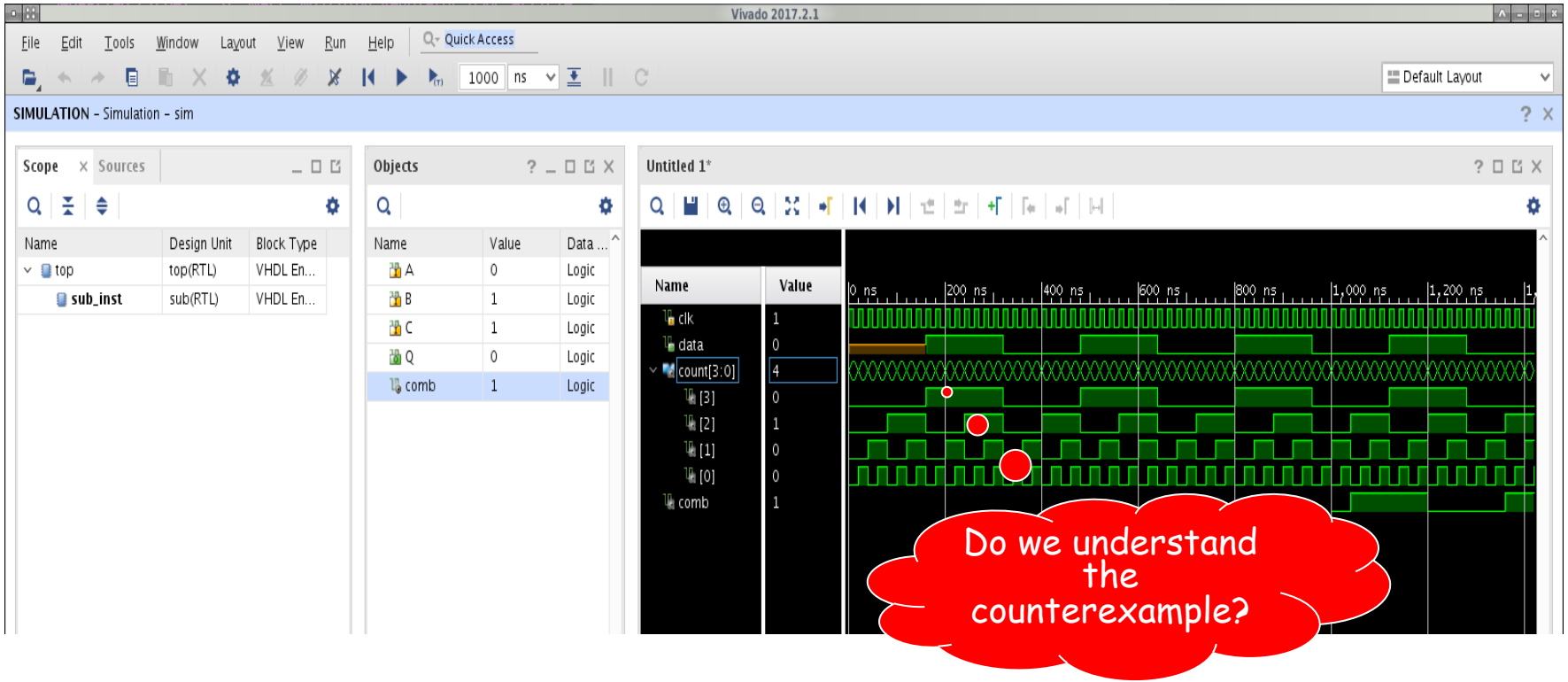
Is the system correct?



Counterexamples

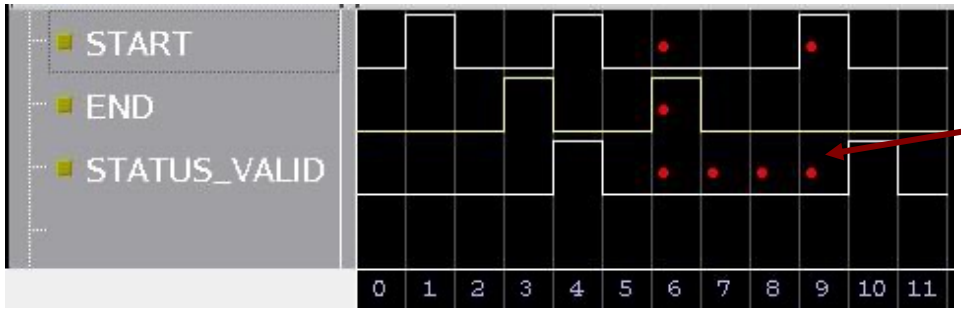
Counterexamples in hardware

A huge timing diagram that is very difficult to understand



Explaining counterexamples using causality (Red Dots) part of IBM tool

A timing diagram of a buggy hardware execution



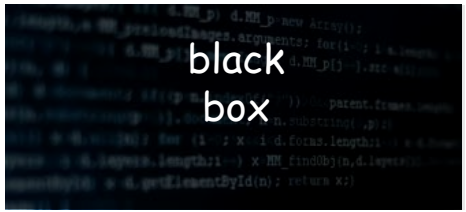
causes marked as red dots

φ = *always* ((!START *and* !STATUS_VALID *and* END (->
next(!START Until (STATUS_VALID *and* READY))

works and is really useful!

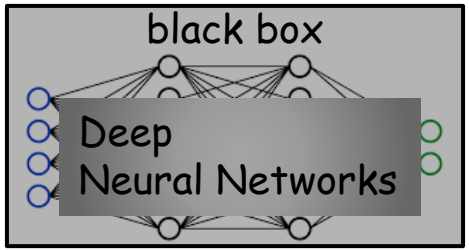
Motivation: Modern computerized systems are huge and difficult or even impossible to understand

Can we understand and fix errors?

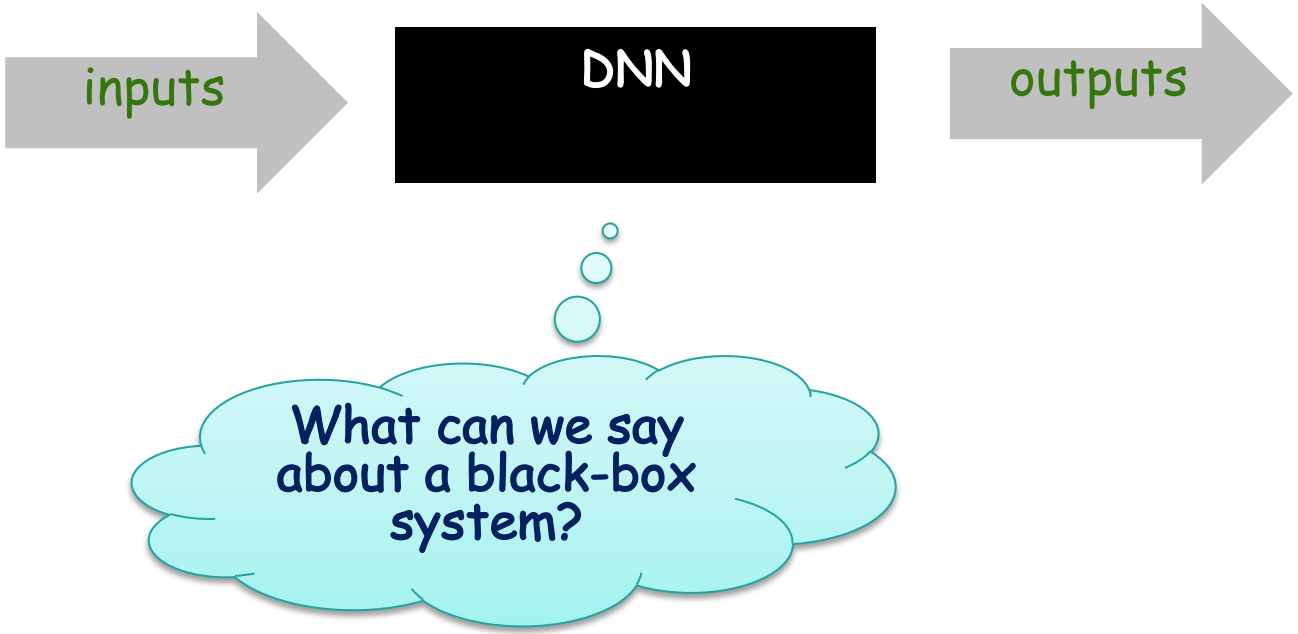


What does the system do?

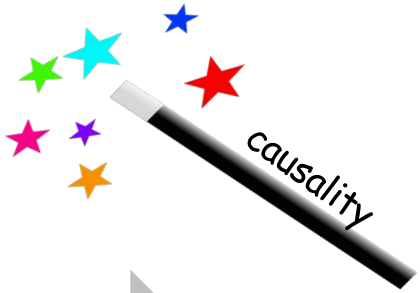
Can we be sure it is correct?

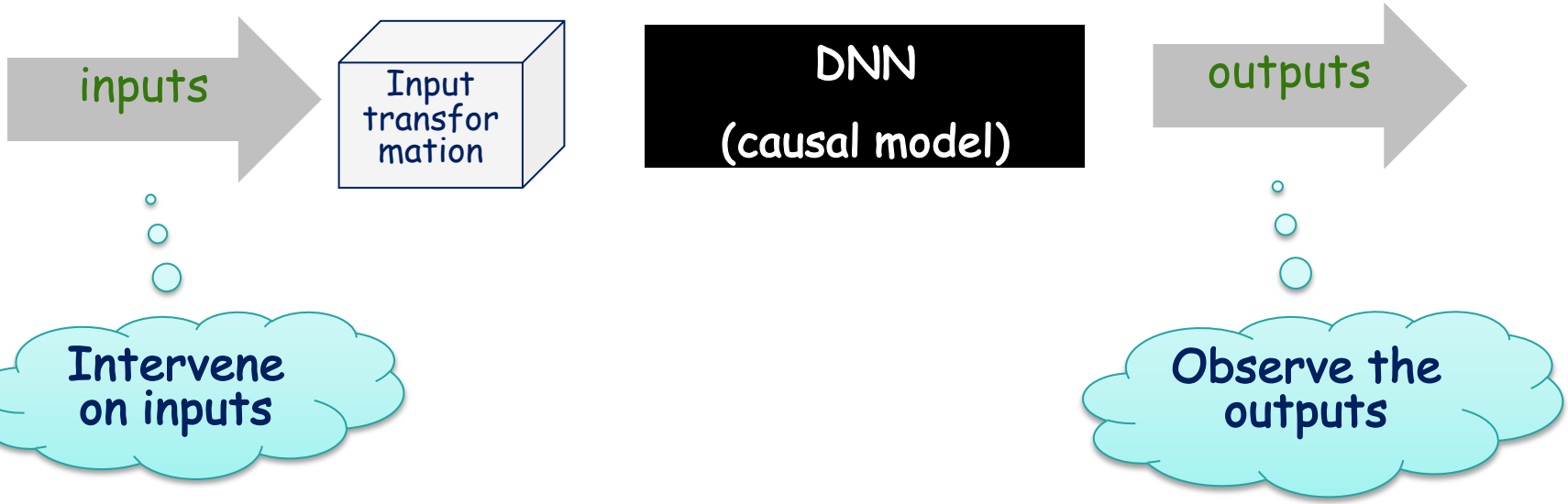


Reasoning about black-boxes



Reasoning about black-boxes





We can reason about various properties of the system without opening the black box

Explanations for Deep Neural Network's decisions

DNN for classifying animals

red panda

Explanation: minimal, sufficient, non-trivial subset of the pixels of the image

Because of this part:

Subtle misclassification – uncovered by explanations

DNN for classifying images

seems ok

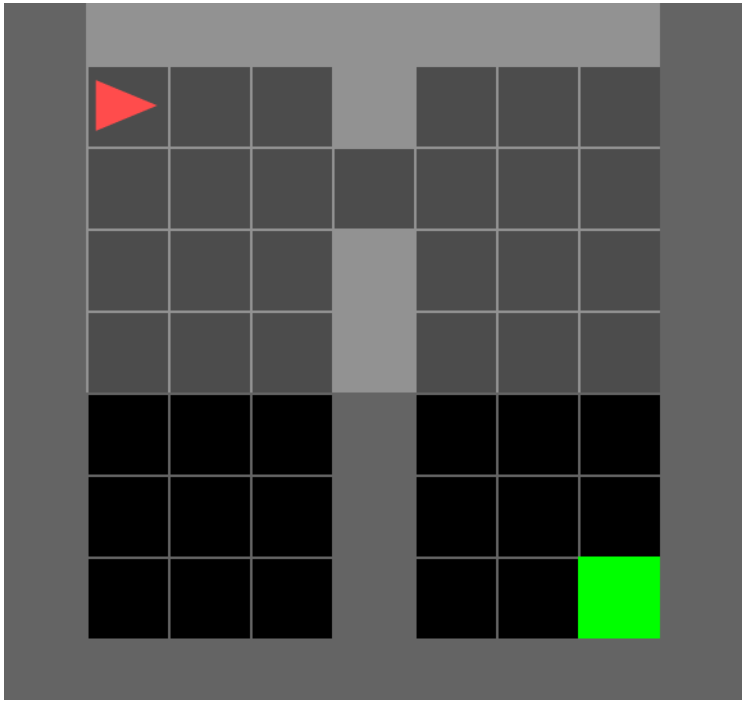
cowboy hat

Because of this part:

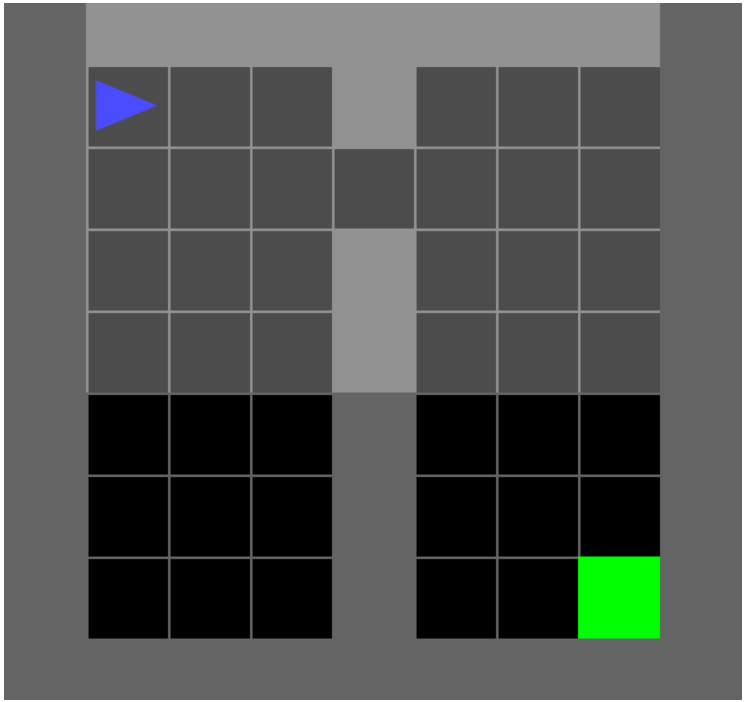
Explanation uncovered misclassification!

retrain

Reinforcement learning – causal simplification of policies

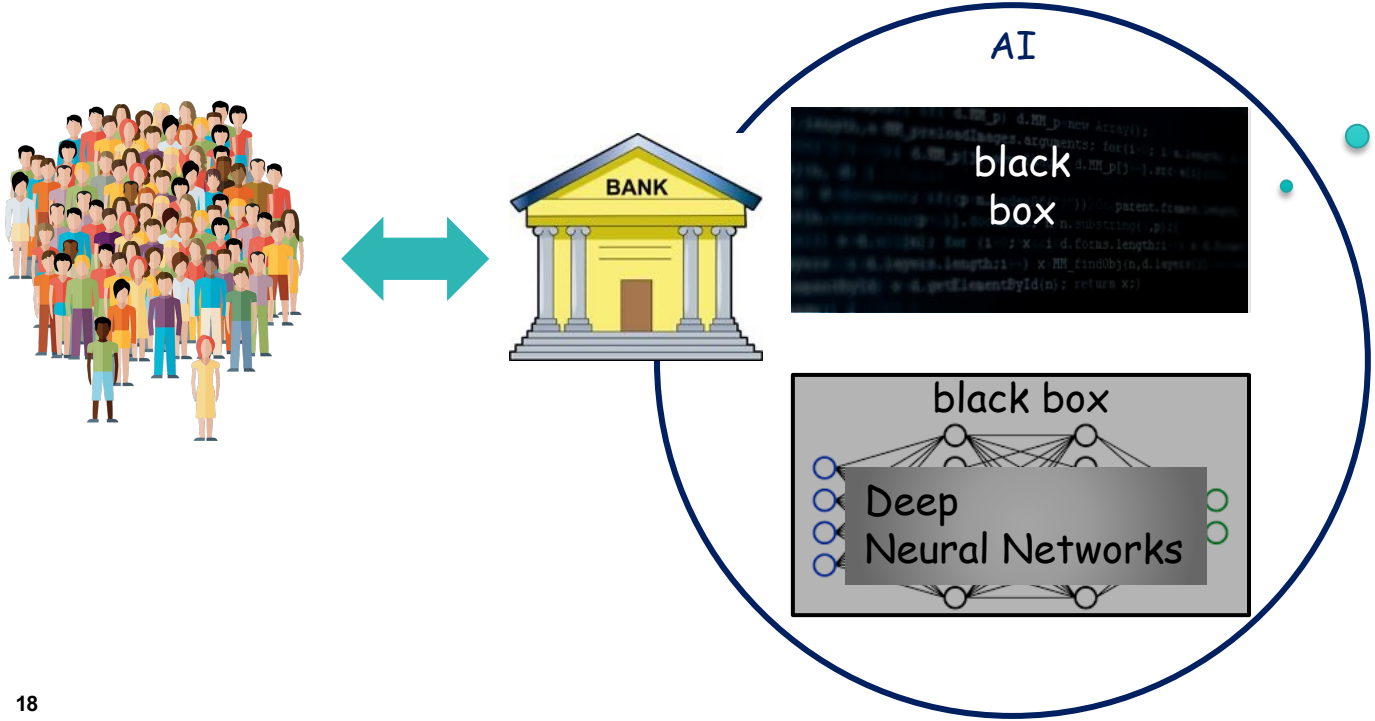


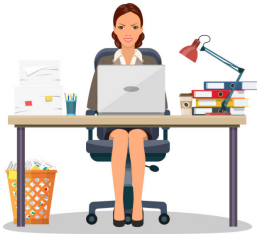
Original policy



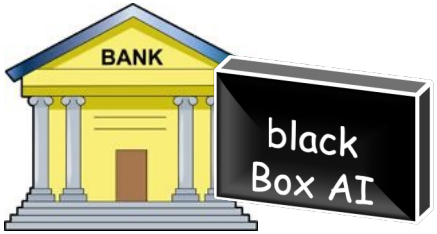
Simplified policy

AI black-box systems are widely used
Their decisions affect people





Proposed regulatory framework



Mortgage approval

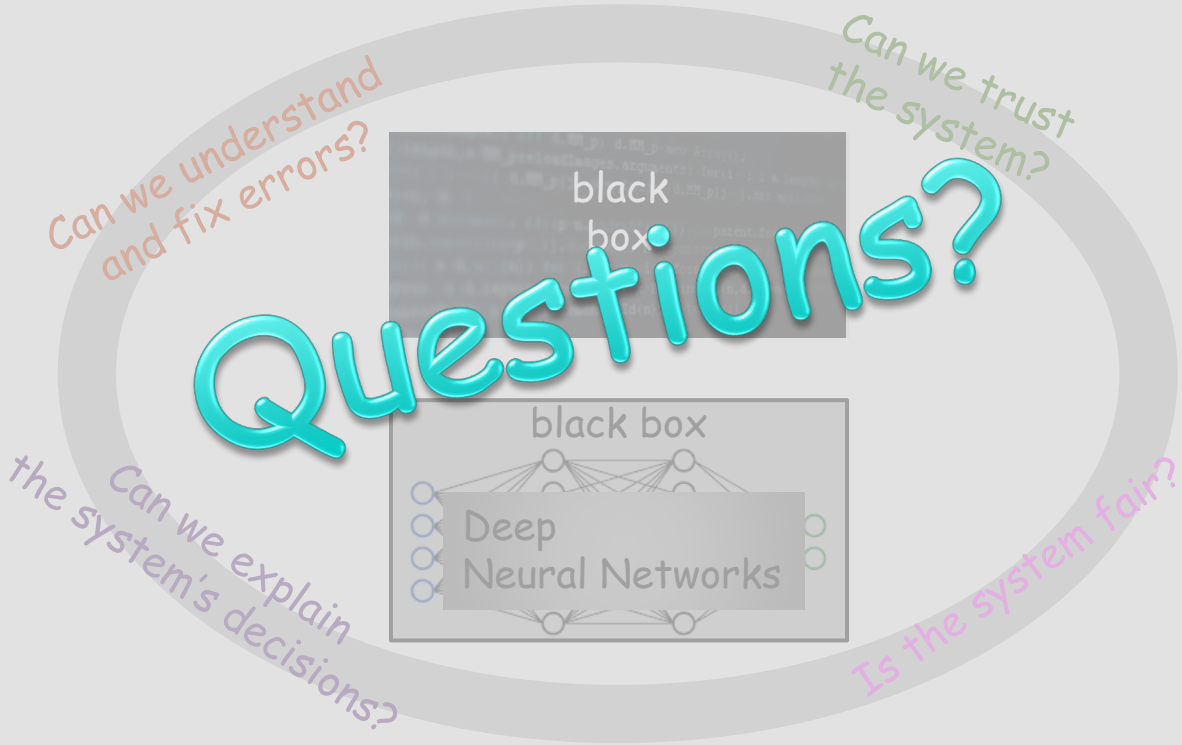
Fairness + no proxy variables

Precise algorithm or approximation

Statistical independence

Bibliography

- Chockler and Halpern. "Responsibility and Blame: A Structural-Model Approach". J. Artif. Intell. Res. 22: 93-115 (2004)
- Beer, Ben-David, Chockler, Orni, Trefler. "Explaining Counterexamples Using Causality". FMSD (2012)
- Aleksandrowicz, Chockler, Halpern, Ivrii. "The Computational Complexity of Structure-Based Causality". AAAI'14: 974-980.
- Alrajeh, Chockler, Halpern. "Combining Experts' Causal Judgments". Artif. Intell. (2020).
- Sun, Chockler, Huang, Daniel Kroening. "Explaining Image Classifiers Using Statistical Fault Localization". ECCV'20: 391-406.
- Chockler, Kroening, Sun. "Explanations for Occluded Images". ICCV'21: 1234-1243.
- Pouget, Chockler, Sun, Kroening. "Ranking Policy Decisions". NeurIPS'21.
- McNamee and Chockler: Causal policy ranking. OSC ICLR'2022.
- Chockler and Halpern. "On Testing for Discrimination Using Causal Models". AAAI'22.


Can we understand and fix errors?
Can we trust the system?
black box
Questions?
black box
Deep Neural Networks
Can we explain the system's decisions?
Is the system fair?