ACU
AUSTRALIAN CATHOLIC UNIVERSITY

Explainable AI: Challenges and Opportunities

Sam Baron

Dianoia Institute of Philosophy

Roadmap

1. Introduce AI
2. Overview of Explainable AI (XAI)
3. Challenges for XAI
4. Example: Causation
5. Concluding Thoughts

ACU
AUSTRALIAN CATHOLIC UNIVERSITY

Artificial Intelligence

Two Kinds of AI

General Artificial Intelligence

- Continuous with human level intelligence
- E.g., Skynet

Algorithmic Artificial Intelligence

- Algorithms developed for various purposes, in various ways.
- E.g., takes in pictures and tells you which ones are images of dogs.

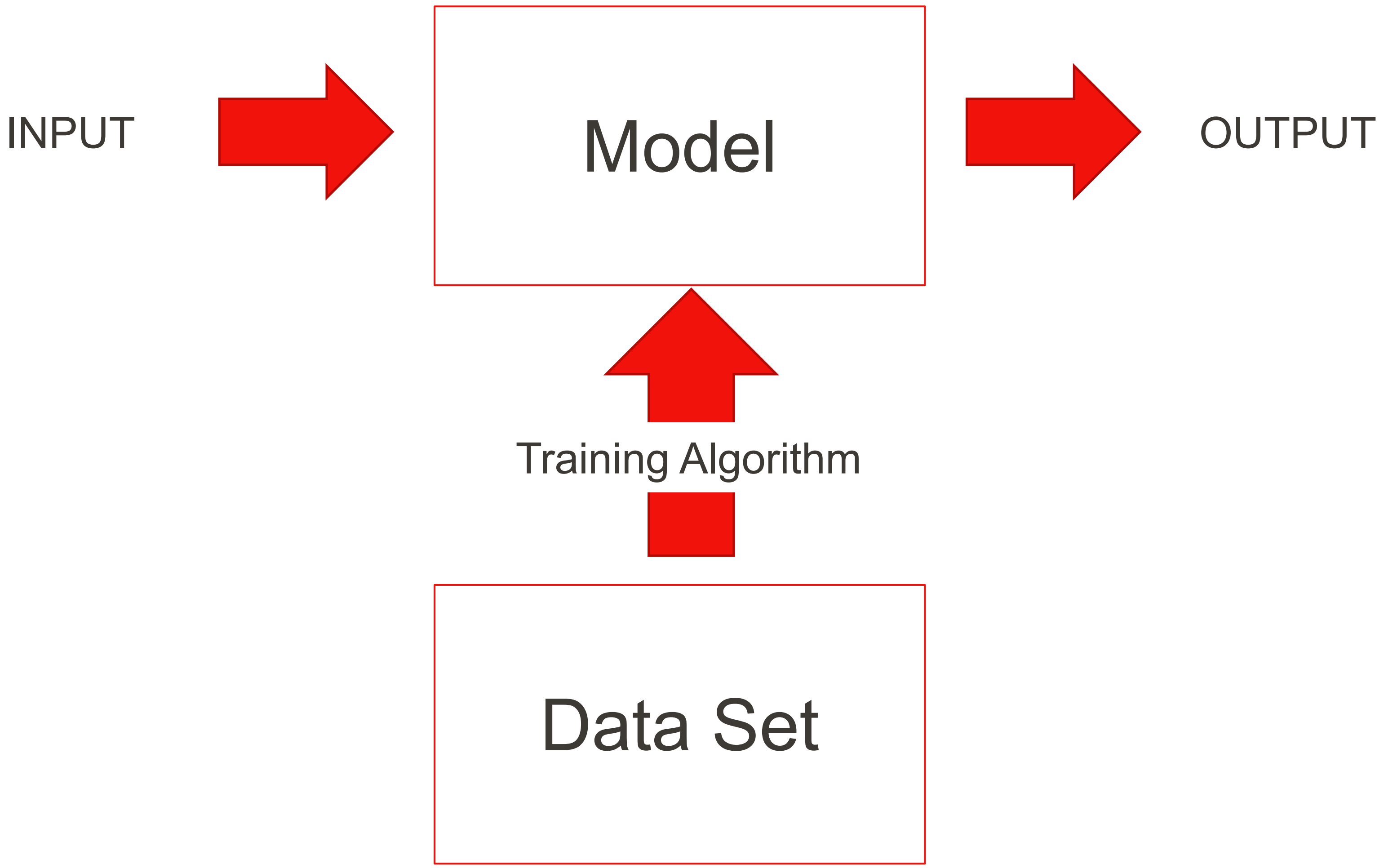
Two Kinds of AI

Where’s the Danger?

- Many are worried about general AI
- Worries about potential existential risk
- But algorithmic AI exists and is harming people *right now*
- E.g., loan decisions, criminal recidivism, medical diagnosis, social media, advertising, streaming services
- The study of algorithmic AI in the humanities is lagging

Explainable AI (XAI)

Explanation and Artificial Intelligence



INPUT

Model

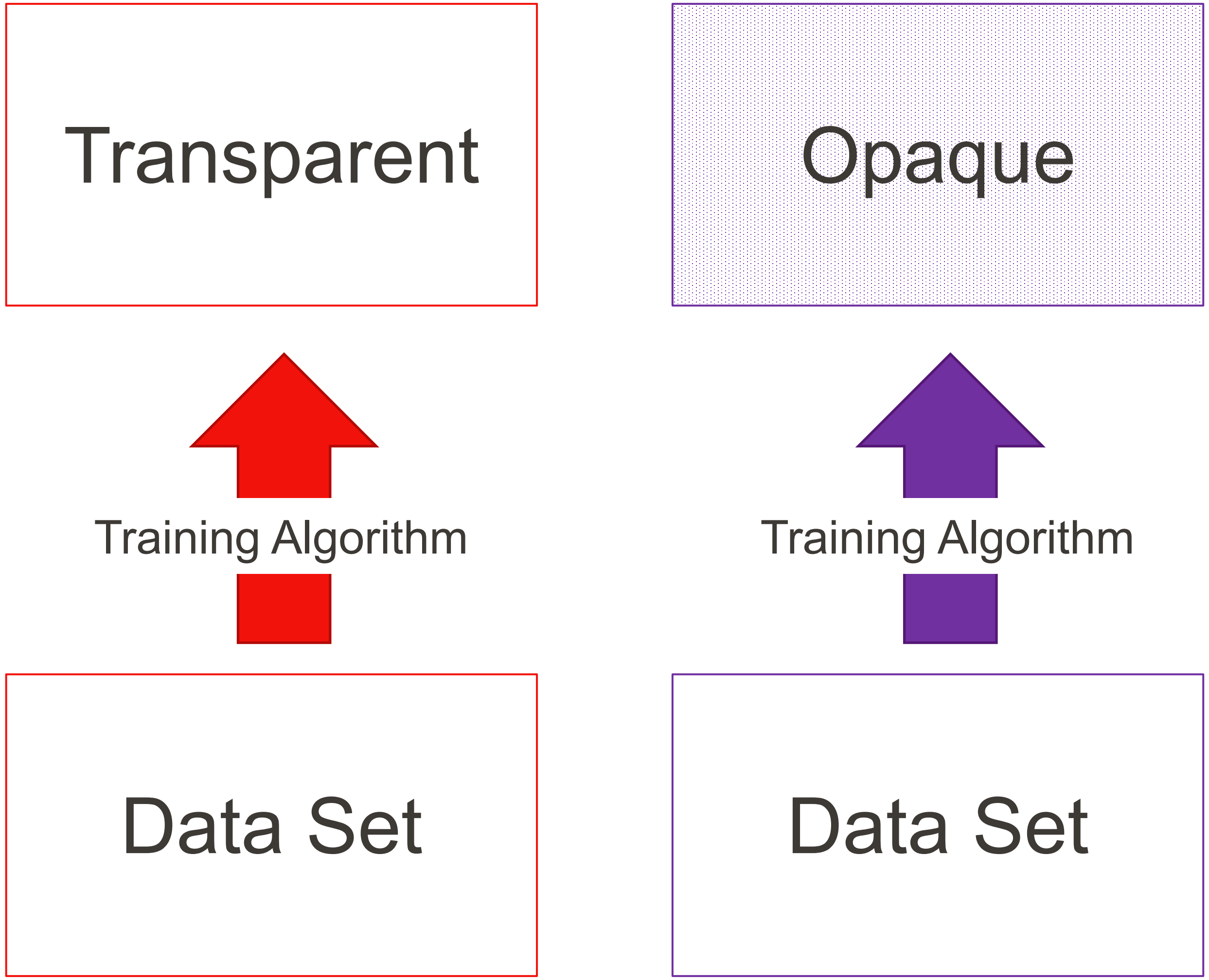
OUTPUT

Training Algorithm

Data Set

Explanation and Artificial Intelligence

Understand how specific inputs yield specific outputs



Transparent

Training Algorithm

Data Set

Opaque

Training Algorithm

Data Set

No understanding of how specific inputs yield specific outputs

Explanation and Artificial Intelligence

Demand for Explanation

- We are on the receiving end of AI decision-making
- We have a right to understand why a decision was reached
- Continuous with similar rights in e.g. legal contexts
- Right is being recognised in e.g. European Data Regulation
- XAI is the project of providing such explanations

XAI Challenges

Challenges for Explainable Artificial Intelligence

Three Goals for XAI

1. Help individuals understand why particular decisions were reached.
2. Equip individuals with the capacity to contest adverse decisions.
3. Equip individuals with the capacity to get better outcomes.

Explanation and Artificial Intelligence

1. **Transparency: in principle, AI systems must be explainable;**
2. **Inclusion: the needs of all human beings must be taken into consideration so that everyone can benefit and all individuals can be offered the best possible conditions to express themselves and develop**
3. Responsibility: those who design and deploy the use of AI must proceed with responsibility and **transparency**
4. Impartiality: do not create or act according to bias, thus safeguarding fairness and human dignity
5. Reliability: AI systems must be able to work reliably
6. Security and privacy: AI systems must work securely and respect the privacy of users.

Vatican statement on AI Ethics, 2020

Challenges for Explainable Artificial Intelligence

The Demand for Explanation

Example: Sara applies for a loan

1. Why did the model deliver a low credit score for Sara?
2. Can Sara challenge the bank’s decision?
3. What could Sara do next time to get a better outcome?

Genuine explanations of how the system works should provide answers to these questions.

Challenges for Explainable Artificial Intelligence

Spurious and Genuine Explanation

Why did Alex’s house burn down?

1. Because there was an electrical fault.
2. Because aliens shot it with a laser.
3. Because he set fire to it.

Challenges for Explainable Artificial Intelligence

Spurious and Genuine Explanation

1. What is a genuine explanation of an AI system?
2. How do we provide genuine explanations of AI systems?

We currently lack good answers to these questions.

So it is unclear how to make progress.

ACU
AUSTRALIAN CATHOLIC UNIVERSITY

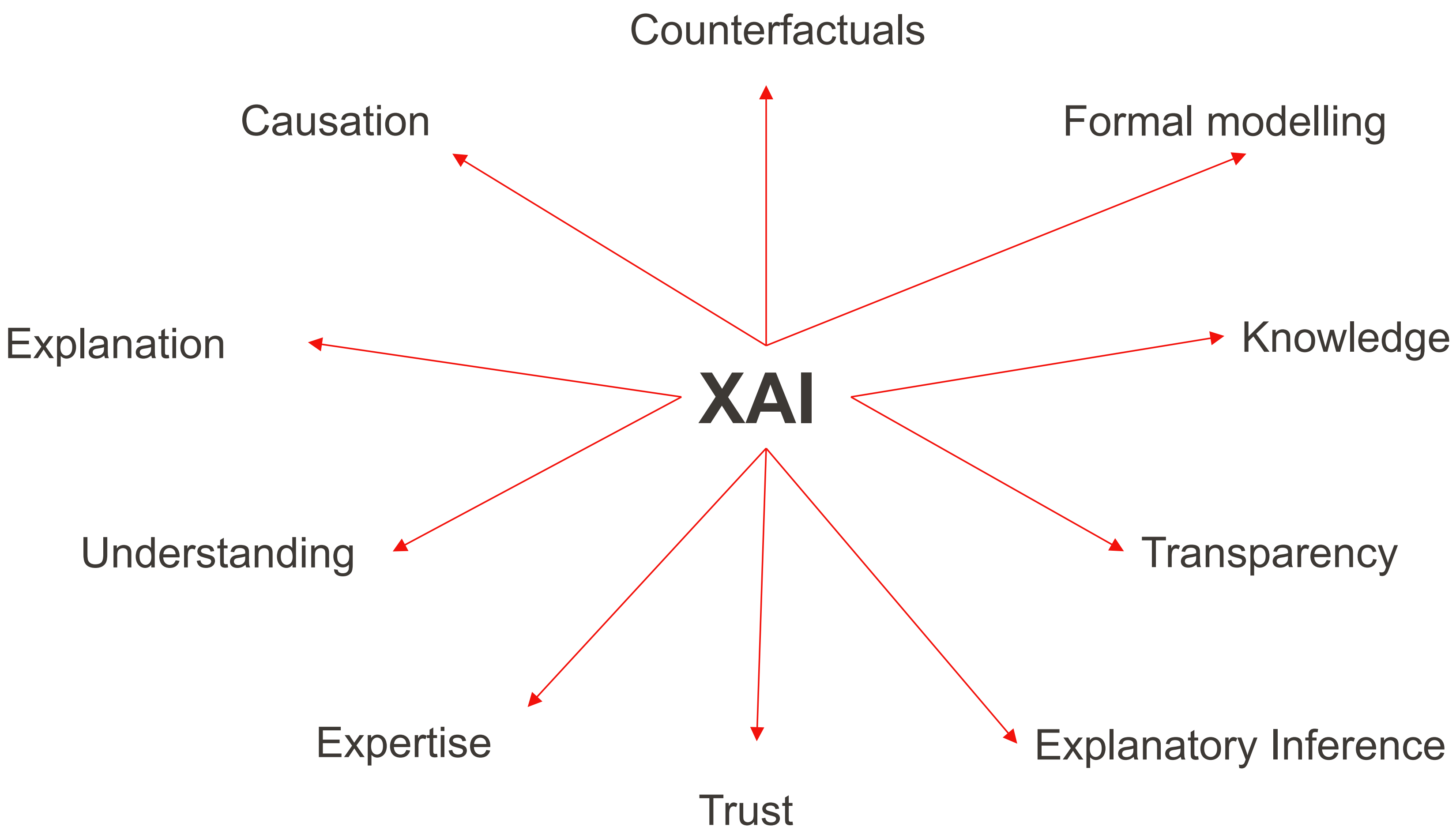
Interdisciplinary Opportunities

Philosophy and XAI

The Search for Explanation

- Explanation has been a central theme in the philosophy of science for a century
- Philosophers have produced a number of theories of explanation and how it works
- Philosophers have also developed sophisticated techniques for finding explanations
- Philosophy can *help*.

Philosophy and XAI



XAI

Counterfactuals

Causation

Formal modelling

Explanation

Knowledge

Understanding

Transparency

Expertise

Explanatory Inference

Trust

Causal Explanation

Example: Parole Decision

Alex is up for parole. Information about Alex including criminal history, sentencing, offenses, age, race, gender, time served are fed into a machine learning model. The model predicts that Alex's probability of re-offense is high. Alex is denied parole.

Causal Explanation

Example: Parole Decision

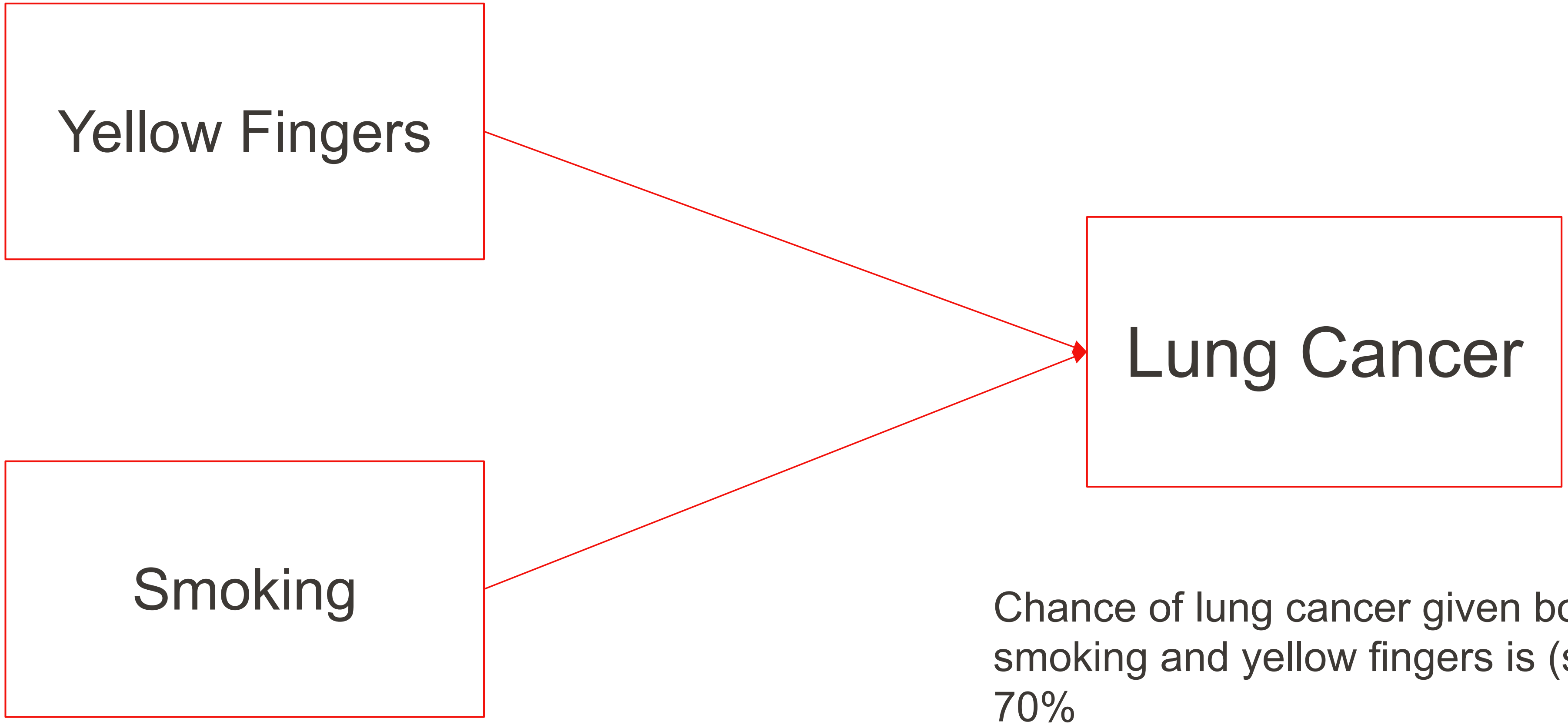
- If information about Alex’s race caused the system to yield its output, Alex can contest
- To help Alex, we need information about *which* inputs caused the system to yield a specific output in his case
- Finding this out is not straightforward
- We can find out which inputs are correlated with his output
- But that’s not enough!

Causal Explanation

Example: Parole Decision

- A method is needed to differentiate causation from mere correlation
- Philosophers have developed a method
- The basic idea: ‘wiggle’ each input one-by-one while holding the others fixed
- If the output wiggles too, then a variable is causal; if not then not.

Causal Explanation



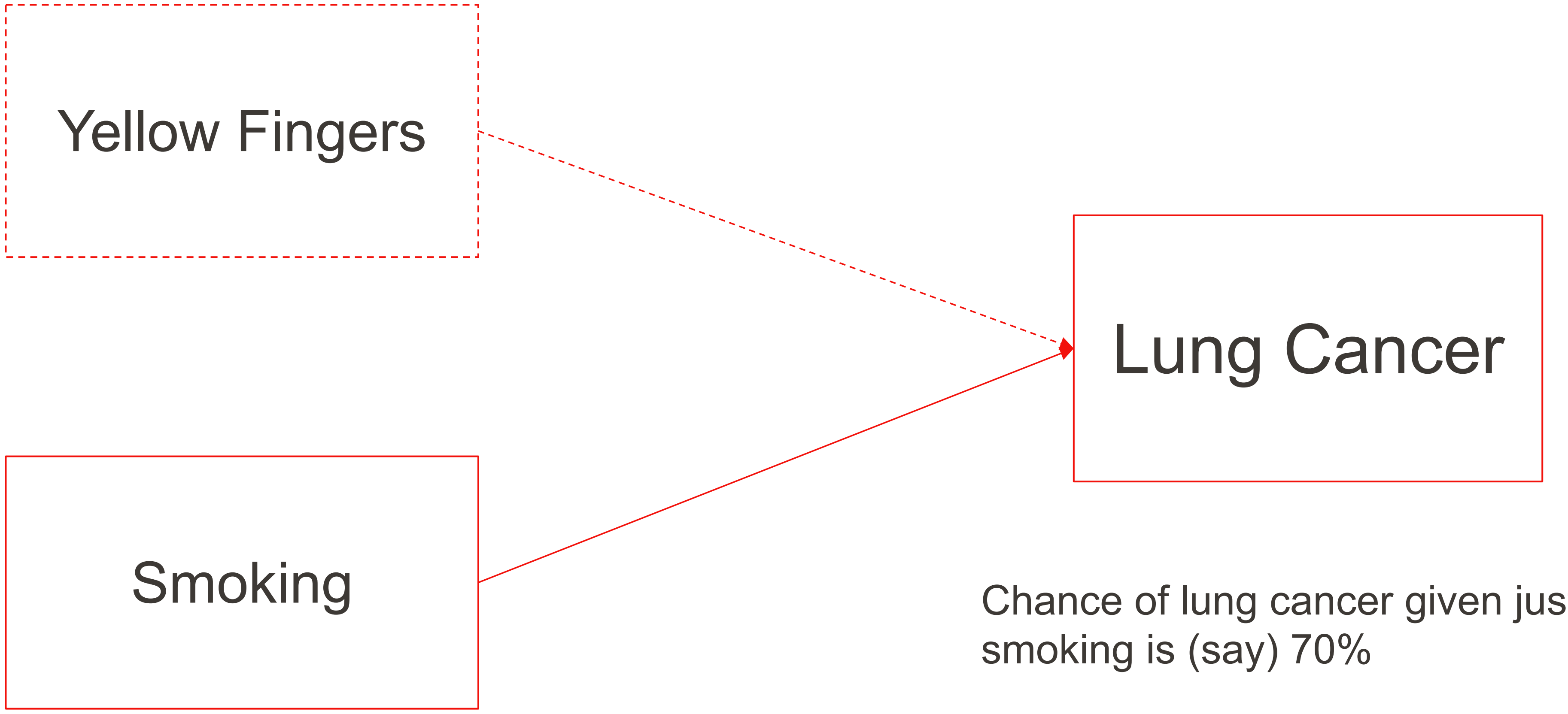
Yellow Fingers

Smoking

Lung Cancer

Chance of lung cancer given both smoking and yellow fingers is (say) 70%

Causal Explanation



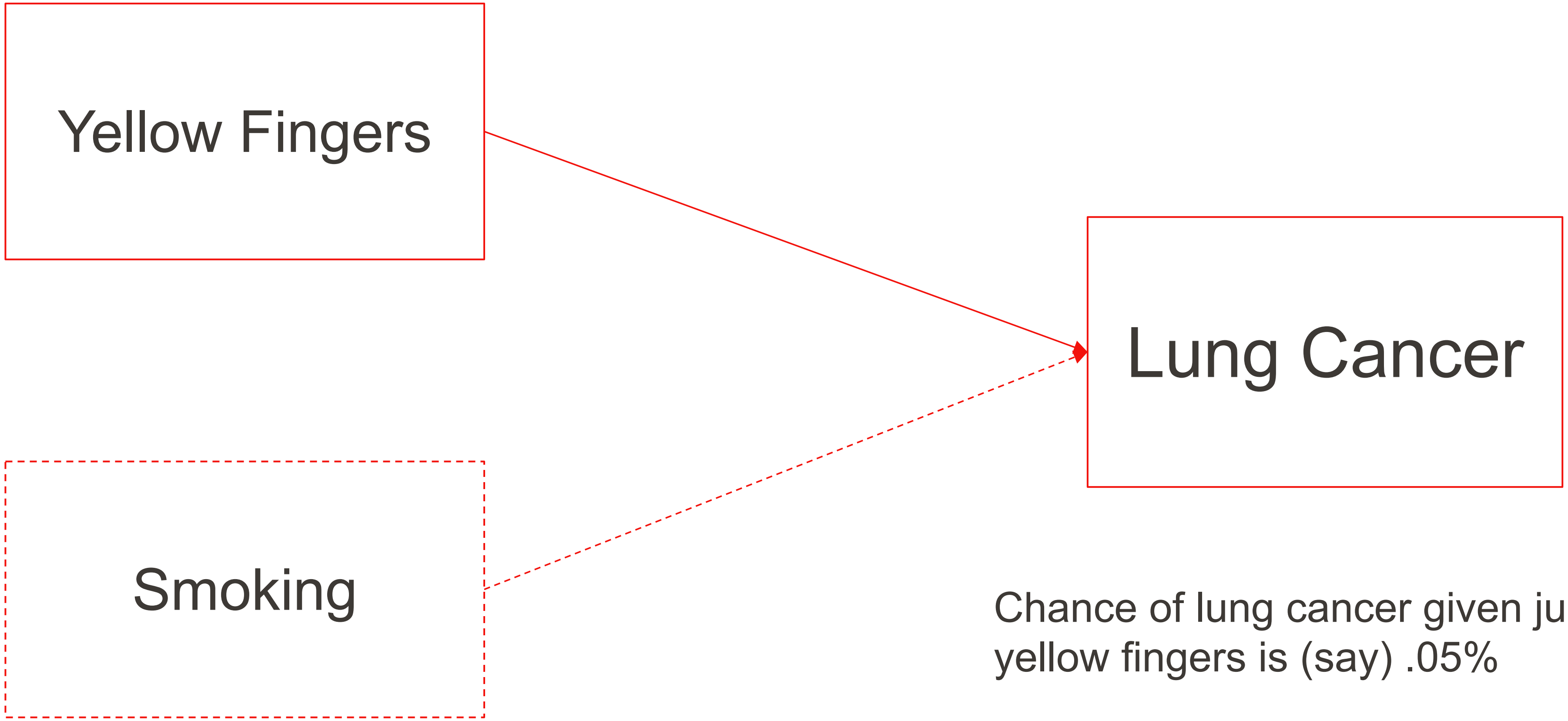
Yellow Fingers

Lung Cancer

Smoking

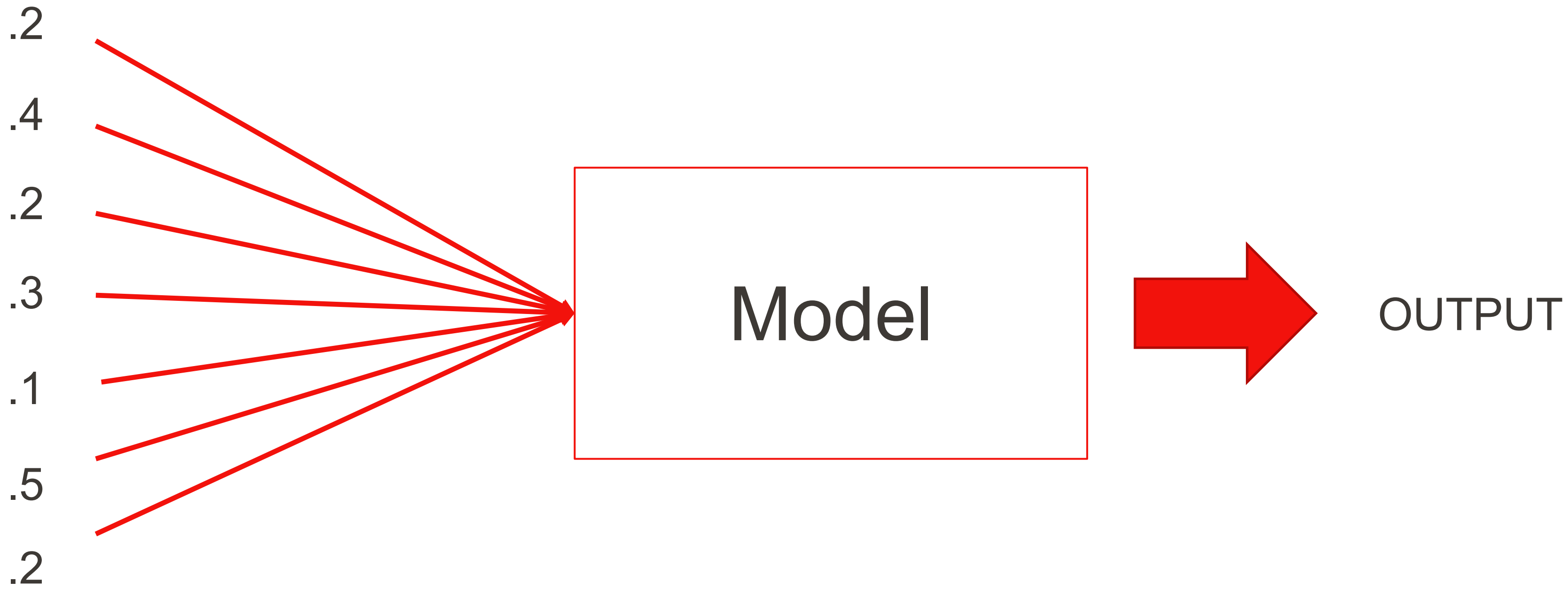
Chance of lung cancer given just smoking is (say) 70%

Causal Explanation



Chance of lung cancer given just yellow fingers is (say) .05%

Causal Explanation



Which correlations are causal?

Causal Explanation

Causal Certification

1. Guarantee that each explanation of the form ‘output P was caused by input Q’ is true.
2. Guarantee that individuals have been given complete information about causes.

Causal certification is needed for the goals of XAI.

AUSTRALIAN CATHOLIC UNIVERSITY

Concluding Thoughts

AUSTRALIAN CATHOLIC UNIVERSITY

Concluding Thoughts

- XAI rests on notions of explanation, understanding, causation
- Philosophy is well placed to develop these notions
- Philosophy and computer science must work together to ensure rights to explanation are respected
- This is necessary to minimise the harms of algorithmic AI
- We need to work out how to develop this partnership.