Sjur K Dyrkolbotn

# A scalable approach to causal responsibility

Western Norway
University of
Applied Sciences

# The liberal consensus on causality

1. Causality is not relevant to nature and metaphysics.

   *[T]he reason why physics has ceased to look for causes is that, in fact, there are no such things.*

   Bertrand Russell, *On the notion of cause*, 1919.

2. Causality is a common-sense notion for allocating moral and legal responsibility *post hoc.*

   *Each case must be judged in the light of its own facts and by resorting not to the refinements of the philosophical doctrine of causation but to the commonplace tests which the ordinary business men conversant with such matters would adopt.*

   Lord MacMillan, *Yorkshire Dale Steamship Co v Minister of War Transport (1942, A.C. (H.L.) 691, 706)).*

# The consensus is breaking down

- Why?
- Complexity and interconnectedness.
- Dependence and lack of (political) control.
- Inequality and polarisation.

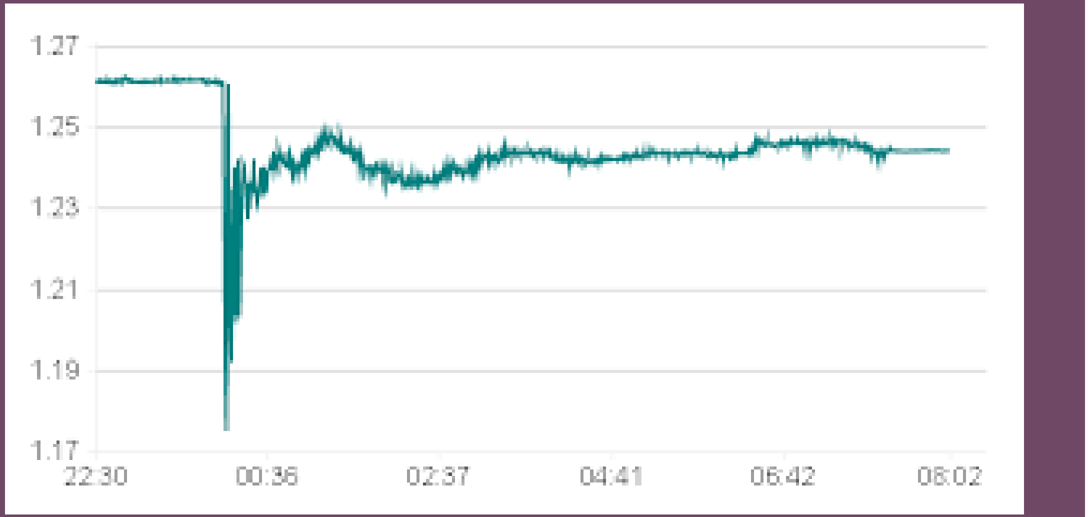
# Consequences

1. Overattribution of responsibility.
    - Climate change, wealth, microaggresion, hate speech, #metoo, algorithmic trading.

2. Underattribution of responsibility.
    - CO2 emissions, inequality, systemic bias, discrimination, rape, algorithmic trading.

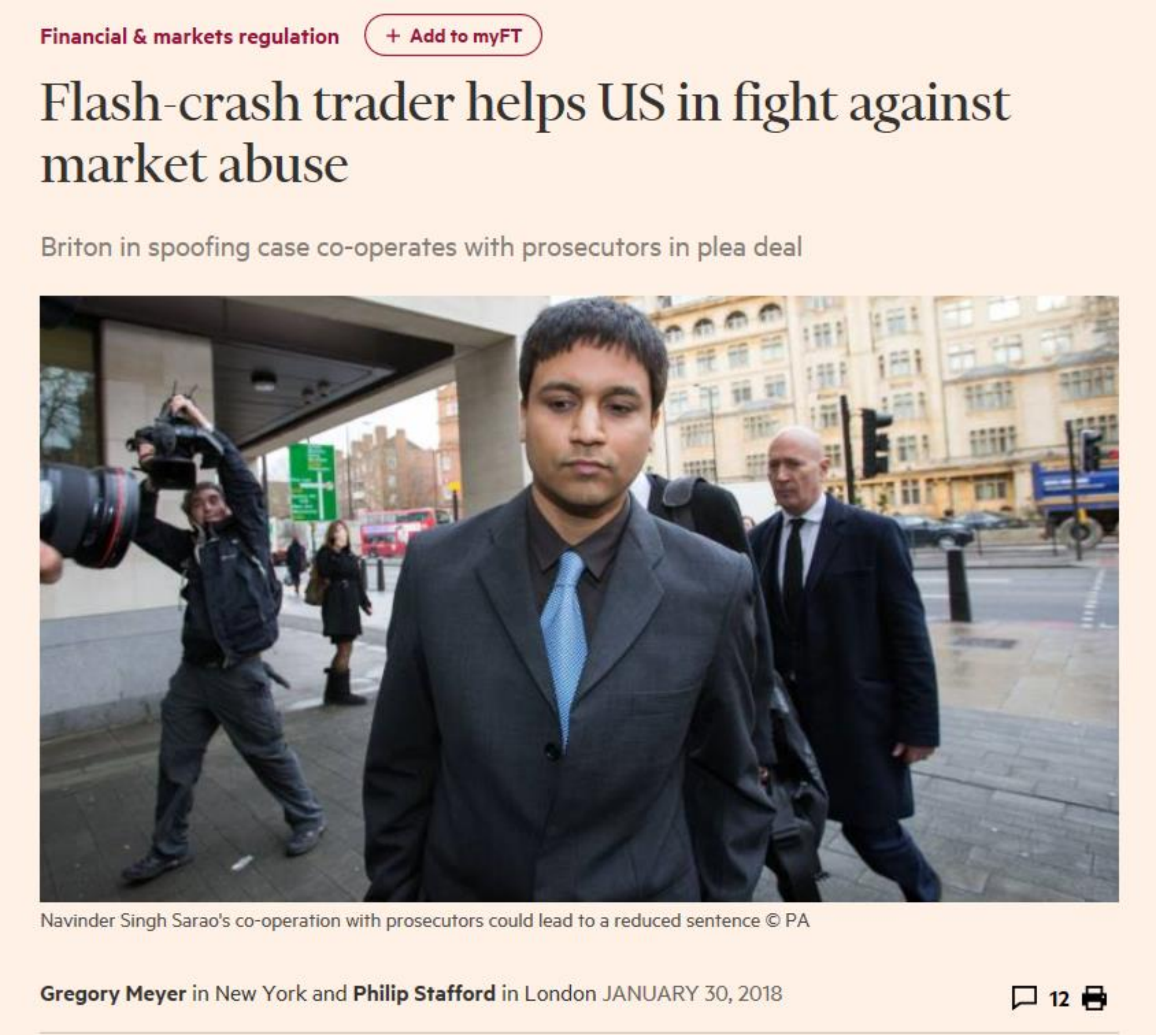
- Growing instability of responsibility judgements.

- Is the breakdown of causal consensus a contributing cause?



# A red herring

Why no action against flash traders – is the problem only caused by «spoofers»?



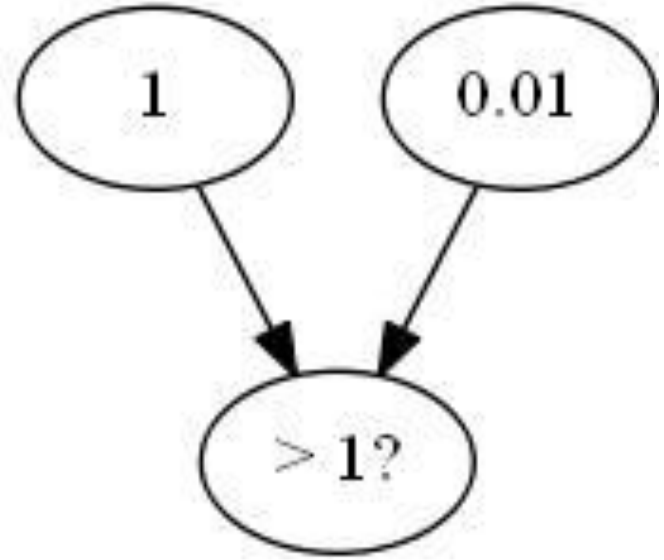
Financial & markets regulation

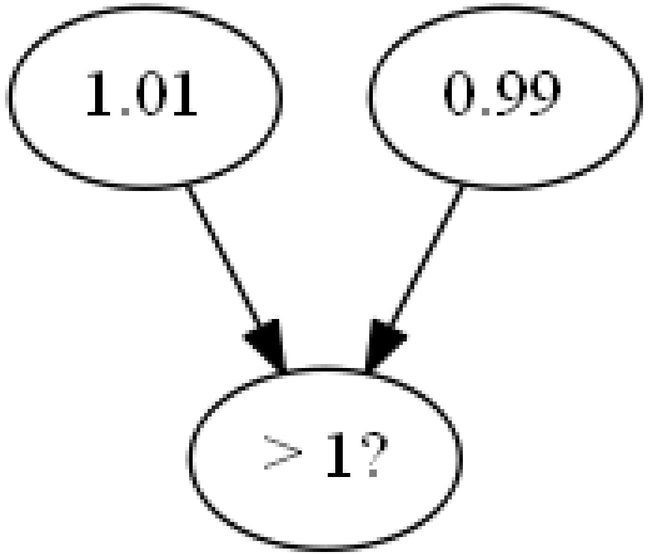
## Flash-crash trader helps US in fight against market abuse

Briton in spoofing case co-operates with prosecutors in plea deal

Navinder Singh Sarao's co-operation with prosecutors could lead to a reduced sentence © PA

**Gregory Meyer** in New York and **Philip Stafford** in London JANUARY 30, 2018





# The commonplace notion of cause

- A causes B if, and only if, B would not have occurred *but for* A.

- Dominant in law, described as the «common understanding of cause» in *Burrage v. US, 134 S. Ct. 881 - Supreme Court 2014*.

- Common, but problematic – mere trifles can be but-for causes and major contributions can fail the test.

# Overattribution – the traditional approach

The distinction between legal causation and factual causation.

Legal causation (proximate causation) is *not* about causation, but about ethics, policy and modelling (domain knowledge).

Foreseeability and risk.

# Underattribution – the traditional approach

Causal contribution is enough, provided it is material.

- *Williams v The Bermuda Hospitals Board (Bermuda) [2016] UKPC 4*.
- Many sources of asbestos; all contribute to pneumoconiosis (factual causes), *Bonnington Castings Ltd v Wardlaw [1956] AC 613.*

Causal contribution requires evidence of contribution to a causal process that actually caused the outcome.

- Asbestos and smoking; asbestos is not necessarily a contributing cause of lung cancer, if smoking actually caused lung cancer by a *different process.*

# Weaknesses of the traditional approach

Reliant on the but-for test as the default rule – a presumption that but-for causation must be proven.

Corrections are *ad hoc* – no precise and coherent theory.

Corrections create *slippery slopes* – invite opportunistic views on responsibility.

# Distinct principles to prevent underattribution in the law

THE SUFFICIENT CONDITION TEST.

THE NESS TEST.

THE ACCELLERATION PRINCIPLE.

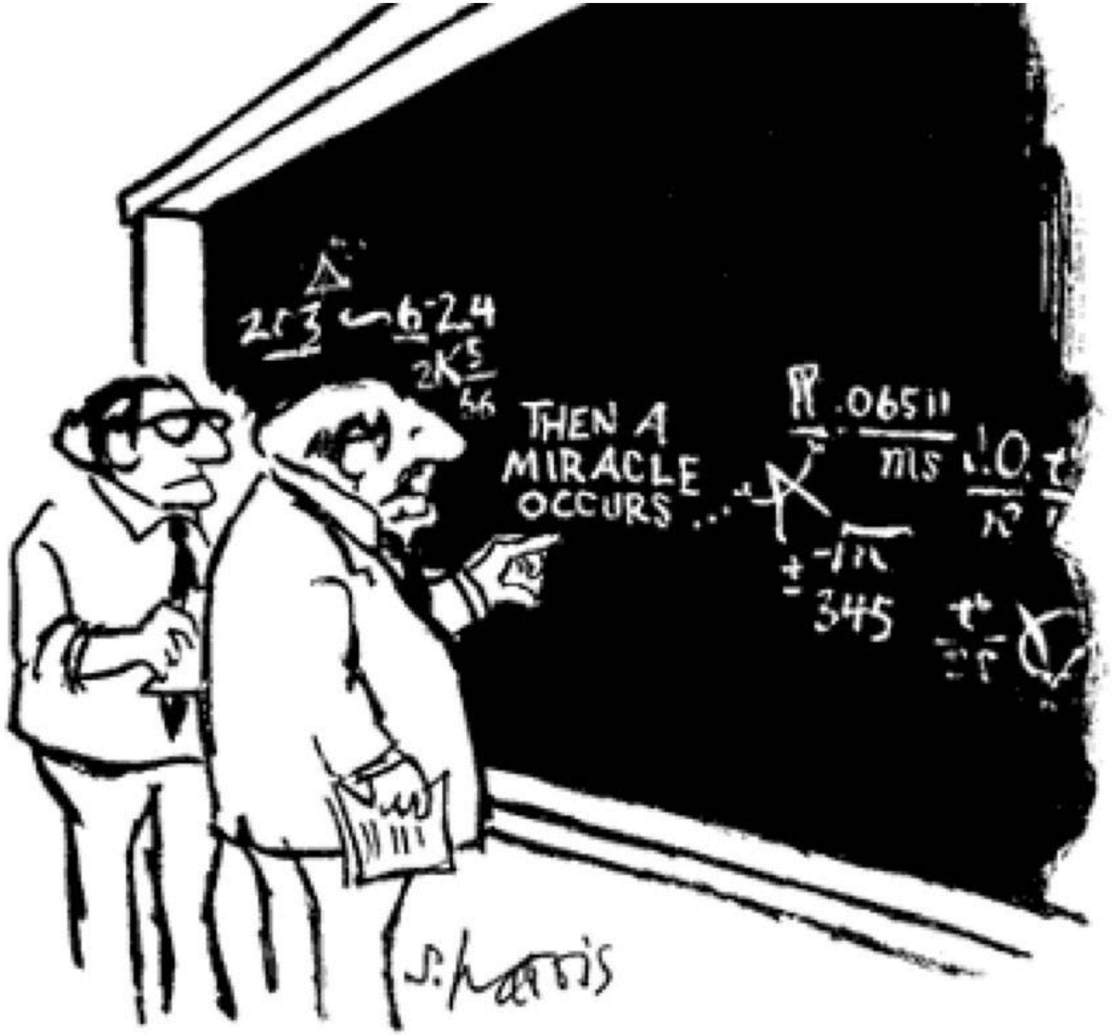
THE CONTRIBUTION PRINCIPLE.

THE SIGNIFICANT FACTOR PRINCIPLE.

**THE GENERALISED BUT-FOR TEST.**

# Formal theories

- Domain experts encode knowledge in formal models, we apply a definition of cause **to the model.**
  - The link with legal and factual causality.
- Great potential, but current definitions are **incorrect** and **intractable**.
- Reliance on miracles.



# Overdetermination and preemption

- Two challenges for formal models of actual causality.

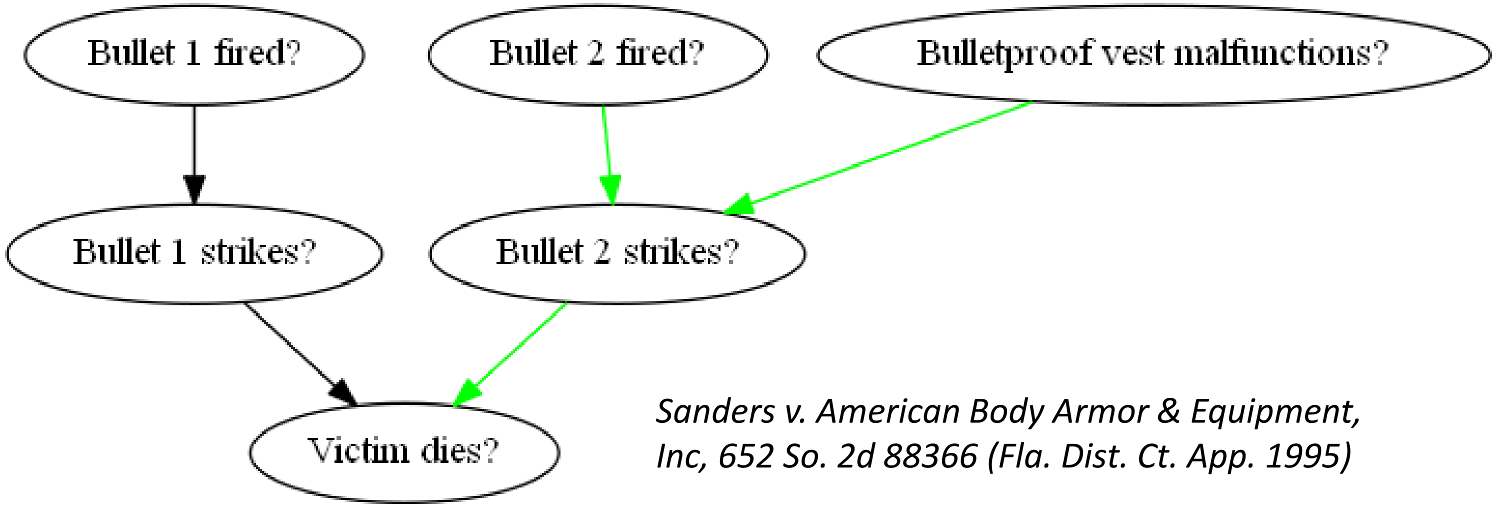
Causes that are not necessary for the outcome – causes that fail the but-for test.

- How to identify such causes correctly and efficiently?
- **Contributing** causes.

Putative causes that are prevented by other causes from having an effect on the outcome.

- How to distinguish between contributing causes and putative causes that have been preempted?

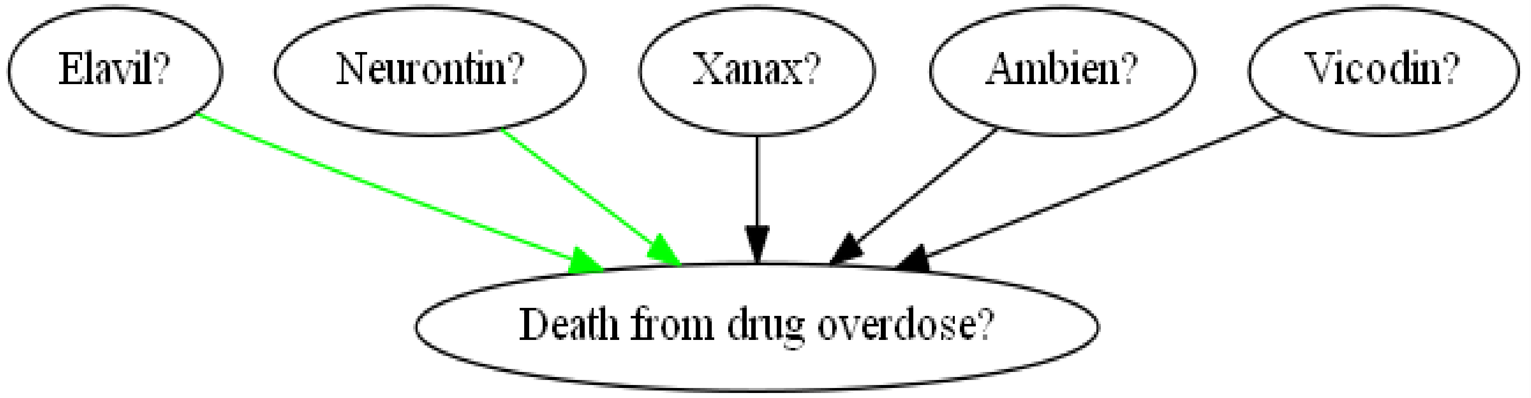
# A simple example



*Sanders v. American Body Armor & Equipment, Inc, 652 So. 2d 88366 (Fla. Dist. Ct. App. 1995)*

- **Both bullets were fired, so the malfunctioning vest was not a but-for cause of death. But it would have been if the first bullet had not been fired, and this was sufficent for factual causation.**

# A more challenging example

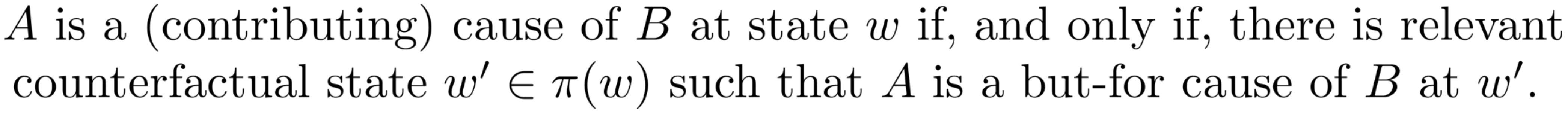


- *South Coast Framing v. WCAB - S215637 (Supreme Court of California 2015).*

**Neither Elavil nor Neurontin were but-for causes of death, but each would have been together with *some counterfactual combination of dosages* of the other drugs. This was a marginal causal contribution, but sufficient to establish factual causality in a workers compensation dispute.**

# Intractability

- The standard approach to overdetermination in formal models:

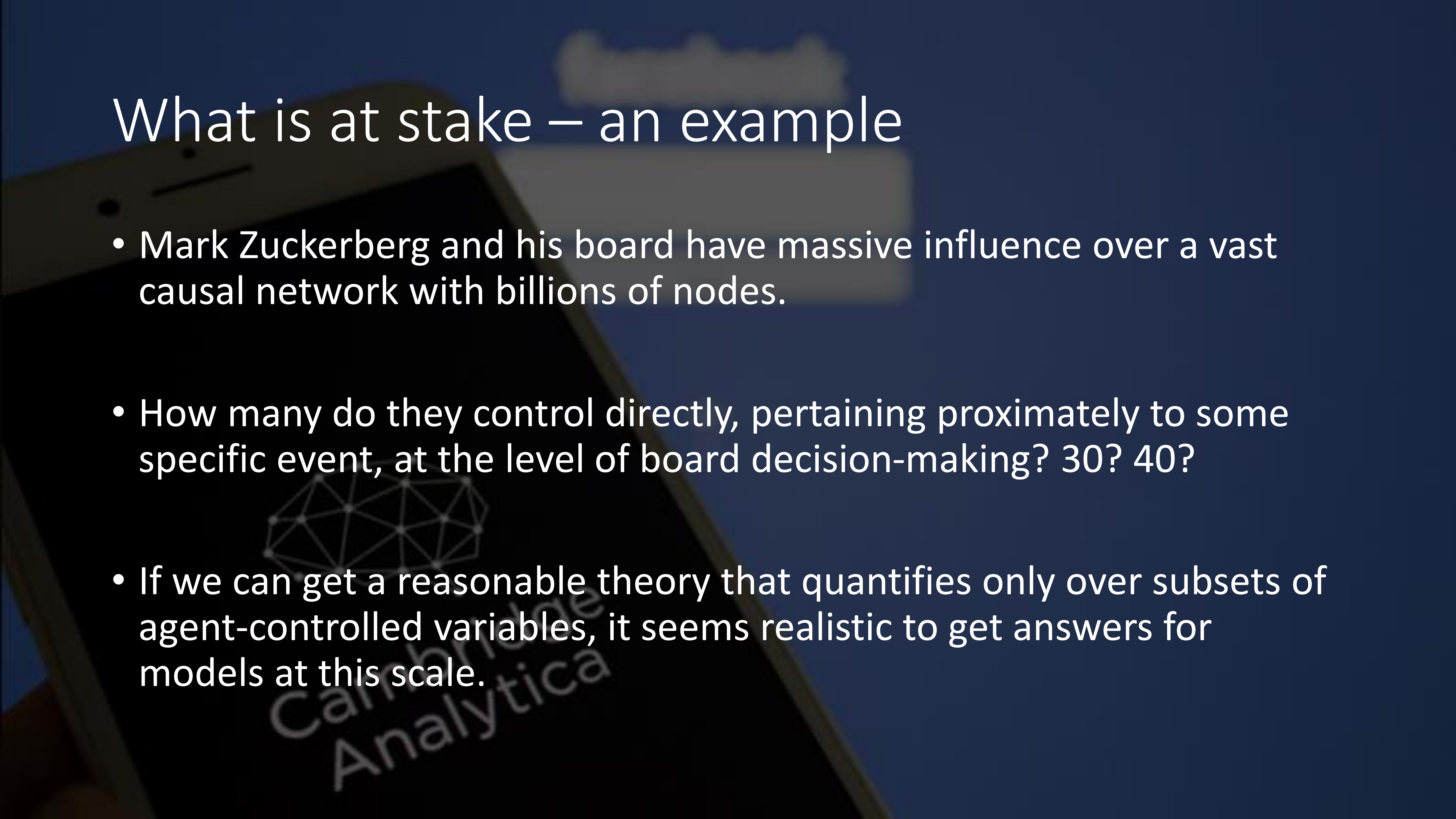


$A$ is a (contributing) cause of $B$ at state $w$ if, and only if, there is relevant counterfactual state $w' \in \pi(w)$ such that $A$ is a but-for cause of $B$ at $w'$.

- Structurally reasonable, but existing theories ask us to quantify over **too many** counterfactual states (at least as many as the number of subsets of variables in the model).

- To recognise overdetermination is suffices to quantify over all subsets of **agent-controlled** variables.

# What is at stake – an example

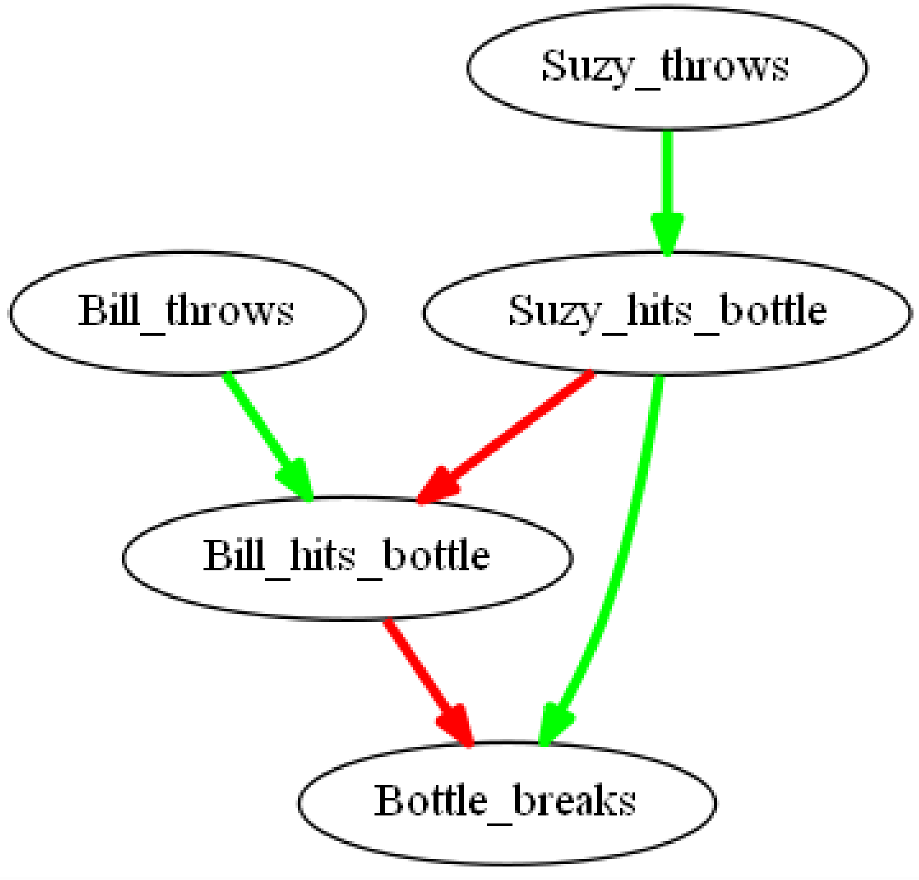


- Mark Zuckerberg and his board have massive influence over a vast causal network with billions of nodes.

- How many do they control directly, pertaining proximately to some specific event, at the level of board decision-making? 30? 40?

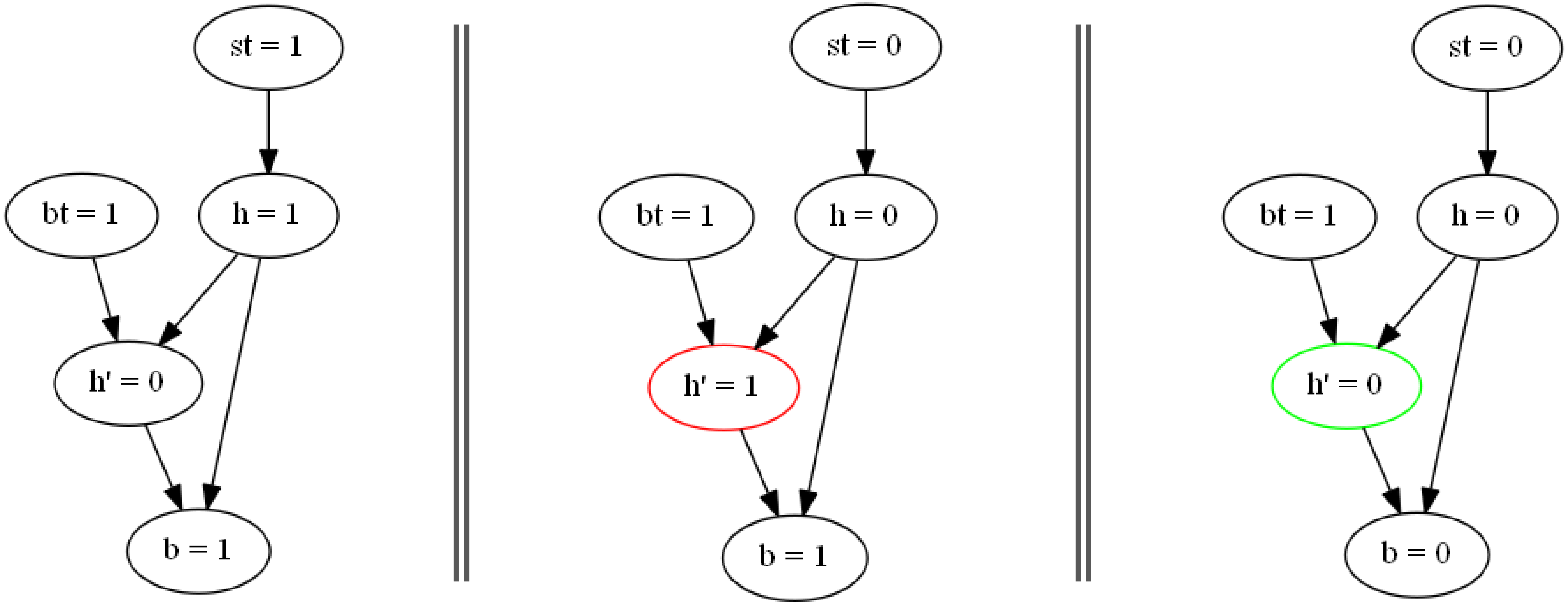
- If we can get a reasonable theory that quantifies only over subsets of agent-controlled variables, it seems realistic to get answers for models at this scale.

# The challenge – modelling preemption



- Suzy and Bill both throw a rock. Suzy strikes the bottle first, but Bill would have struck it if Suzy had not done so. Hence, Suzy is not a but-for cause of the bottle shattering.
  - Overdetermination principles not enough, since they imply that Bill is a cause as well.

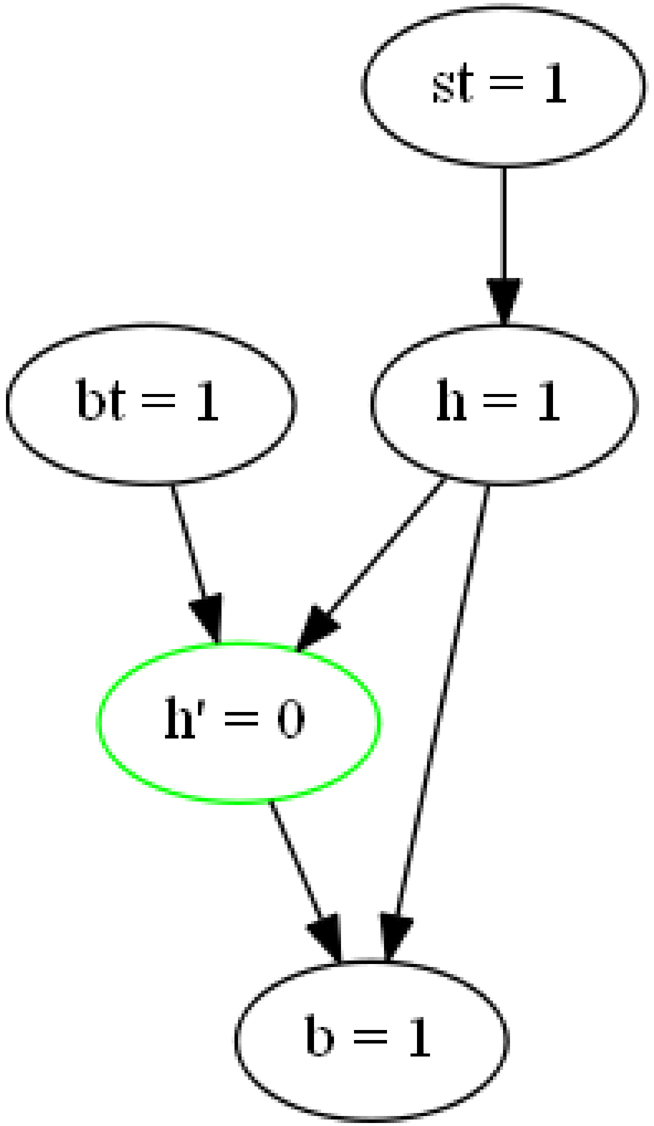
- The *actus interveniens* principle – how to recognise preemption?



# The (generalised) *actus interveniens*

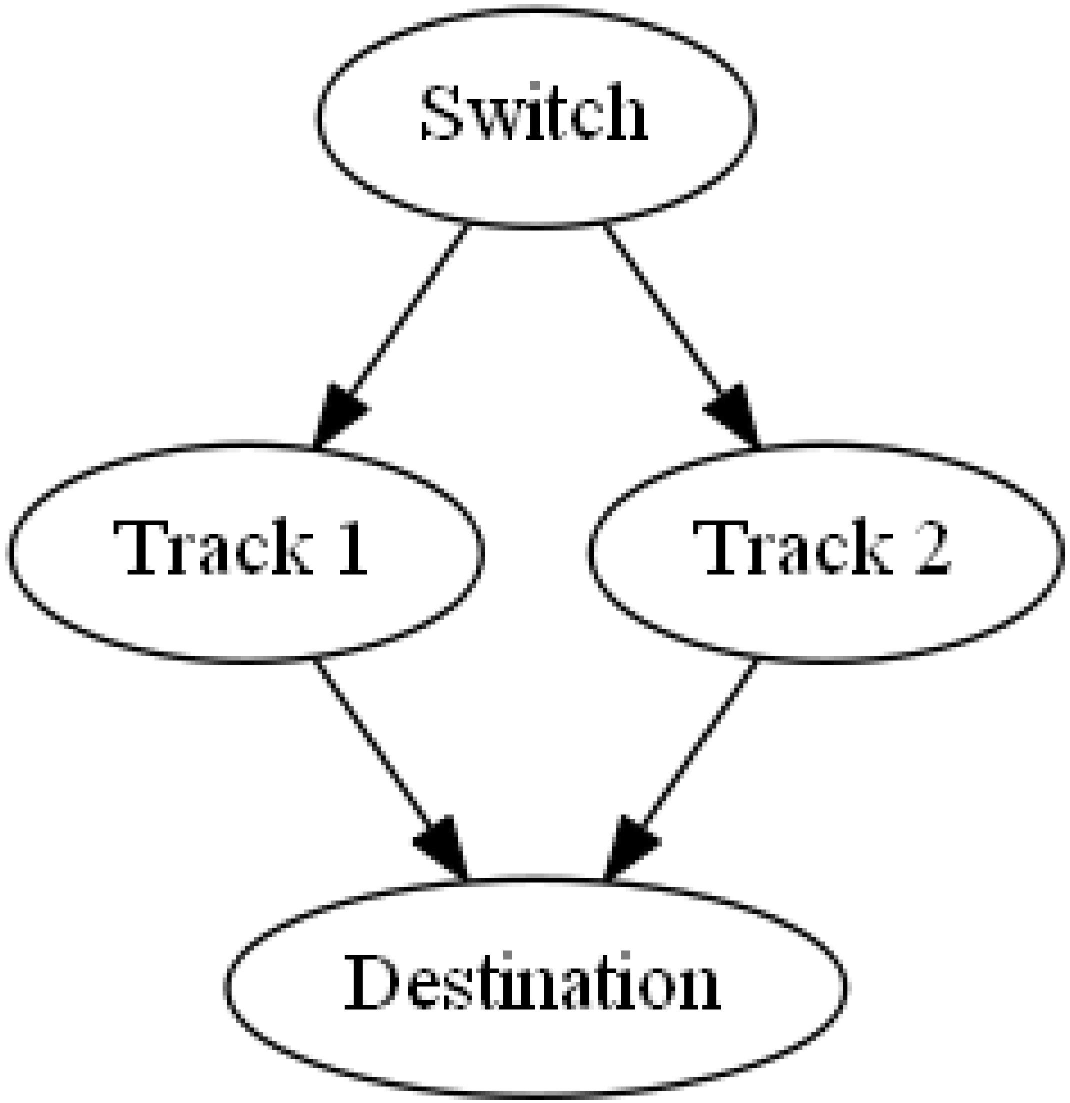
# The standard formal approach

- Formal theories do not use *actus interveniens* reasoning; they do not say that Suzy prevents Bill's action from having the putative effect.
- Instead, they say Suzy is necessary for b = 1 on the assumption that Bill would have missed the bottle, so Bill is ruled out by the minimality principle.



# Incorrectness

- Existing theories of causality quantify over counterfactual states that are **impossible according to the model.**
- Impossible counterfactuals result in overattribution.



# The challenge

- Counterfactual theories are still the right way forward.
- Legal and moral responsibility **presuppose** a counterfactual account.

# Why allow miracles in the first place?

Introduced (mainly) to model preemption.

But to what extent are they really necessary?

We could try to *identify* the potential witnesses of preemption *before* looking for causes.

# Towards a scalable approach

Consider only source variables (variables controlled by the agent), not variables whose values are determined by the model.

Efficiently identify the set of variables that witness to preemption.

Only allow miracles when testing for preemption, not when looking for causes.

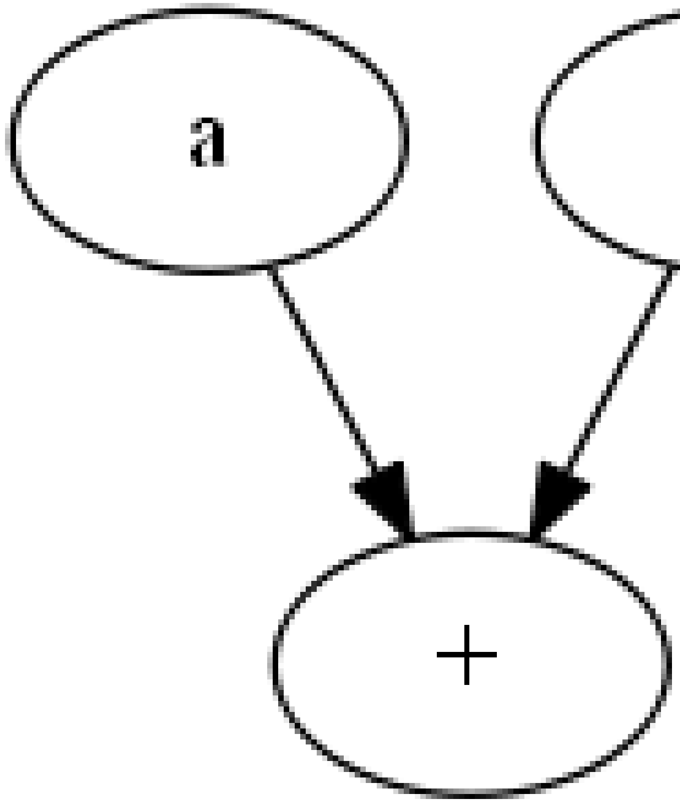
# A formalisation

- Frames: $G = (N, E)$ with $E \subseteq N \times N$ (let $E^{-}(n) = \{n \mid (m, n) \in E\}$).
- Sources: $S = \{n \in N \mid E^{-}(n) = \emptyset\}$.
- Possible states: the set $W_G$ of all functions $w$ such that:

$$\begin{array}{l} \forall n \in N : w(n) \in D_n (\text{domain of } n) \\ \forall n \in N \setminus S : w(n) = n\Big( \prod\limits_{m \in E^{-}(n)} w(m) \Big) \end{array}$$

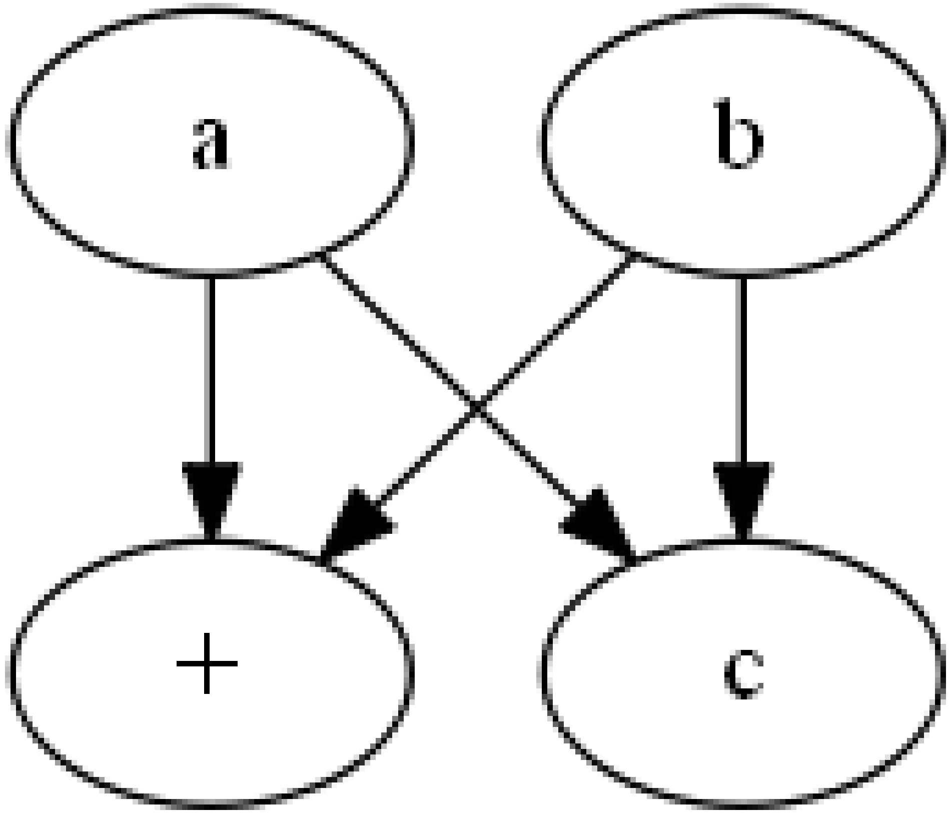
- Agreement: $a(w_1, w_2) = \{s \in S \mid w_1(s) = w_2(s)\}$.
- Distance: $d_G : W_G \times W_G \to \mathbb{N}$ defined by:

$$\forall w_1, w_2 \in W_G : d_G(w_1, w_2) = |S \setminus a(w_1, w_2)|$$



- $N = \{a, b, +\}$
- $D_a = \{1, 0\}, D_b = \{1, 0\}, D_+ = \{0, 1, 2\}$.
- $S_G = \{a, b\}$.
- $W_G = \{w_1, w_2, w_3, w_4\}$ with
- $$\begin{array}{ll} w_1(a) = 1 = w_1(b), & w_1(+) = 2 \\ w_2(a) = 0 = w_2(b), & w_2(+) = 0 \\ w_3(a) = 1, w_3(b) = 0, & w_3(+) = 1 \\ w_4(a) = 0, w_4(b) = 1, & w_4(+) = 1 \end{array}$$
- $a(w_1, w_4) = \{b\}$, $d(w_1, w_4) = 1$.

# Example



- $N = \{a, b, +, c\}$ with $D_c = \{1, 0\}$.
- $c : D_a \times D_b \to D_c$.
- $S_G = \{a, b\}$.
- $W_G = \{w_1, w_2, w_3, w_4\}$ with
- $$\begin{array}{l} w_1(a) = 1, w_1(b) = 1, w_1(+) = 2, w_1(c) = 1 \\ w_2(a) = 0, w_2(b) = 0, w_2(+) = 0, w_2(c) = 0 \\ w_3(a) = 1, w_3(b) = 0, w_3(+) = 1, w_3(c) = 1 \\ w_4(a) = 0, w_4(b) = 1, w_4(+) = 1, w_4(c) = 1 \end{array}$$
- $a(w_1, w_4) = \{b\}, d(w_1, w_4) = 1$.

# Example contd.

# Interventions

- Given a function $w \in W_G$ and some $M \subseteq N$ we define the (unique, by recursion) function that satisfies the following constraint:

$$w[f \leftarrow M](n) = \begin{cases} f(n) \text{ if } n \in M, \\ n\Big( \prod_{m \in E^-(n)} w[f \leftarrow M](m)\Big) \text{ if } n \in N \setminus (S \cup M), \\ w(n) \text{ otherwise.} \end{cases}$$

- Given $W \subseteq W_G$ we define $W[f \leftarrow M] = \{w[f \leftarrow M] \mid w \in W\}$

# More about interventions

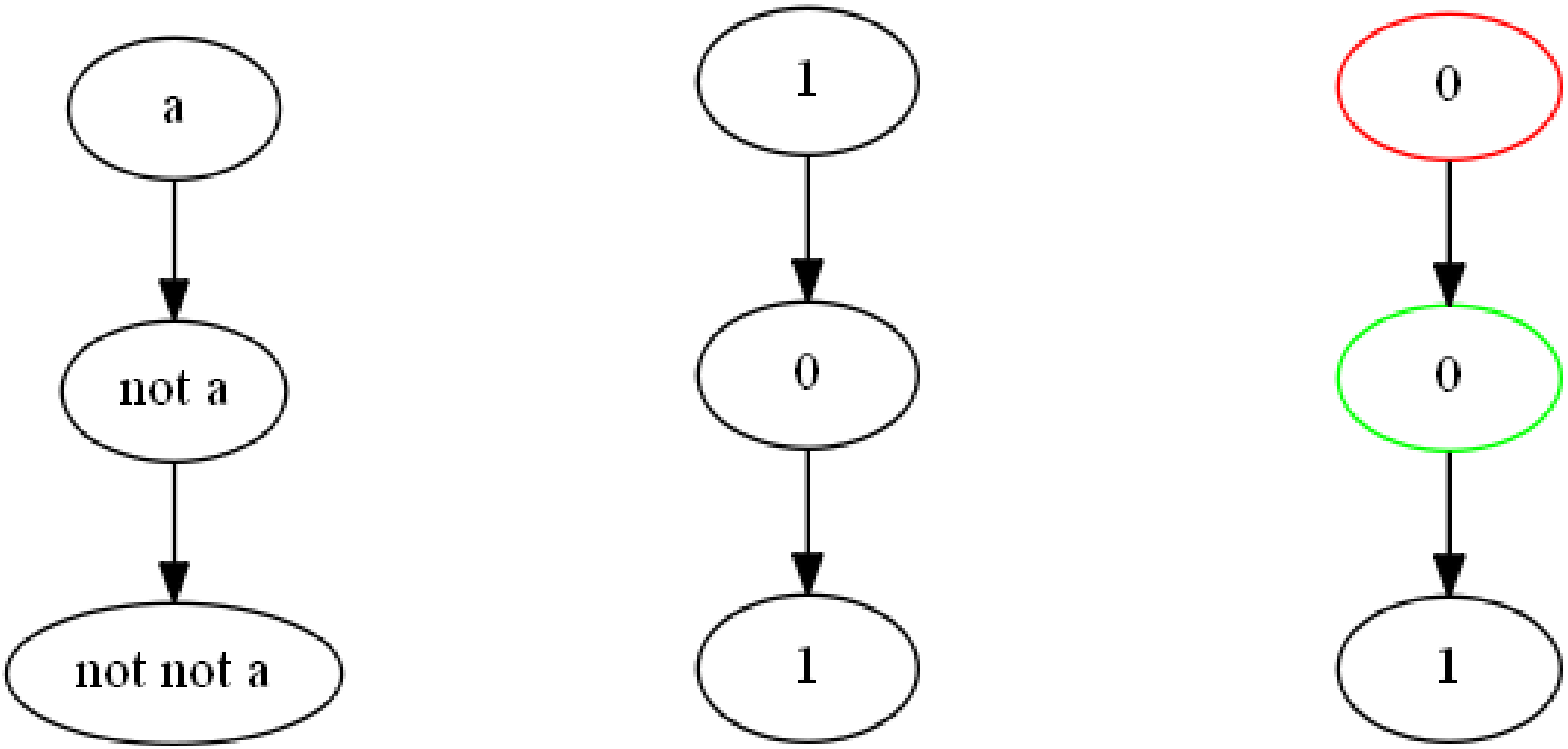
- Interventions at the source lead to harmelss counterfactuals – a unique new state of the same model.

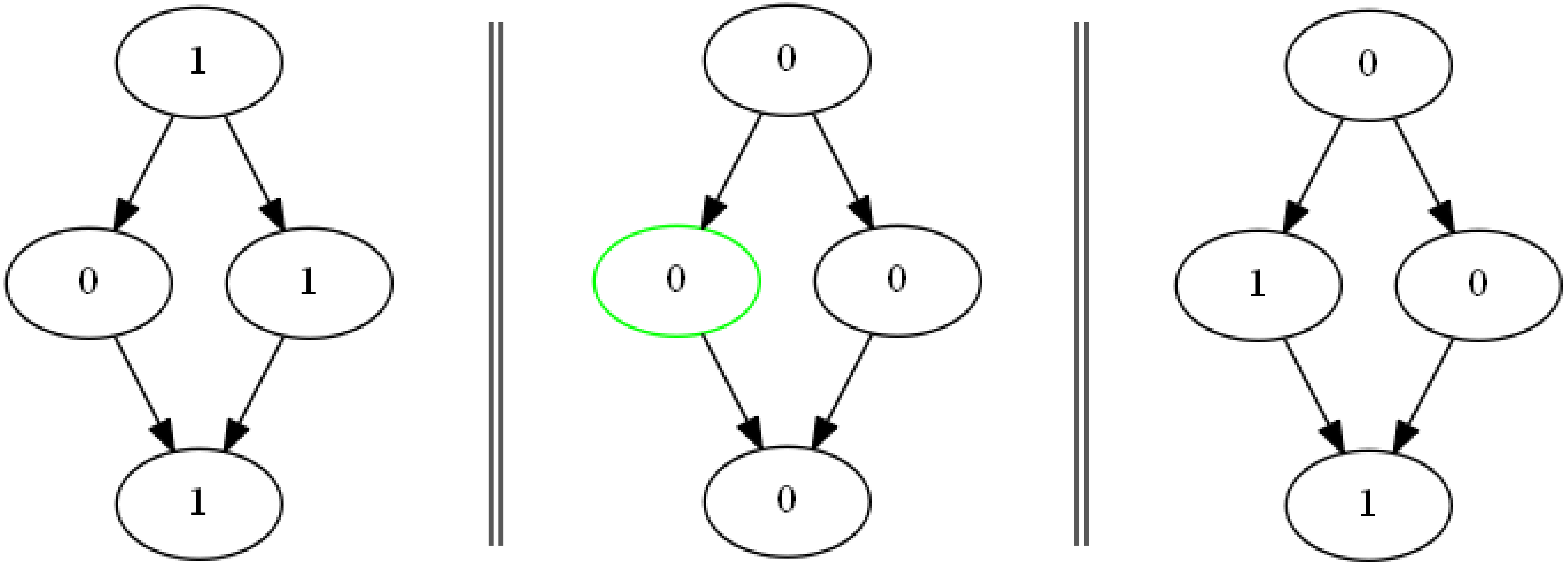
- Interventions at dependent variables are miracles – they generally contradict the information encoded in the model by the domain expert.

- When should a causal expert (e.g., a judge) be allowed to do such a thing?

- **Not** when looking for causes!

# Example of a miracle – disappearing negation





# The miraculous switch

# A generalisation of the NESS test

- The traditional NESS test: A is a cause of B if, and only if, A is a necessary element of a sufficient set of conditions for B.
  - Only works when «being a condition for» is well-defined.
  - For dependent variables, it is not (without miracles).

- The generalised NESS test: A is **correlate** of B if, and only if, A is a necessary **consequence** of a sufficient set of conditions for B.
  - Sets of sufficient conditions are always subsets of S.

- Reduces to standard NESS test when attention is restricted to a single equation (a game, as in Braham and Van Hees’ work).

# A formal definition

- Given a model $G$ and a state $w \in W_G$.

- $m \in N$ passes the generalised NESS test for $n \in N$ if, and only if, there is some $w' \in W_G$ such that:

$$\begin{array}{ll} (1) & w'(n) \neq w(n) \;\&\; w'(m) \neq w(m) \\ (2) & \forall w'' \in W_G : d(w'', w) < d(w', w) \Rightarrow w''(n) = w(n) \end{array}$$

- Let $NESS(w, n)$ denote the set of all $m \in N$ that pass the generalised NESS test for $n$ at $w$.

- For models with a single equation, $NESS(w, n)$ is exactly the standard NESS causes.

# A fixed-point definition of causality

- Given a model $G$ and a set of states $W_G$.
- For $n \in N, w \in W_G$, denote by $Cause(w,n)$ the contributing causes of $w(n)$ at $w$.
- Definition: $Cause(w,n) = L \cap S$, where $L$ is **a** greatest set such that for all $m \in L$ there is some $w' \in W_G[w \leftarrow (N \setminus L)]$ satisfying:

$$\begin{array}{ll} (1) & w'(n) \neq w(n) \;\&\; w'(m) \neq w(m) \\ (2) & \forall w'' \in W_G[w \leftarrow (N \setminus L)] : d(w'',w) < d(w',w) \Rightarrow w''(n) = w(n) \end{array}$$

- **Algorithm:** start from $NESS(w,n)$ and check against conditions (1) and (2) above to prune the set until you reach a fixed point.

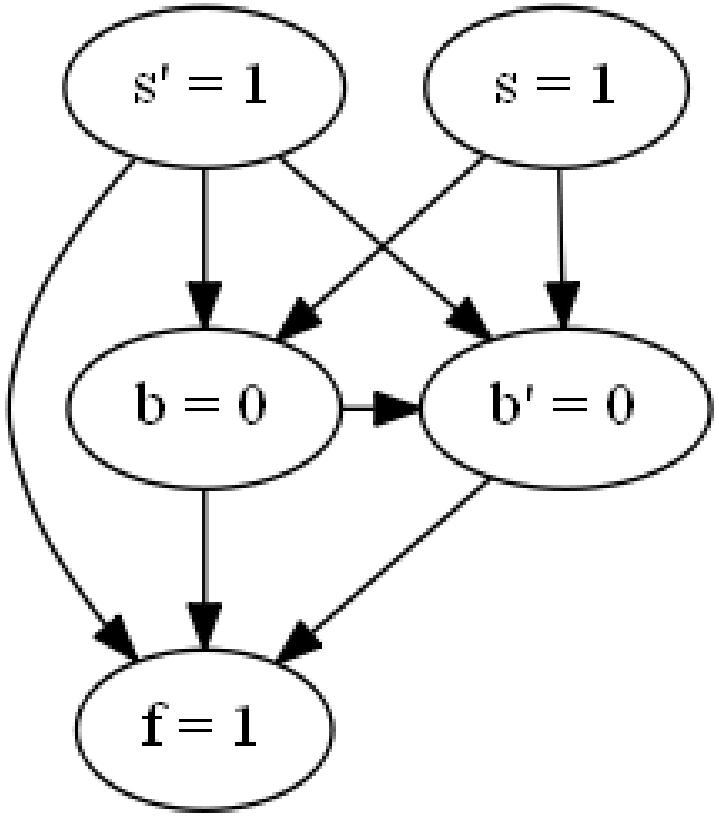
- **Claim**: Pruning $NESS(w, n)$ once is enough to determine $S \cap L$.
- Sketch of argument: Assume $s \in S$ is removed in iteration $i > 1$ of the algorithm. So there are enough interventions like this:

  $$s \longrightarrow \ldots \longrightarrow o_1 \longrightarrow \ldots \longrightarrow n$$

  where $o_1$ is held fixed at $w(o_1)$, blocking contribution from $s$. At least some such $o_1$ must have been removed in iteration $i - 1$, since otherwise $s$ would have been removed earlier. But then we have interventions like this:

  $$s \longrightarrow \ldots \longrightarrow o_1 \longrightarrow \ldots \longrightarrow o_2 \longrightarrow \ldots \longrightarrow n$$

  where $w(o_2)$ blocks the contribution from $o_1$. This blocks contribution from $s$ as well, so $s$ would have been removed at iteration $i - 1$ after all.
- **Corollary**: $S \cap L$ is unique (the set is well-defined).



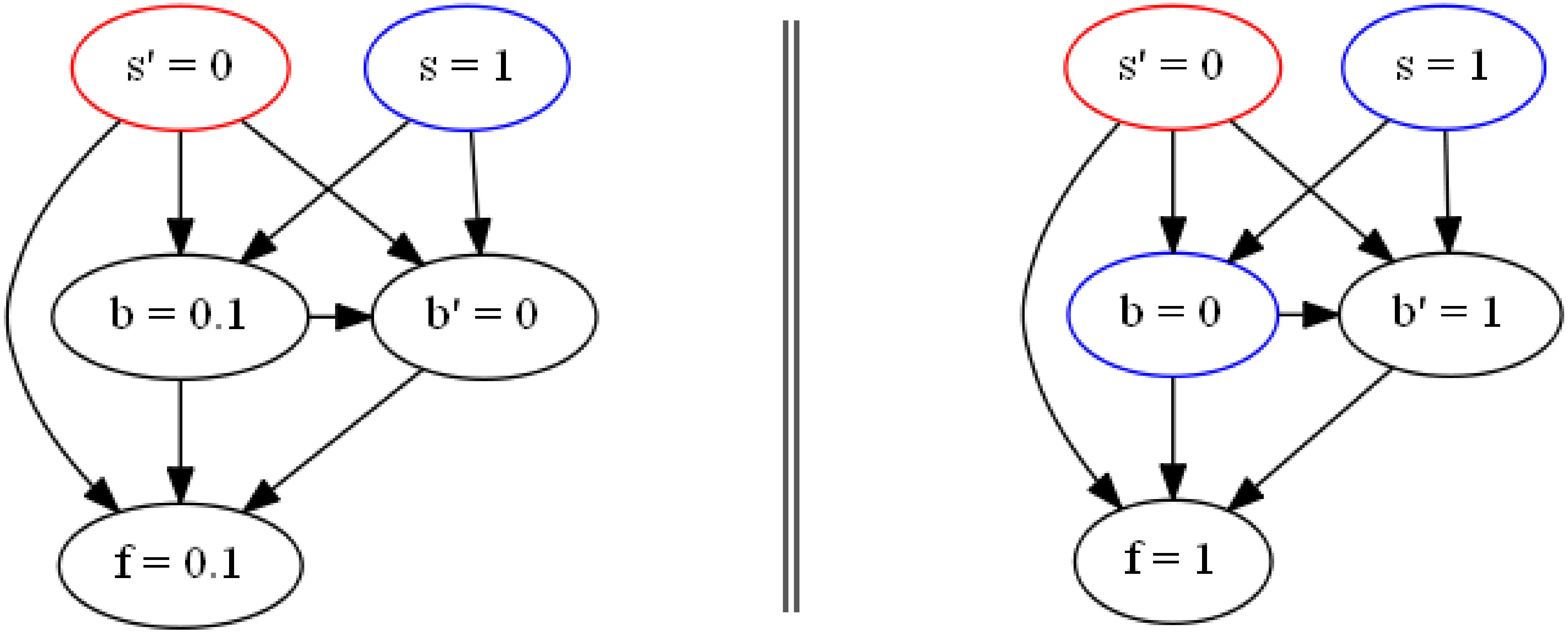
$$b(s', s) = \begin{cases} 0 \text{ if } s' = 1 \\ s - 0.9 \text{ otherwise} \end{cases}$$

$$b'(b, s', s) = \begin{cases} 0 \text{ if } s' = 1 \text{ or } b = 0.1 \\ 1 \text{ otherwise} \end{cases}$$

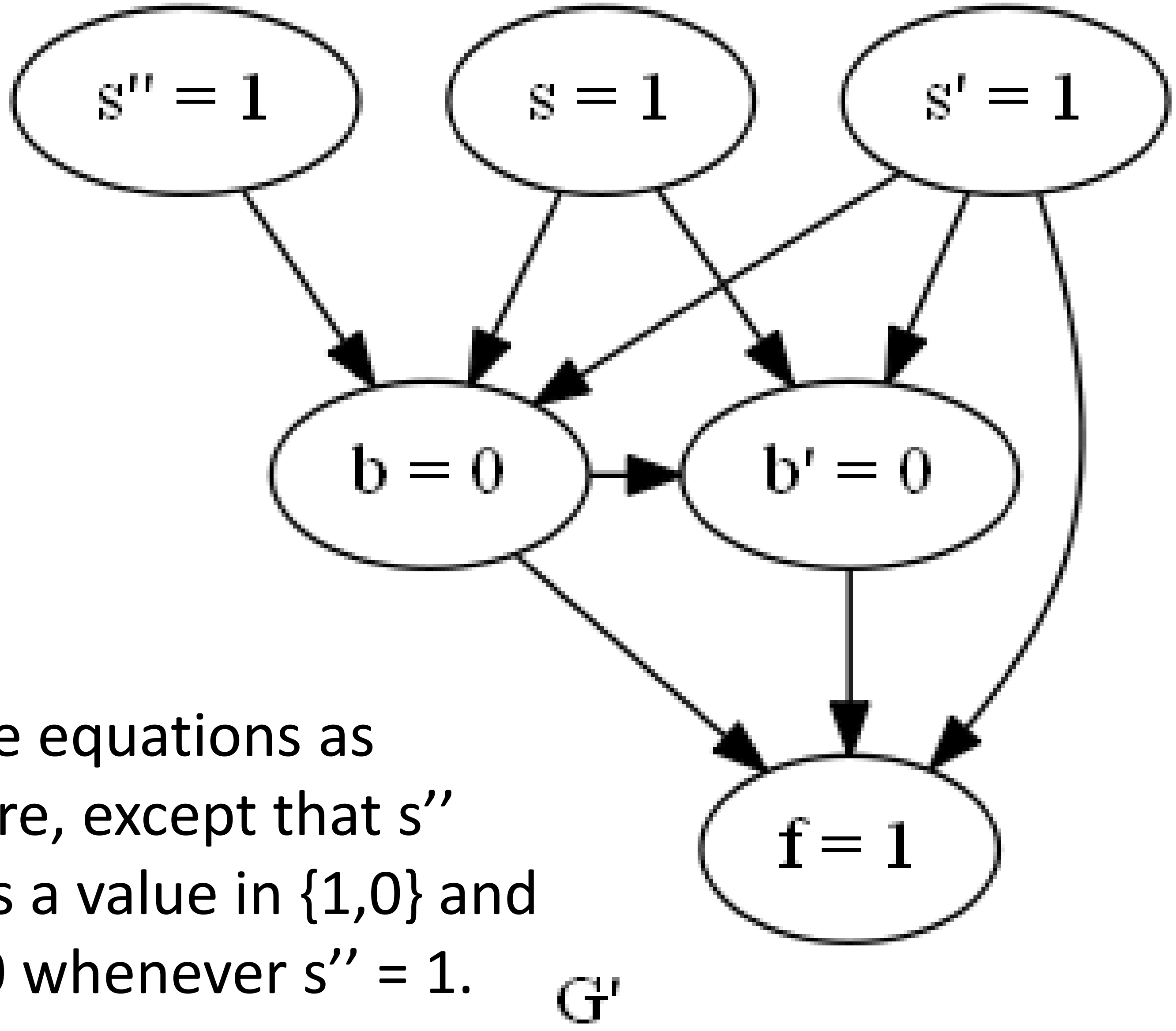
$$f(s', b, b') = s' + b + b'$$

# Example – flood in the Netherlands

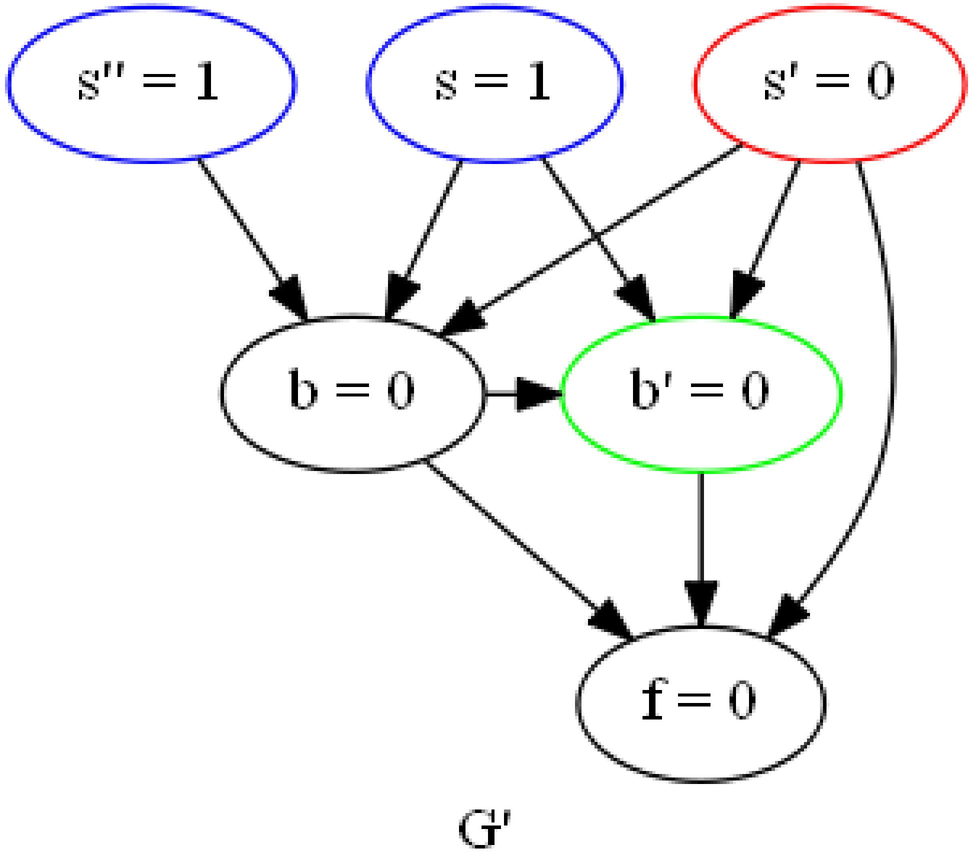
# NESS without and with miracles



# Giving more control to the agent

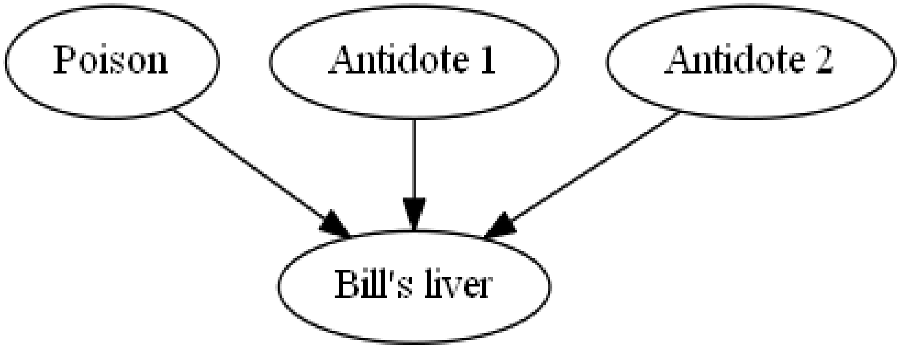


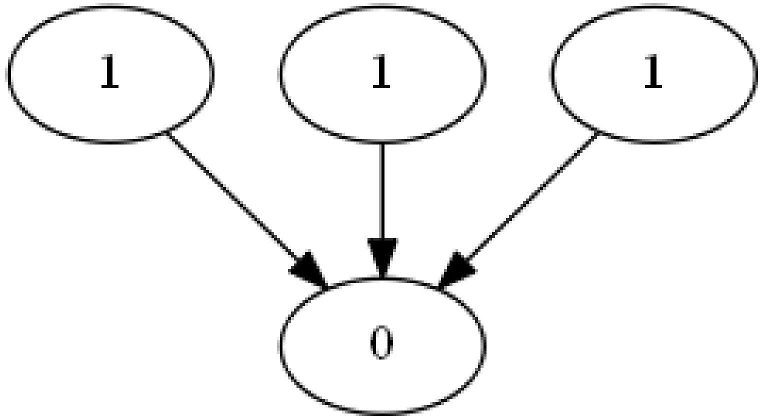
Same equations as before, except that s'' takes a value in {1,0} and b = 0 whenever s'' = 1.



- The NESS correlates of f = 1 from the previous slide are s’’, s’, s and b.
- b’ is **not** a NESS correlate of f = 1.
- Pruning is illustrated on the left: s’’, s and b are all removed as spurious.
- **Verdict:** s’ is the only cause of f = 1.

# The fixed-point analysis





# Why NESS is not right

- A dose of poison is deadly, but so are two doses of antidote. Moreover, two doses of antidote are ineffective against the poison, so the poison contributes along with the antidotes in the actual state. But it is not a NESS cause of Bill’s liver failing...
-
- Makes NESS useless also for attributing responsibility in simple games, like public goods games.

# The generalised but-for test

- A is a cause of B, if and only if, B would not have happened **as it did**, but for A.

- Not circular, but (arguably) too permissive.

- **Solution:** restrict set of permissible counterfactuals by a fixed-point construction.
  - **Only allow changes to source variables that are themselves regarded as making causal contributions.**