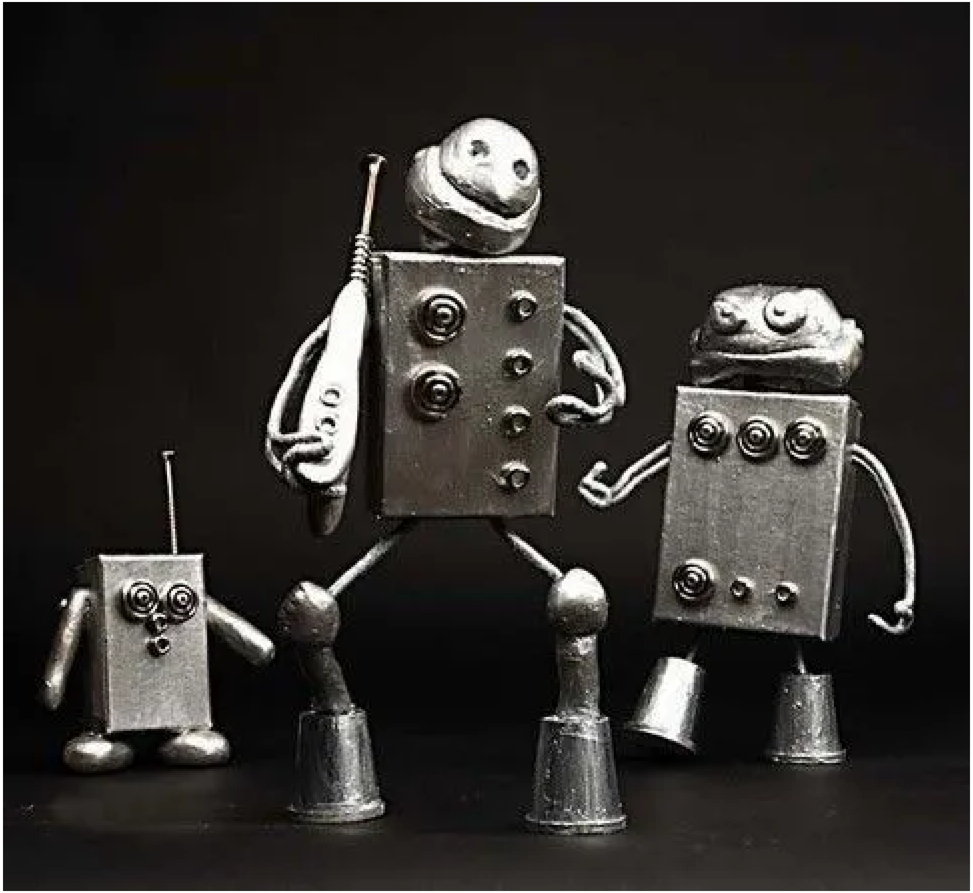
XAI from the perspective of Evidential Pluralism

Jon Williamson
Department of Philosophy & Centre for Reasoning, University of Kent

Causal XAI
Queen Mary, 26 October 2023

Contents

1 The need for explainable AI



Key questions

Will the AI work?

- Will it work in general? (developer)
- Will it work for me? (client affected by the AI)

Will the AI work fairly?

- Will it treat groups fairly? (regulator)
- Will it treat me fairly? (client)

XAI: how do we explain whether the AI will work / work fairly?

Key questions

Will the AI work? **I.e., is it effective? A causal question.**

Will it work in general? (developer, owner) **Generic causation.**

Will it work for me? (client affected by the AI) **Single-case causation.**

Will the AI work fairly? **Need to understand how it works.**

Will it treat groups fairly? (regulator) **Generic fairness.**

Will it treat me fairly? (client) **Single-case fairness.**

XAI: how do we explain whether the AI will work / work fairly?

I’ll argue that Evidential Pluralism can help us answer all these questions.

2 Evidential Pluralism

Correlation is not Causation

A correlation between A and B might be due to:

Causation. A is a cause of B.

Other causal explanations. Reverse causation, confounding, performance bias, detection bias, ...

Statistical explanations. Chance, fishing, temporal trends.

Non-causal connections. Semantic, constitutive, mereological, logical, nomological or mathematical relationships between A and B.

If A is a cause of B, then there is some complex of mechanisms that:

Explains instances of B by invoking instances of A, and

Can account for the magnitude of the observed correlation.

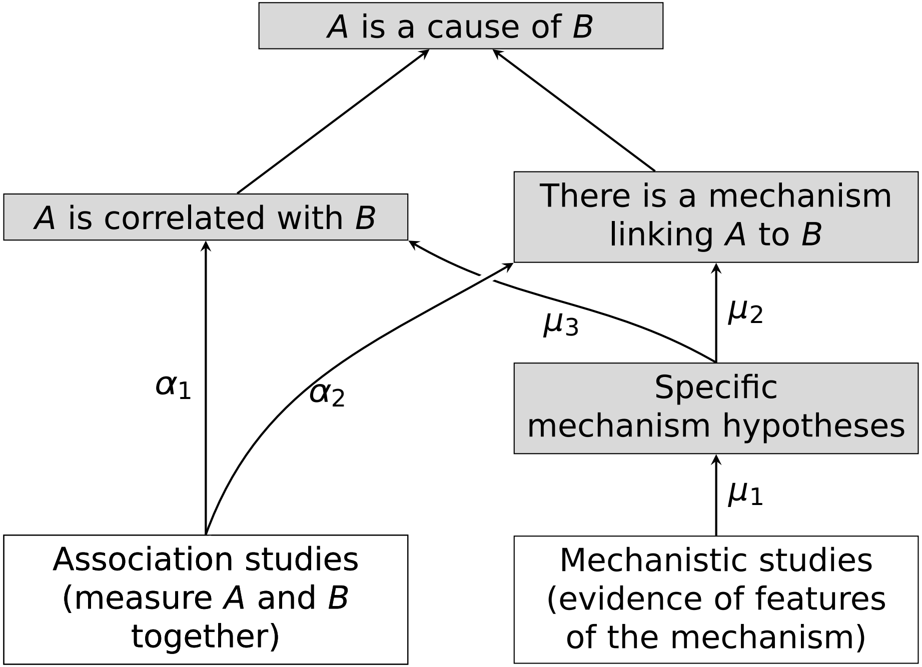
So, in order to establish causation one needs to establish both:

Correlation. The existence of an appropriate correlation.

Mechanism. The existence of an appropriate mechanism that can explain that correlation.

(Russo, F. and Williamson, J. (2007). Interpreting causality in the health sciences. *International Studies in the Philosophy of Science*, 21(2):157–170.)

This observation motivates **Evidential Pluralism**, a theory of causal enquiry:

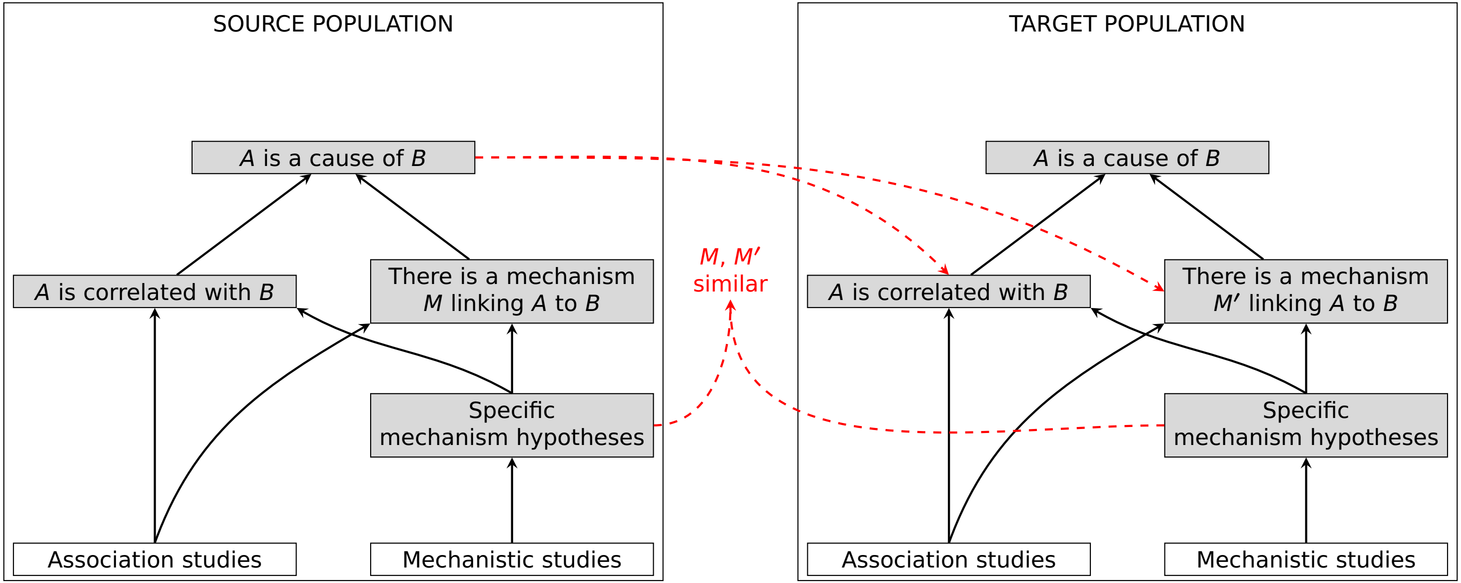

A is a cause of *B*
A is correlated with *B*
There is a mechanism linking *A* to *B*
α_1
α_2
μ_3
μ_2
Specific mechanism hypotheses
μ_1
Association studies (measure *A* and *B* together)
Mechanistic studies (evidence of features of the mechanism)

Object Pluralism. In order to establish a causal claim one normally needs to establish the existence of an appropriate conditional correlation and the existence of an appropriate mechanism complex.

Study Pluralism. So, when assessing a causal claim one ought to consider relevant association studies and mechanistic studies, where available.

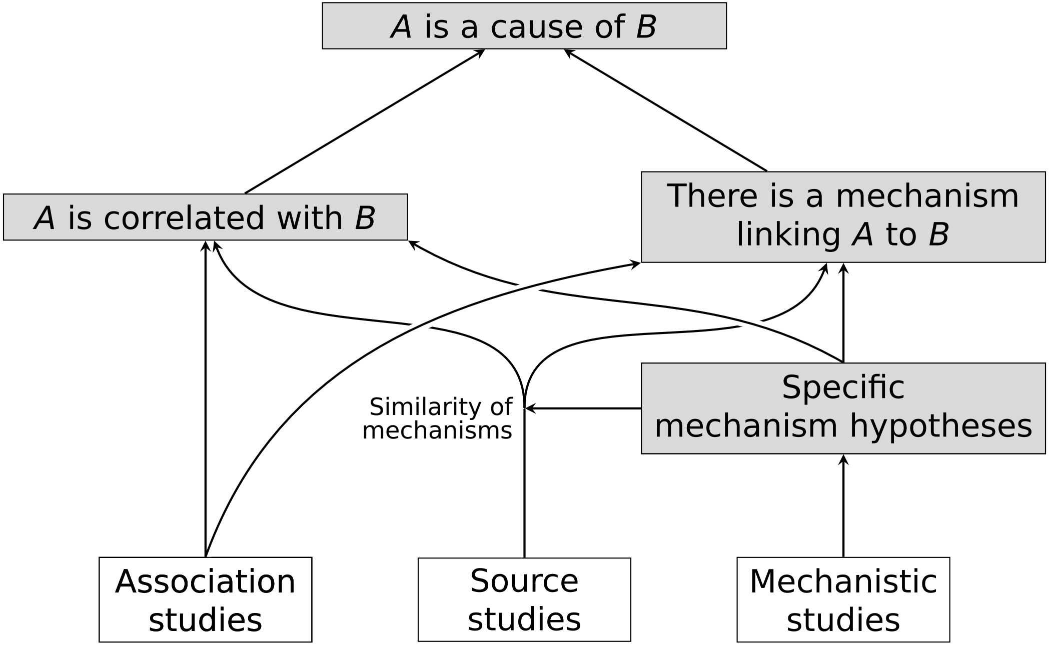
External validity

Evidential Pluralism also offers an account of extrapolating a causal claim to a new context:


SOURCE POPULATION
A is a cause of *B*
A is correlated with *B*
There is a mechanism *M* linking *A* to *B*
Specific mechanism hypotheses
Association studies
Mechanistic studies
M, *M′* similar
TARGET POPULATION
A is a cause of *B*
A is correlated with *B*
There is a mechanism *M′* linking *A* to *B*
Specific mechanism hypotheses
Association studies
Mechanistic studies

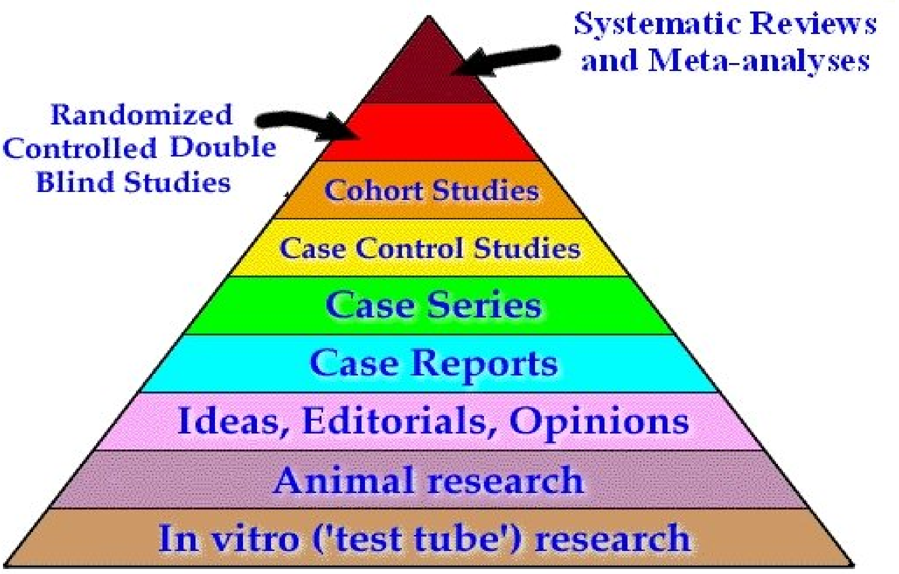
(Williamson, J. (2019). Establishing causal claims in medicine. *International Studies in the Philosophy of Science*, 32(2):33–61.)

Bringing efficacy and external validity together:

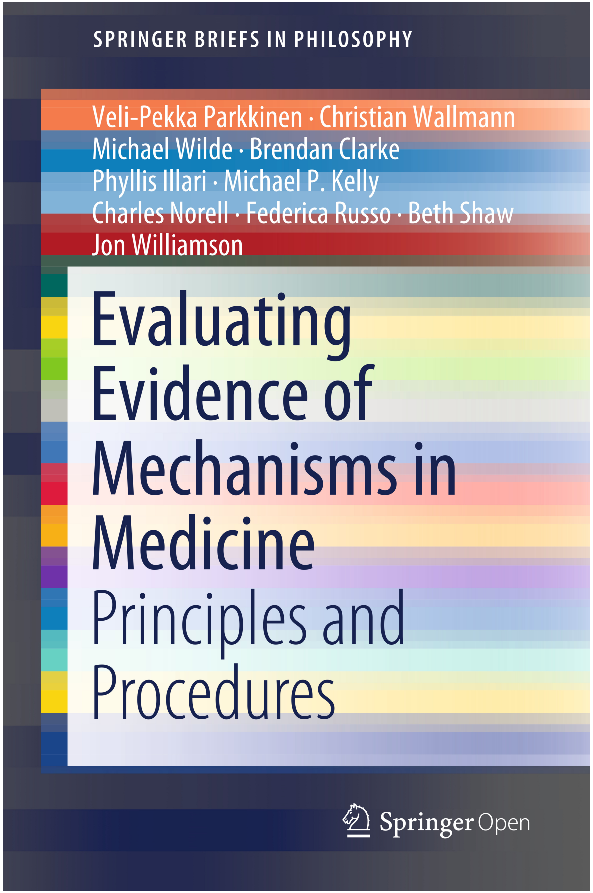


EP contrasts with the standard approach to EBM & EBP:

> Evidence-based medicine de-emphasizes intuition, unsystematic clinical experience, and pathophysiologic rationale as sufficient grounds for clinical decision making and stresses the examination of evidence from clinical research. (Guyatt et al., 1992, p. 2420)


Systematic Reviews and Meta-analyses
Randomized Controlled Double Blind Studies
Cohort Studies
Case Control Studies
Case Series
Case Reports
Ideas, Editorials, Opinions
Animal research
In vitro ('test tube') research

Evidential Pluralism motivates EBM+



SPRINGER BRIEFS IN PHILOSOPHY

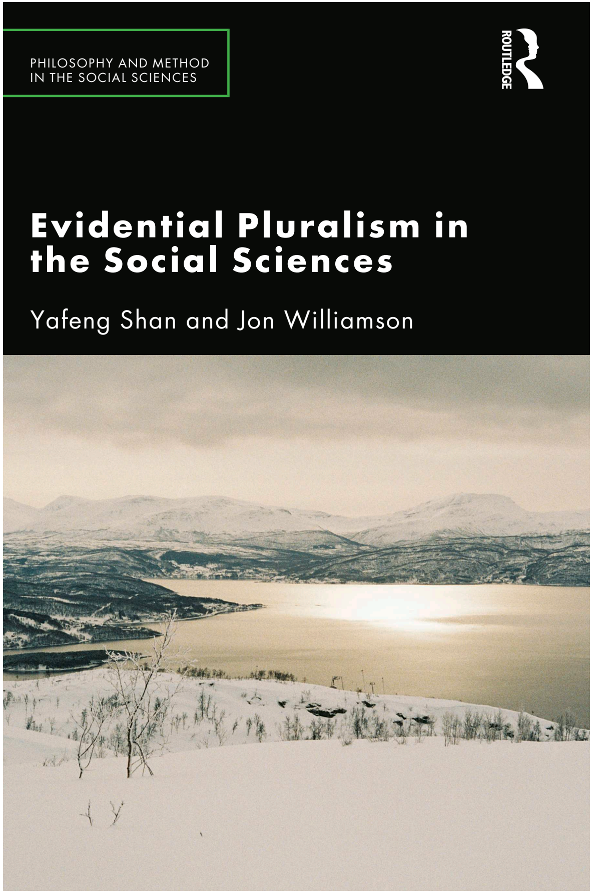
Veli-Pekka Parkkinen · Christian Wallmann
Michael Wilde · Brendan Clarke
Phyllis Illari · Michael P. Kelly
Charles Norell · Federica Russo · Beth Shaw
Jon Williamson

Evaluating Evidence of Mechanisms in Medicine

Principles and Procedures

Springer Open

… and EBP+



PHILOSOPHY AND METHOD IN THE SOCIAL SCIENCES

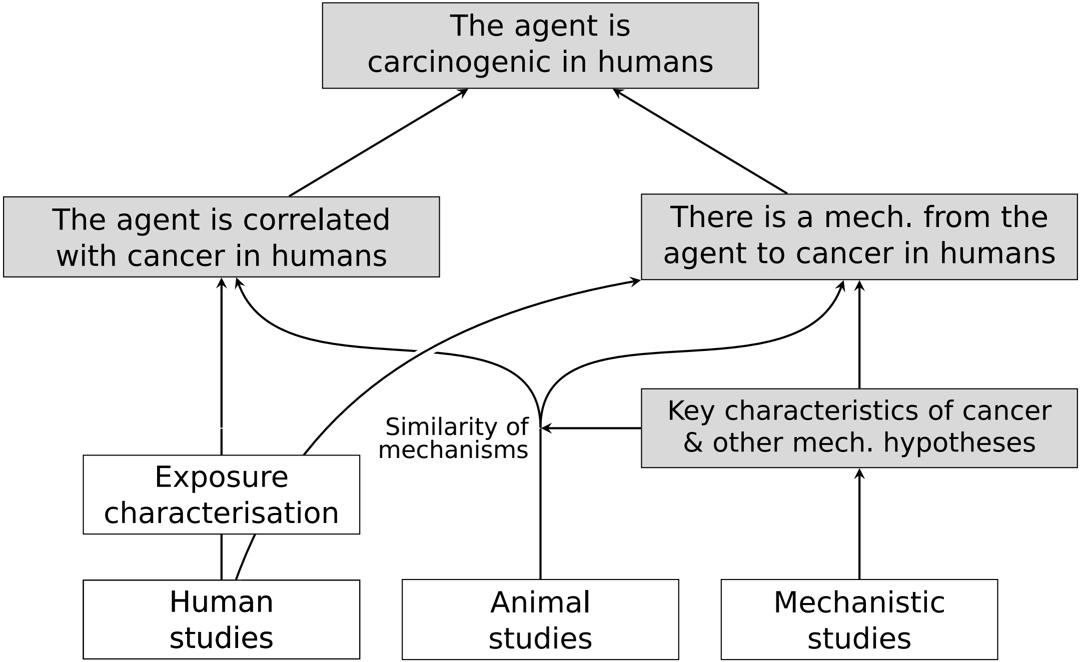
ROUTLEDGE

Evidential Pluralism in the Social Sciences

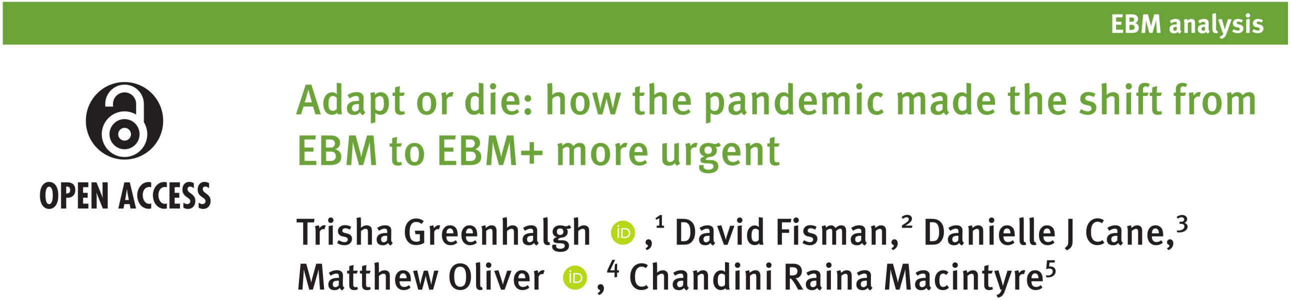
Yafeng Shan and Jon Williamson

Mechanistic evidence is now assessed in this way in exposure assessment

EG IARC integrates human, animal and mechanistic studies:


The agent is carcinogenic in humans
The agent is correlated with cancer in humans
There is a mech. from the agent to cancer in humans
Key characteristics of cancer & other mech. hypotheses
Similarity of mechanisms
Exposure characterisation
Human studies
Animal studies
Mechanistic studies

Would we have done better with EBM+ during the pandemic?



EBM analysis

OPEN ACCESS

Adapt or die: how the pandemic made the shift from EBM to EBM+ more urgent

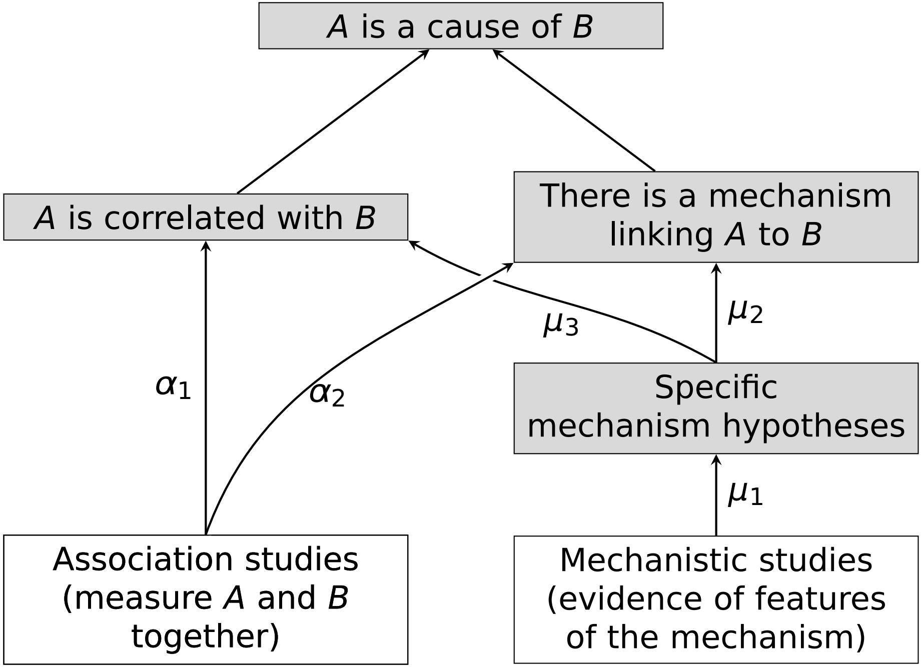
Trisha Greenhalgh,[1] David Fisman,[2] Danielle J Cane,[3] Matthew Oliver,[4] Chandini Raina Macintyre[5]

> Our central argument is that for some aspects of the pandemic, especially those characterised by a combination of complexity (multiple variables interacting dynamically with a high degree of uncertainty), urgency (decisions needed in days not years) and threat (the consequences of not acting could be catastrophic), mechanistic evidence has been mission-critical and RCTs difficult or impossible. Thousands of lives were likely lost as a result of what was incorrectly claimed to be an "evidence-based" approach—dismissing or downgrading mechanistic evidence, overvaluing findings from poorly designed or irrelevant RCTs, and advocating for inaction where RCT evidence was lacking. The pandemic is an epistemic opportunity for the EBM movement to come to better understand, debate and embrace EBM+. (Greenhalgh et al., 2022, p. 253.)

3 AI from the perspective of Evidential Pluralism

Why do we need to move from EBM to EBM+?

Because EBM focusses on the α-channels to the exclusion of the μ-channels.



A is a cause of *B*
A is correlated with *B*
There is a mechanism linking *A* to *B*
α_1
α_2
μ_3
μ_2
Specific mechanism hypotheses
μ_1
Association studies (measure *A* and *B* together)
Mechanistic studies (evidence of features of the mechanism)

Similar worries arise with machine learning.

An AI system can be viewed as an intervention to achieve some desirable outcome.

We can ask whether the AI system is effective at achieving the outcome:

- **?** Will it work in general? (developer) **Generic causation.**
- **?** Will it work for me? (client affected by the AI) **Single-case causation.**

The problem is that current machine learning validation methods are α-channel methods:

- IE Test the system on a test dataset or datasets.

London (2019) suggests that this EBM-motivated methodology is all that is required.

He suggests that there's no further need for interpretability or explainability.

But, as in the case of EBM, this methodology:

- ✗ Excludes a potentially informative stream of evidence.
- ✗ Is prone to bias.
  - EG Over-fitting the sample population from which training and test data are drawn.
- ✗ Makes successful extrapolation a mystery.
- ✗ Fails to address the important question of effectiveness in the single case.

Evidential Pluralism offers a way out of these problems

(i) Use the standard validation methods to establish correlation.

(ii) Hypothesise key features of mechanisms:

The mechanism of action, by which the intervention (AI) is responsible for the outcome.

Ancillary mechanisms that can counteract or enhance the effect of the AI.

(iii) Collect and scrutinize evidence to evaluate these hypotheses.

If one can establish that there is a mechanism that can account for the extent of the correlation, this confirms causation.

NB There's no need to establish all the details of the mechanism complex.

Features of the relevant mechanisms can also help to detect and assess biases.

They can also tell us whether the sample (training/test) population is sufficiently similar to the target population for success on the sample population to confirm effectiveness in the target population.

∴ Considering both associations and mechanisms allows us to better judge whether the AI system will work.

Compare opaque and interpretable models:

ANNs. Often good performance on the sample population.

✓ Can provide good evidence of correlation.

∴ α-channels can be enough to establish effectiveness on the sample population.

✗ It is hard to see the mechanism by which they work will work on a target population.

∴ There's a risk of bias / confounding / overfitting.

Decisions are being made on the basis of incidental features of the sample.

'[W]e demonstrate that recent deep learning systems to detect COVID-19 from chest radiographs rely on confounding factors rather than medical pathology, creating an alarming situation in which the systems appear accurate, but fail when tested in new hospitals.' (DeGrave et al., 2021, p. 610).)

Confounders include features outside the lungs and even outside the body.

Causal models. Can improve confidence in target population performance.

✓ Evidence of the mechanism of action can provide confidence that decisions are being made on the basis of relevant rather than incidental features.

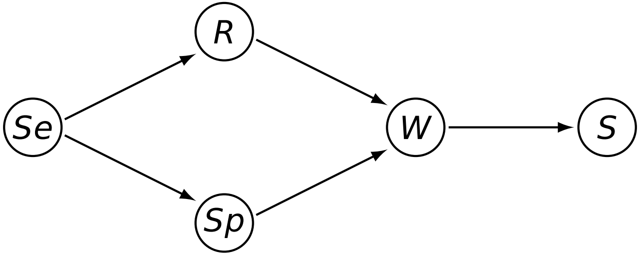
- This confirms effectiveness in both the sample and target populations.

NB Confidence **that** the system works, not the extent to which it works.

✓ Evidence of counteracting and enhancing mechanisms can inform generalisability.

$$P(W_t) = P(W_t|W_s)P(W_s) + P(W_t|\neg W_s)P(\neg W_s)$$

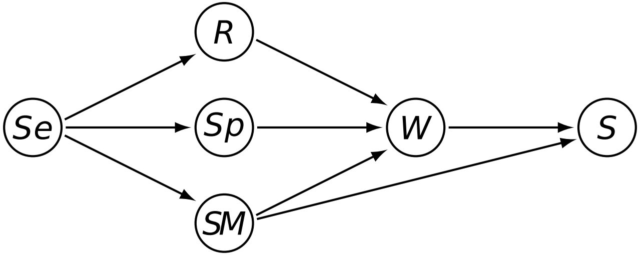
Season (Se) causes rain (R) and sprinkler use (Sp), causing wet (W) and slippery paths (S):



Suppose the sample population is Californian suburbs and the target population is:

India. Evidence of monsoon mechanisms requires reparameterisation.

Scandinavia. Here we have evidence of an interacting mechanism: snow-melt.



Single-case causation

Recall that the client affected by the AI is interested in the single case:

- EG Is the system right to deny me a loan?
- EG Uber's autonomous vehicle that killed a lady crossing a road in Tempe, Arizona, in 2018, kept misclassifying her and failed to determine her trajectory.

Single-case claims:

Causal claim. The system is effective (yields the right outcome) in my case.

Correlation claim. The (single-case) chance of the right outcome is higher when the system is used than when it isn't, conditional on potential confounders.

Mechanistic claim. There is a mechanism by which the system produces the right outcome in my case.

Association studies only provide weak evidence for the correlation and mechanistic claims.

Evidence of the features of the underlying mechanisms is crucial.

Counterfactual reasoning from mechanisms can confirm a single-case correlation claim.

Fairness

Will the AI work fairly?

Will it treat groups fairly? (regulator)

Will it treat me fairly? (client)

EG A 1980s program for screening medical student applicants to St George's discriminated against women and people with non-European sounding names.

This system picked up on pre-existing bias (Lowry and Macpherson, 1988).

The question of fairness is different but analogous to that of effectiveness.

As with effectiveness the concern is that the AI system will exploit 'biasing' features.

In the effectiveness case, the bias is epistemic:

Using features that are epistemically irrelevant can lead to poor outcomes on the target population.

In the fairness case, the bias is ethical:

Using features that are morally irrelevant (e.g., protected characteristics) can lead to unfair outcomes on the target population.

Again, we need to know how the system works to make these judgements.

EP seeks evidence of mechanisms, and this is exactly what we need here.

4 Explanation: by whom and how?

Our last question:

XAI: how do we explain whether the AI will work / work fairly?

First, who should construct the explanations?

The original vision of EBM is that clinicians and patients would be able to assess the evidence and judge for themselves whether something works.

✗ This was unrealistic: it's just too complicated and time-consuming.

Now, trusted regulatory committees evaluate these interventions.

But their evidence summaries can be presented to clinicians and patients.

An EP evaluation is in a sense even more complex.

One needs to evaluate evidence of mechanisms as well as evidence of correlation.

An EP explanation is a kind of conjunctive explanation (Schupbach and Glass, 2023).

∴ The hard work is going to have to be done by trusted evaluators.

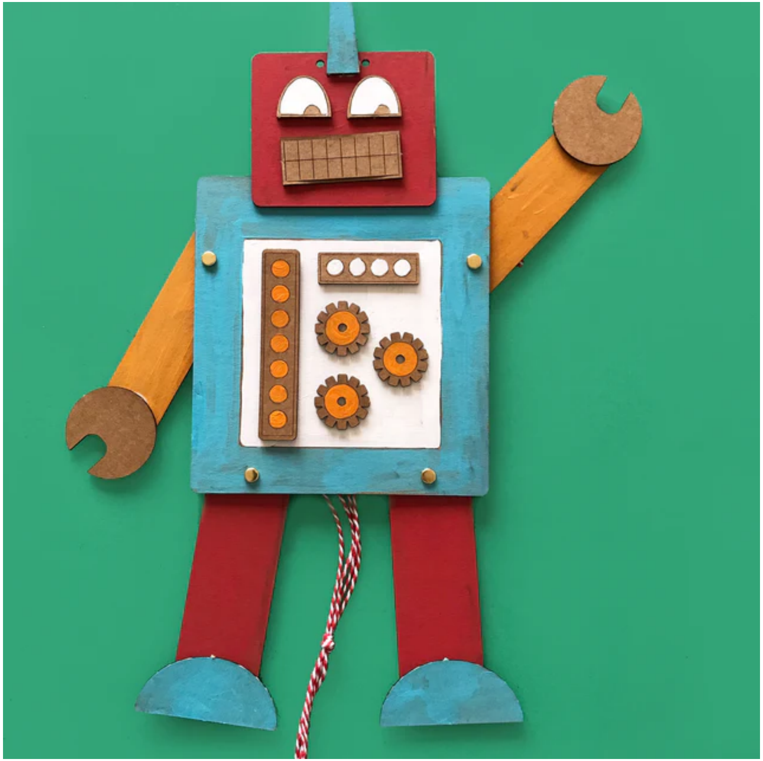
But an EP evaluation is in a sense easier to explain.

Confirmed mechanism hypotheses can underpin an accessible narrative explanation.

An EP evidence summary will be more accessible to the lay person.

∴ An EP evidence summary can meet the needs of XAI.

5 Conclusion



Evidential Pluralism provides a natural account of how to evaluate the effectiveness or fairness of an AI system.

It can handle both generic and single-case questions.

∴ It can meet the needs of developers, clients, and regulators, for example.

It motivates the following model of XAI:

Trusted evaluators review the evidence of effectiveness and/or fairness.

Mechanistic evidence as well as association studies.

An EP evidence summary can then be communicated to stakeholders.

For more on Evidential Pluralism: `https://blogs.kent.ac.uk/jonw/ep`

Bibliography

DeGrave, A. J., Janizek, J. D., and Lee, S.-I. (2021). AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619.

Greenhalgh, T., Fisman, D., Cane, D. J., Oliver, M., and Macintyre, C. R. (2022). Adapt or die: how the pandemic made the shift from EBM to EBM+ more urgent. *BMJ Evidence-Based Medicine*, Online first.

Guyatt, G., Cairns, J., Churchill, D., Cook, D., Haynes, B., Hirsh, J., Irvine, J., Levine, M., Levine, M., Nishikawa, J., Sackett, D., Brill-Edwards, P., Gerstein, H., Gibson, J., Jaeschke, R., Kerigan, A., Neville, A., Panju, A., Detsky, A., Enkin, M., Frid, P., Gerrity, M., Laupacis, A., Lawrence, V., Menard, J., Moyer, V., Mulrow, C., Links, P., Oxman, A., Sinclair, J., and Tugwell, P. (1992). Evidence-based medicine: A new approach to teaching the practice of medicine. *JAMA*, 268(17):2420–2425.

London, A. J. (2019). Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *Hastings Center Report*, 49(1):15–21.

Lowry, S. and Macpherson, G. (1988). A blot on the profession. *BMJ*, 296(6623):657–658.

Russo, F. and Williamson, J. (2007). Interpreting causality in the health sciences. *International Studies in the Philosophy of Science*, 21(2):157–170.

Schupbach, J. N. and Glass, D. H., editors (2023). *Conjunctive explanations: the nature, epistemology, and psychology of explanatory multiplicity*. Routledge, New York & Abingdon.

Williamson, J. (2019). Establishing causal claims in medicine. *International Studies in the Philosophy of Science*, 32(2):33–61.