Teleological Inference with Structural Causal Models

Dario Compagno, Fabio Massimo Zennaro

University of Bergen

May 30, 2025

Disclaimer

This is very much **experimental work-in-progress**!

i.e.: it might not hold together! O_o

Hence:

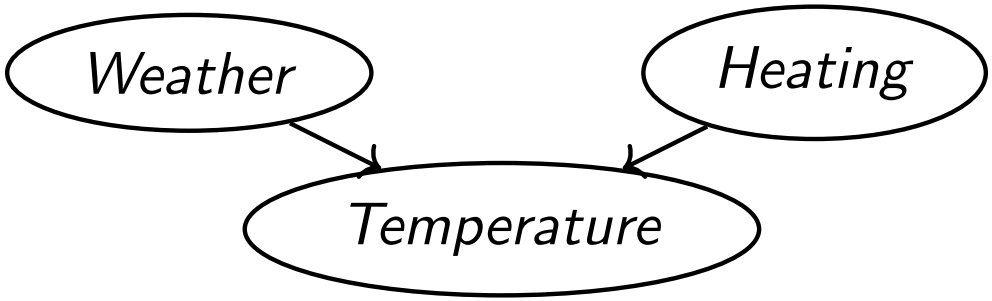
- Do not trust my statements!
- Question my modelling!
- We really appreciate feedback.

1. Motivation

Causal Modelling

We have *established frameworks* for modelling causal systems.

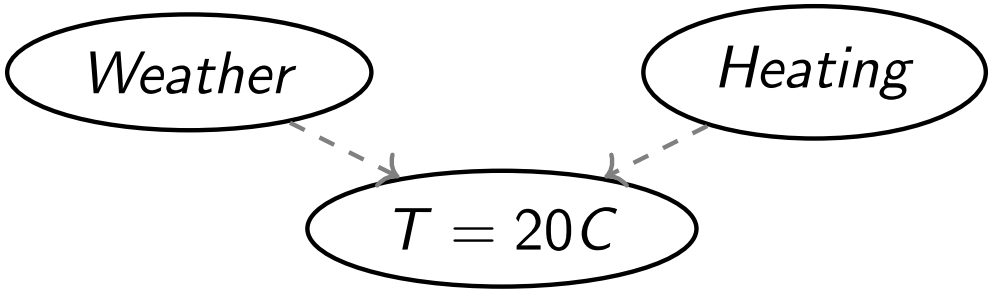
Structural causal models (SCM) join probabilistic and graphical formalism to represent causal systems [12].



We have tools to reason about *observations*, *interventions* and *counterfactuals*.

Interventions

Interventions are well-defined operators.



There are many results on how:

- Evaluate effects of intervention
- Use intervention for causal discovery
- Learn optimal interventions
- Detect interventions

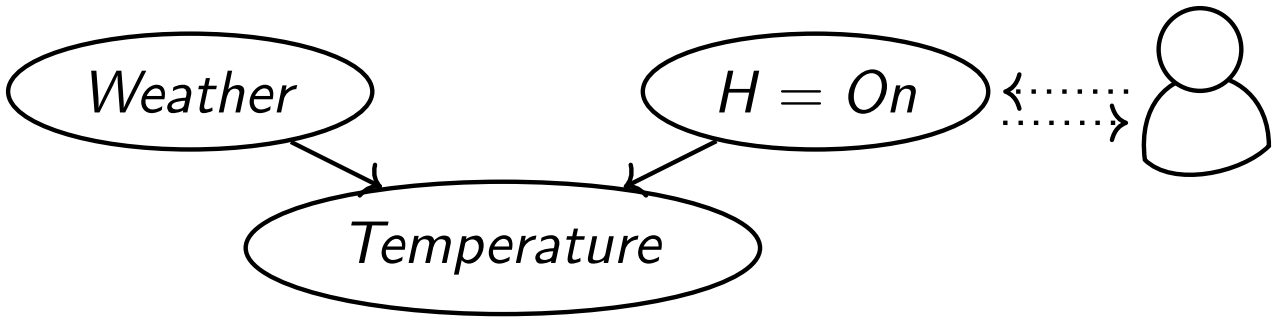
Interventions

Still interventions are external operators not interacting with an SCM.

That is:

- Interventions have *no grounding* in the SCM
- They are treated as *completely arbitrary*

Although, we know that agents often perform interventions **in light of** their knowledge of the system.



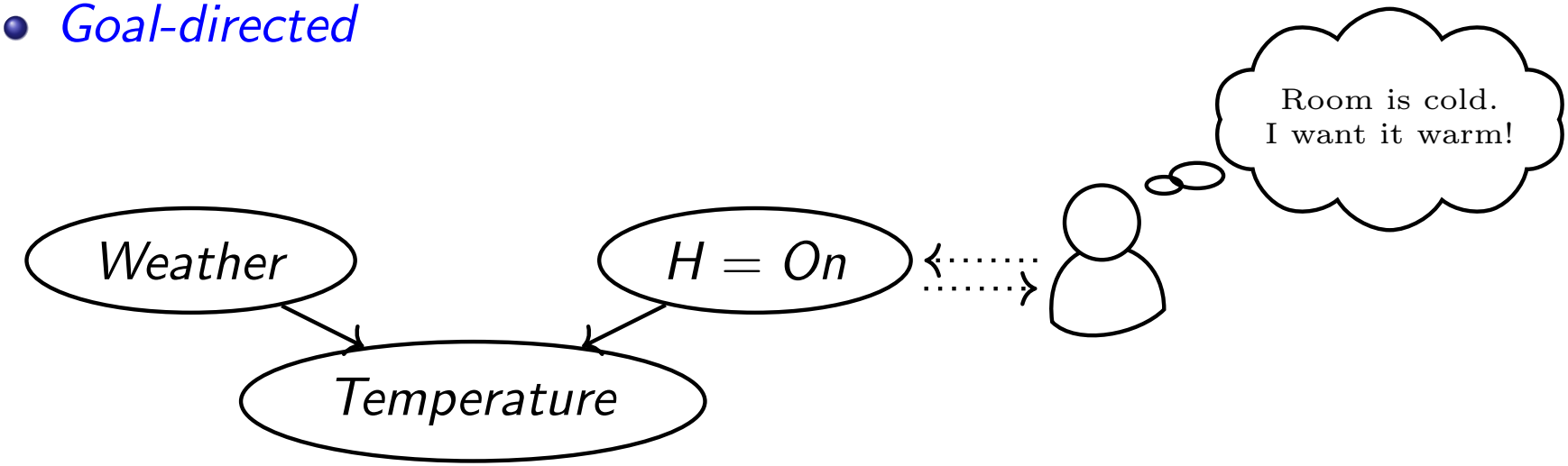
Intentions and Agents

There is no intervention without intention *(Anscombe) [2]*

That is, agents' interventions are *not detached* from the model they are intervening on.

E.g.: intentions makes interventions

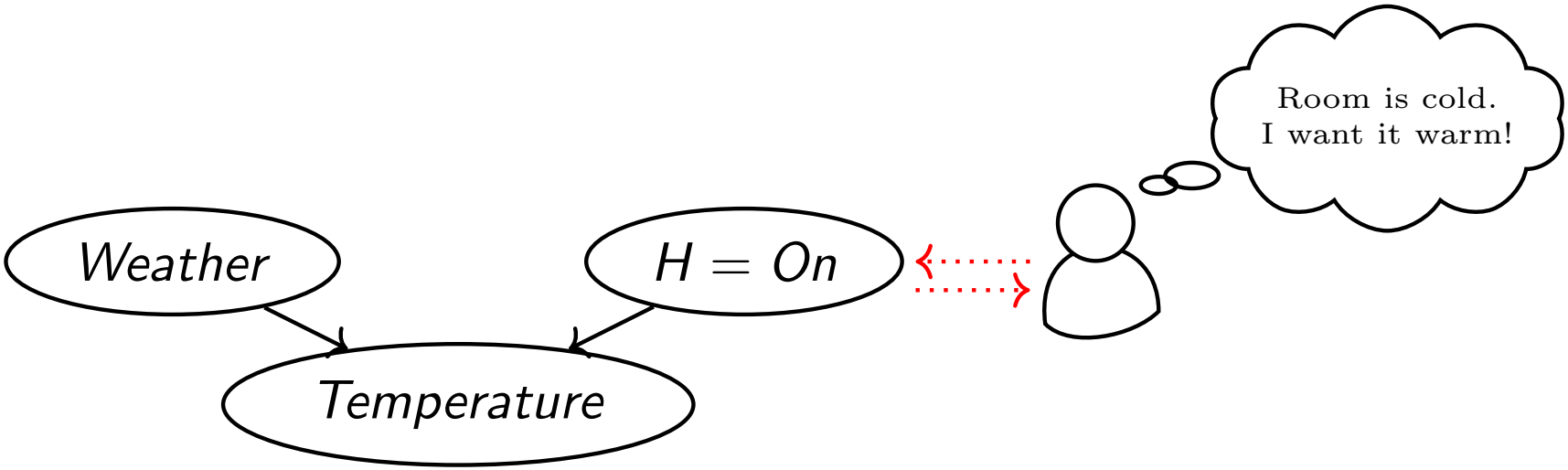
- *Contextual*
- *Goal-directed*



Intentions as objects of study

We want intentions to be a **formal object of study**.

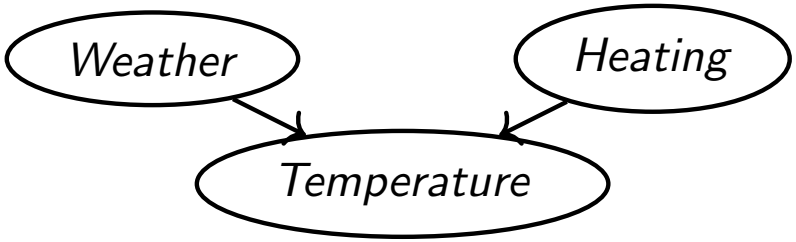
How interventions relate and depend on the causal dynamics of an SCM?



Teleological Reasoning

If the intervention and the causal model (i.e.: the data-generating mechanism) are connected, we could perform **teleological reasoning**:

How is the presence of an agent reflected and inferrable from data?



Has an agent acted on any variable because of an intention?

How do we discriminate between possible intentions from data?

Is an agent's intention in smoking to get pleasure or to get cancer?

Desiderata

We argue that the following are **desirable properties** for a model that would support teleological reasoning:

(D1) *Causal semantics*

(D2) *State-awareness*

(D3) *Goal-directedness*

(D4) *Detection of intentions*

(D5) *Discrimination of intentions*

(D6) *Causal syntax*

- (D6A) *Acyclicity*
- (D6B) *Time-agnosticity*

(D1) Causal semantics

There exists:

1. A *pre-intervention* model that follows the standard causal semantics of SCMs
2. A *post-intentional-intervention* model that is generated from an intentional intervention.

The two models are not **indpendent/ad-hoc** but they are **related** as in the case of standard interventions.

Teleological reasoning happens *across* models!

Previous approaches

There are previous approaches to deal with intentions:

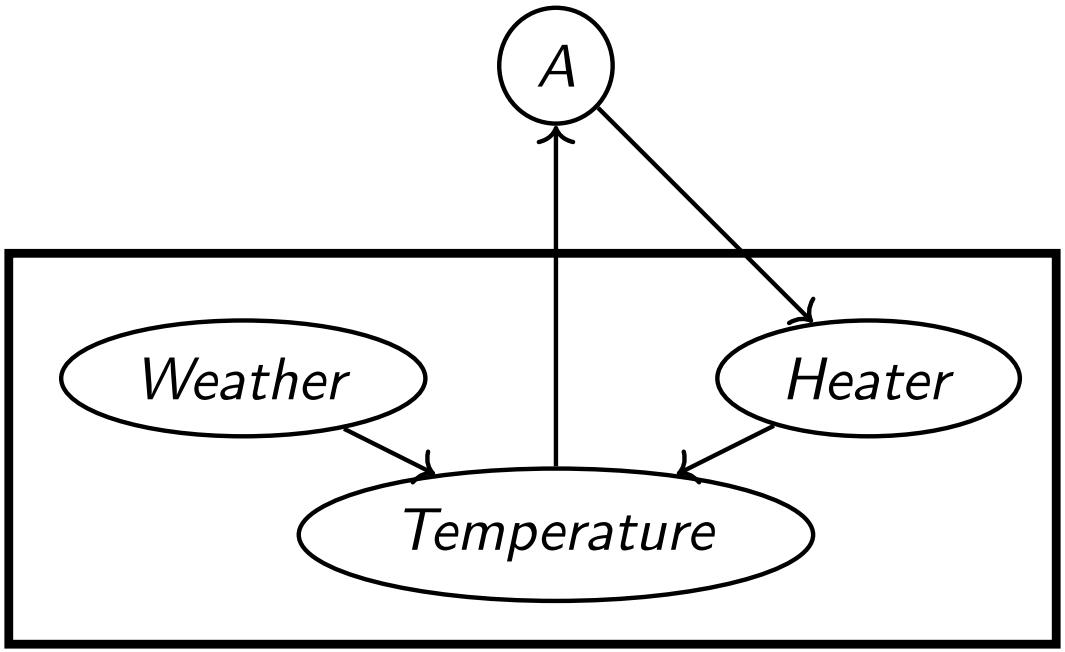
- **Logics**
  - Modal/temporal logic [6]
- **Decision theory**
  - Utility/preference elicitation [9]
- **Reinforcement Learning**
  - Inverse RL [1]
  - RL from Human Feedback [5]
- **Structural Causal Modelling**
  - Intentions in exogenous variables
  - Intentions in endogenous variables
  - Intention in time dimension (dynamic treatments) [11]
  - Extended SCM formalisms [10, 14]

SCM approaches fail on (D1) (and possibly on others)

(D1) Failure

Representing agent *in* the system:

- Violates the assumptions of **closed isolated systems**
- Violates **layering of abstractions**
- Requires **ad-hoc mechanistic modelling**
- Might introduce **cycles**



Intentional Intervention

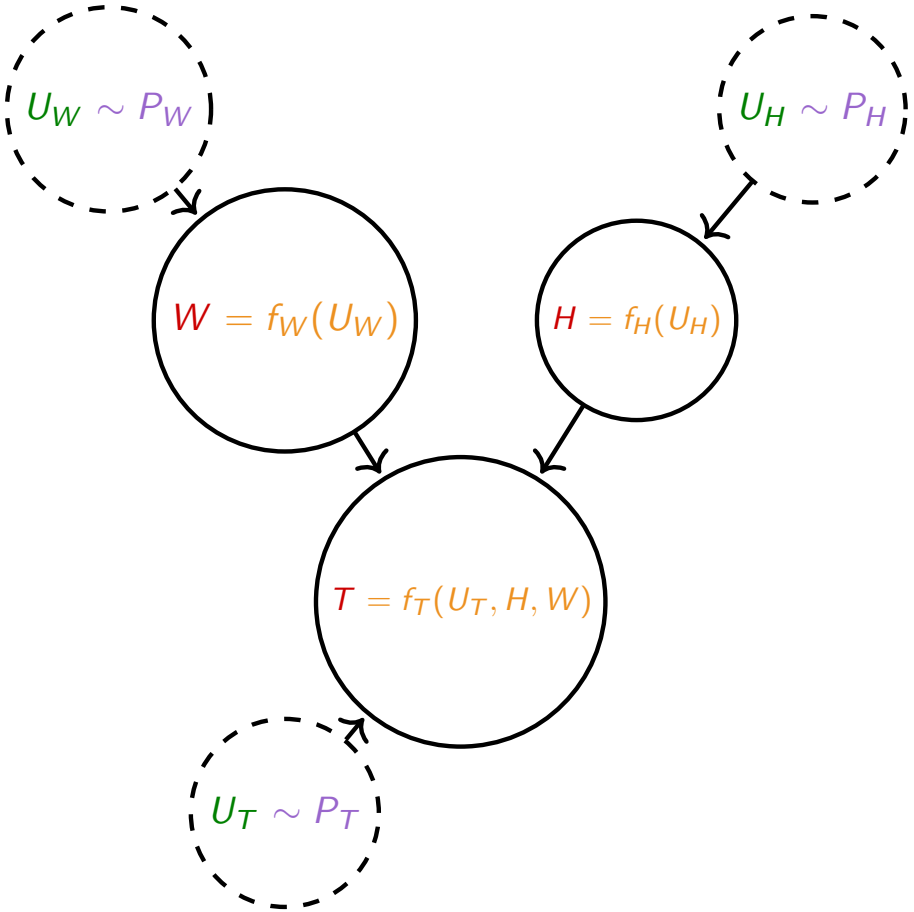
We want to model intentions as **operators** akin to standard intervention, generating and relating different SCMs.

Enter **intentional interventions**.

2. Structural Causal Modelling

Structural Causal Models (SCMs) - Definition

We express a **SCM** as $\mathcal{M} = \langle \mathcal{X}, \mathcal{U}, \mathcal{F}, \mathcal{P} \rangle$ [12, 13]:

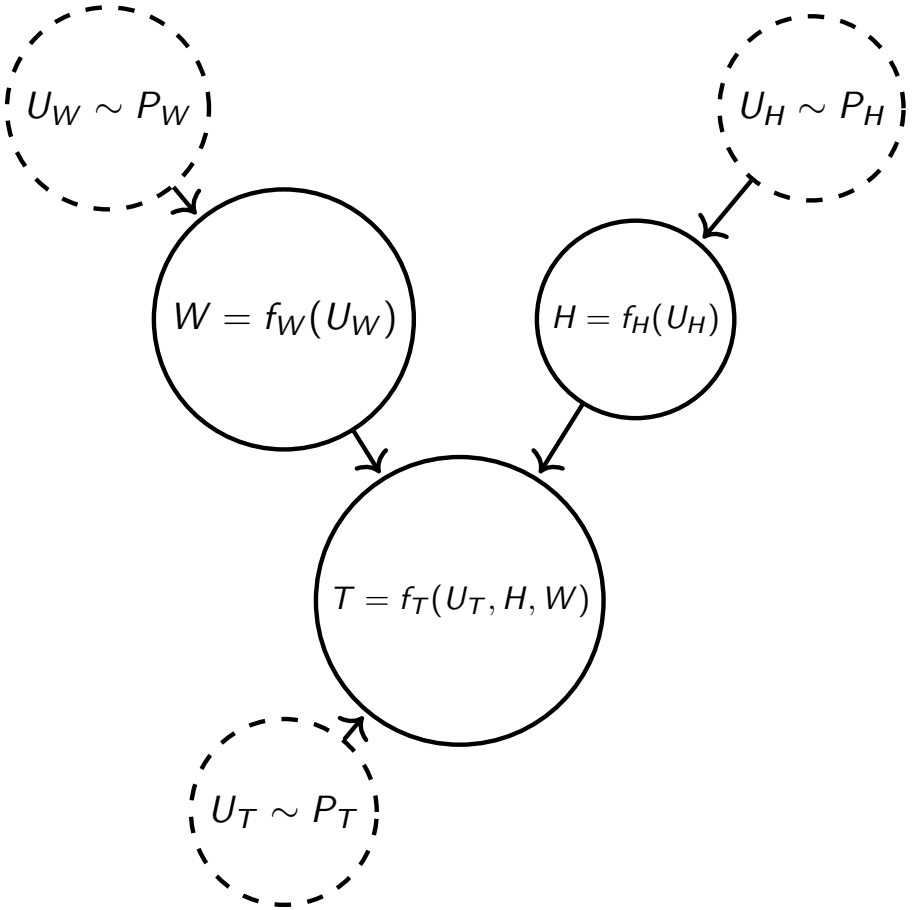


- $\mathcal{X}$: set of *endogenous nodes* (W, T, H) representing variables of interest
- $\mathcal{U}$: Set of *exogenous nodes* (U_W, U_T, U_H) representing stochastic factors
- $\mathcal{F}$: Set of *structural functions* (f_W, f_T, f_H) describing the dynamics of each variable
- $\mathcal{P}$: Set of *distributions* (P_W, P_T, P_H) describing the random factors

Every SCM $\mathcal{M}$ implies a (joint) **distribution** $P_{\mathcal{M}}$: $P_{\mathcal{M}}(W, T, H)$

Structural Causal Models (SCMs) - Interventions

We can perform **interventions** on a causal model [12, 13]:



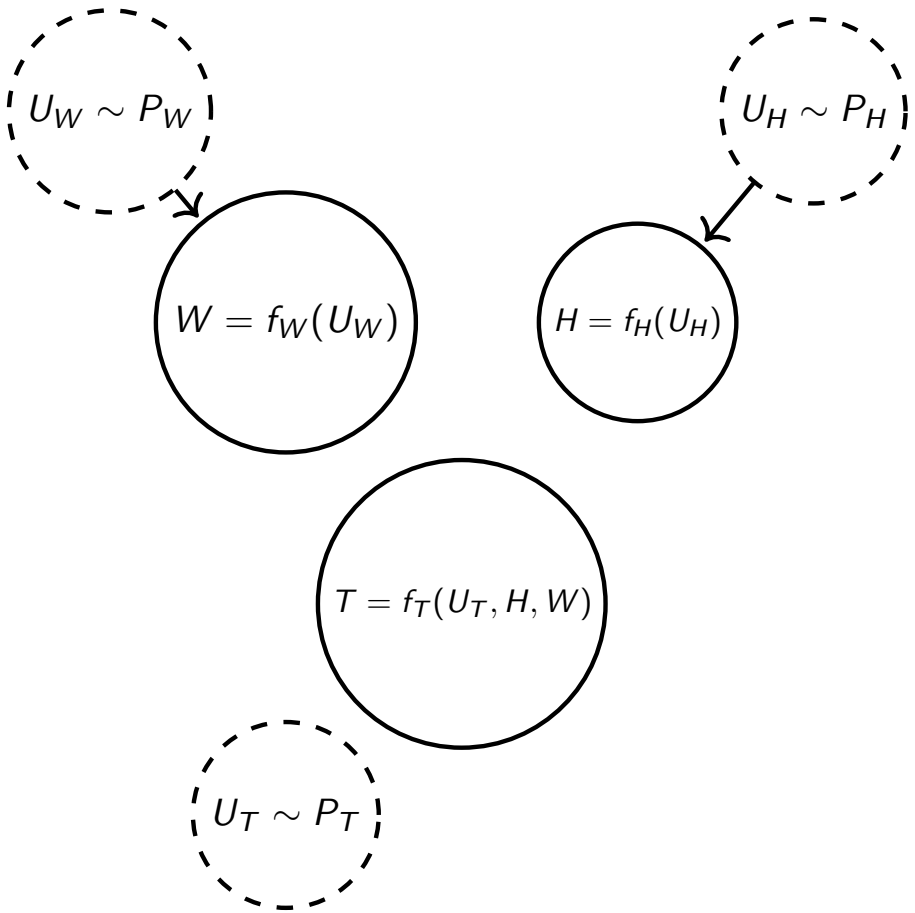
$do(T = 1)$

1.

2.

Structural Causal Models (SCMs) - Interventions

We can perform **interventions** on a causal model [12, 13]:

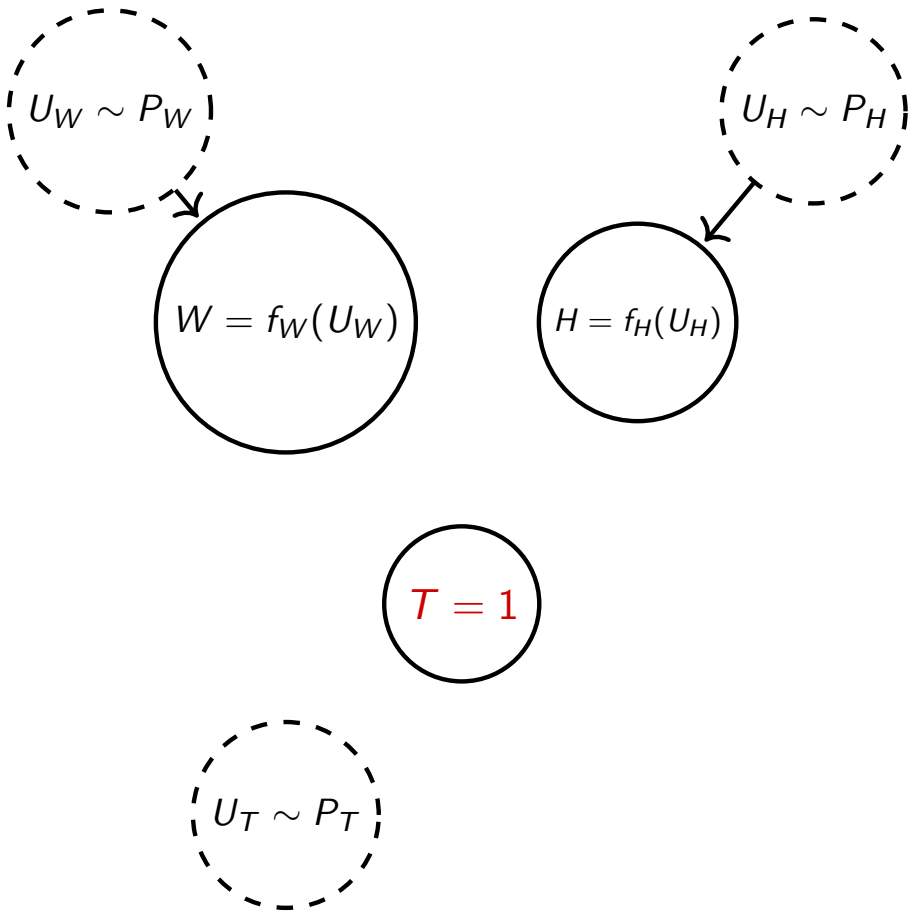


$do(T = 1)$

1. Remove incoming edges in the intervened node
2.

Structural Causal Models (SCMs) - Interventions

We can perform **interventions** on a causal model [12, 13]:

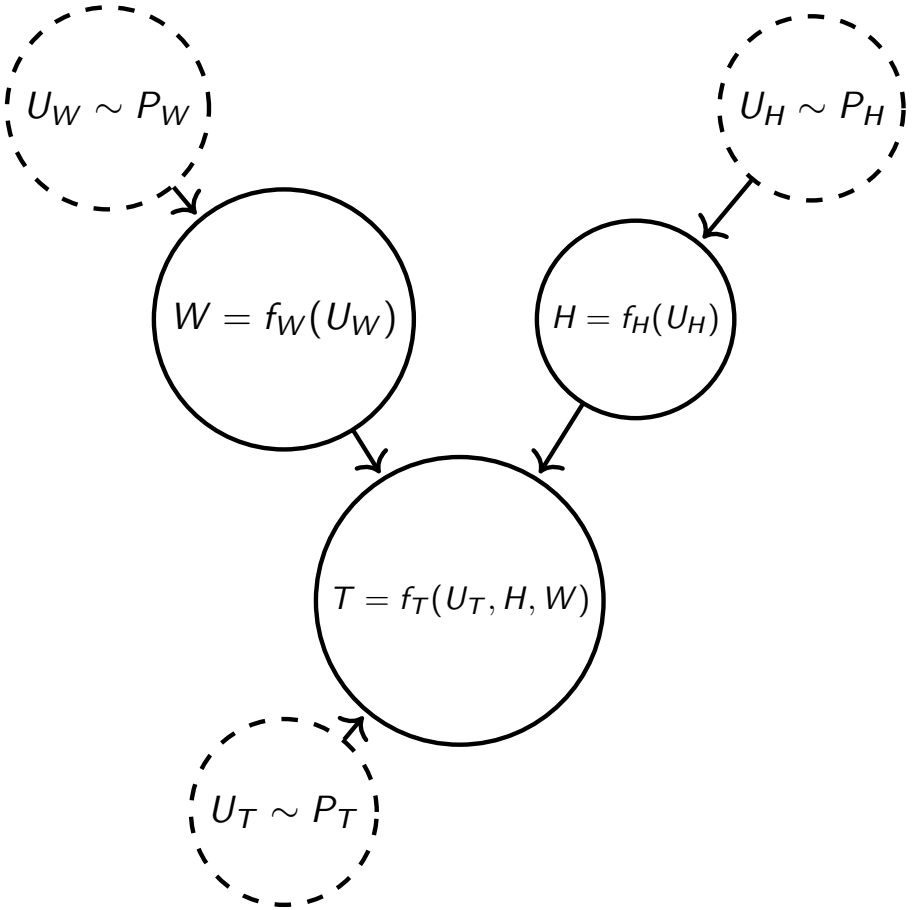


$do(T = 1)$

1. Remove incoming edges in the intervened node
2. Set the value of the intervened node

Counterfactuals

We can compute **counterfactuals** from a causal model [12, 13]:

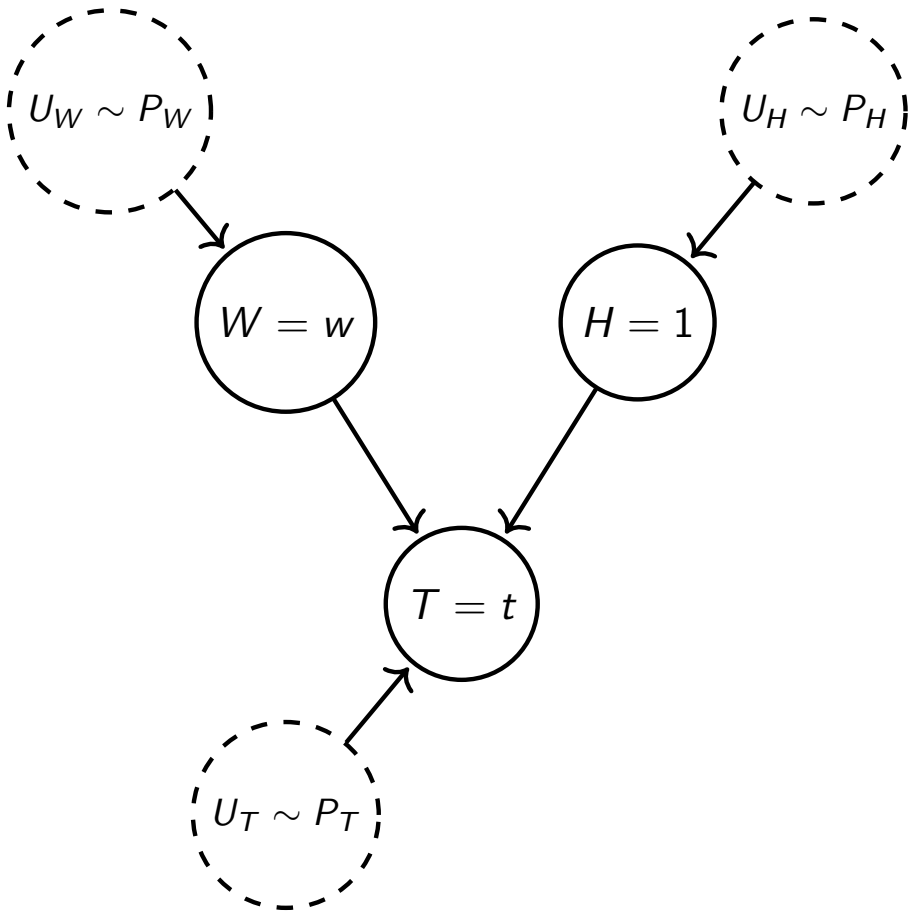


$$P\left(T_{do(H=0)} | H = 1\right)$$

1.
2.
3.
4.

Counterfactuals

We can compute **counterfactuals** from a causal model [12, 13]:

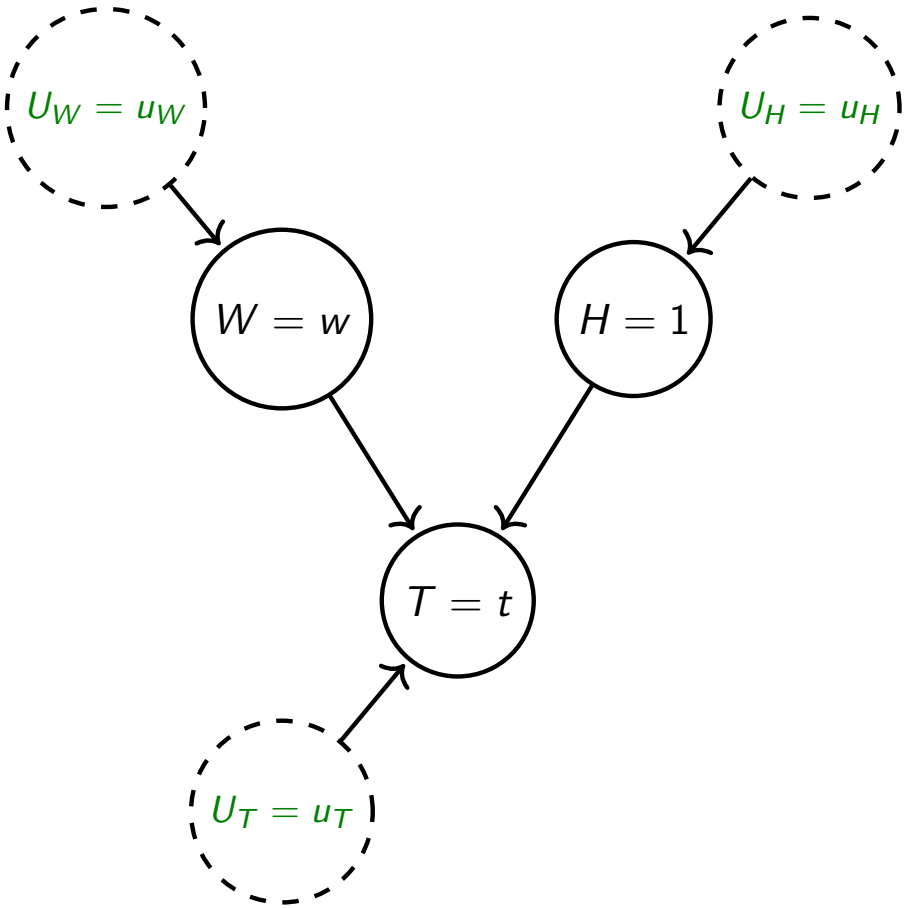


$$P\left(T_{do(H=0)} | H = 1\right)$$

1. *Observe* the real-world system
2.
3.
4.

Counterfactuals

We can compute **counterfactuals** from a causal model [12, 13]:

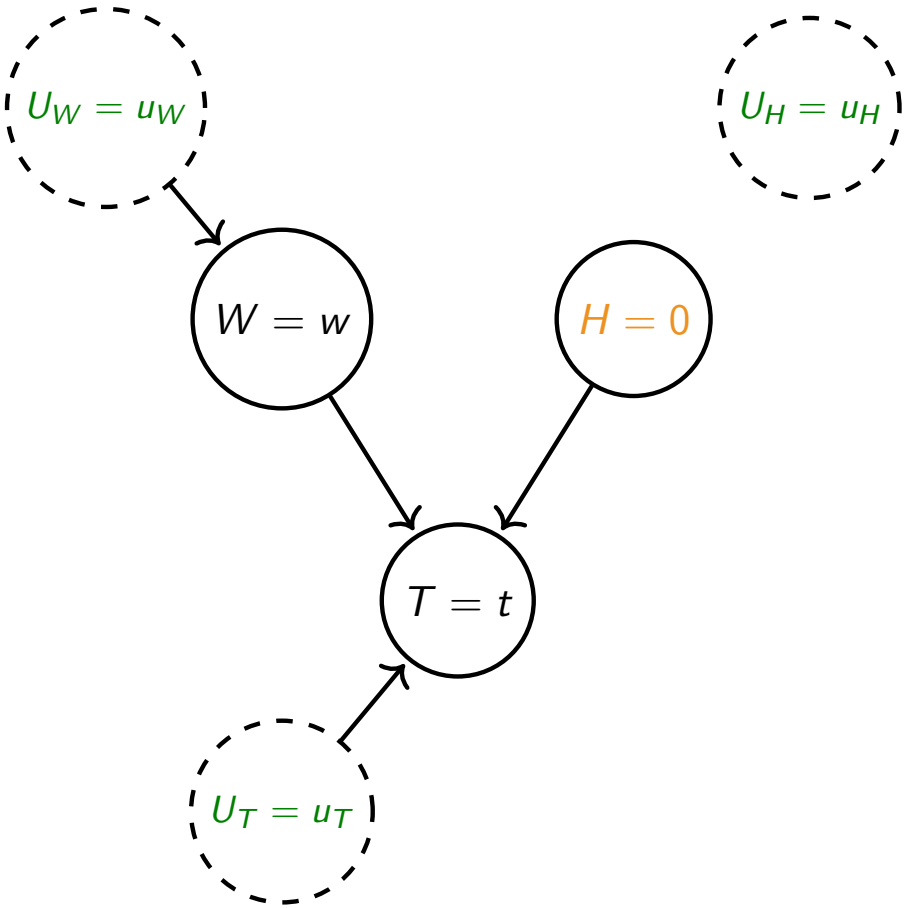


$$P\left(T_{do(H=0)} | H = 1\right)$$

1. Observe the real-world system
2. Abduce the exogenous values
3.
4.

Counterfactuals

We can compute **counterfactuals** from a causal model [12, 13]:

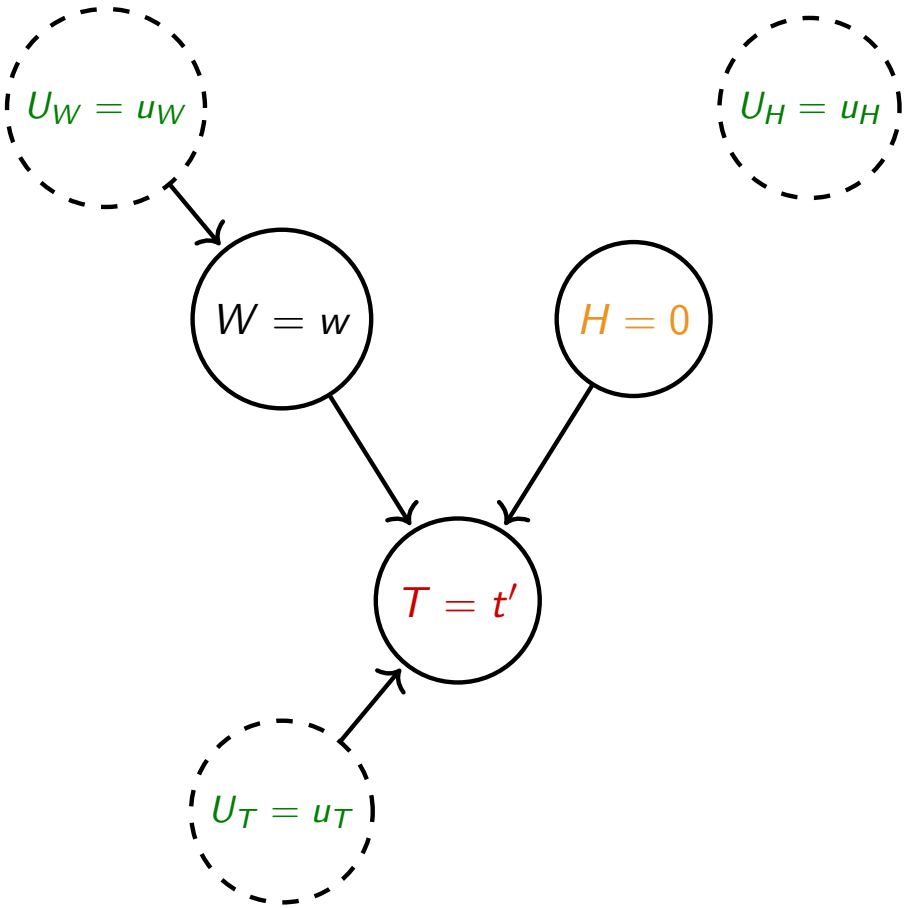


$$P\left(T_{do(H=0)} | H = 1\right)$$

1. Observe the real-world system
2. Abduce the exogenous values
3. Perform the counterfactual intervention
4.

Counterfactuals

We can compute **counterfactuals** from a causal model [12, 13]:

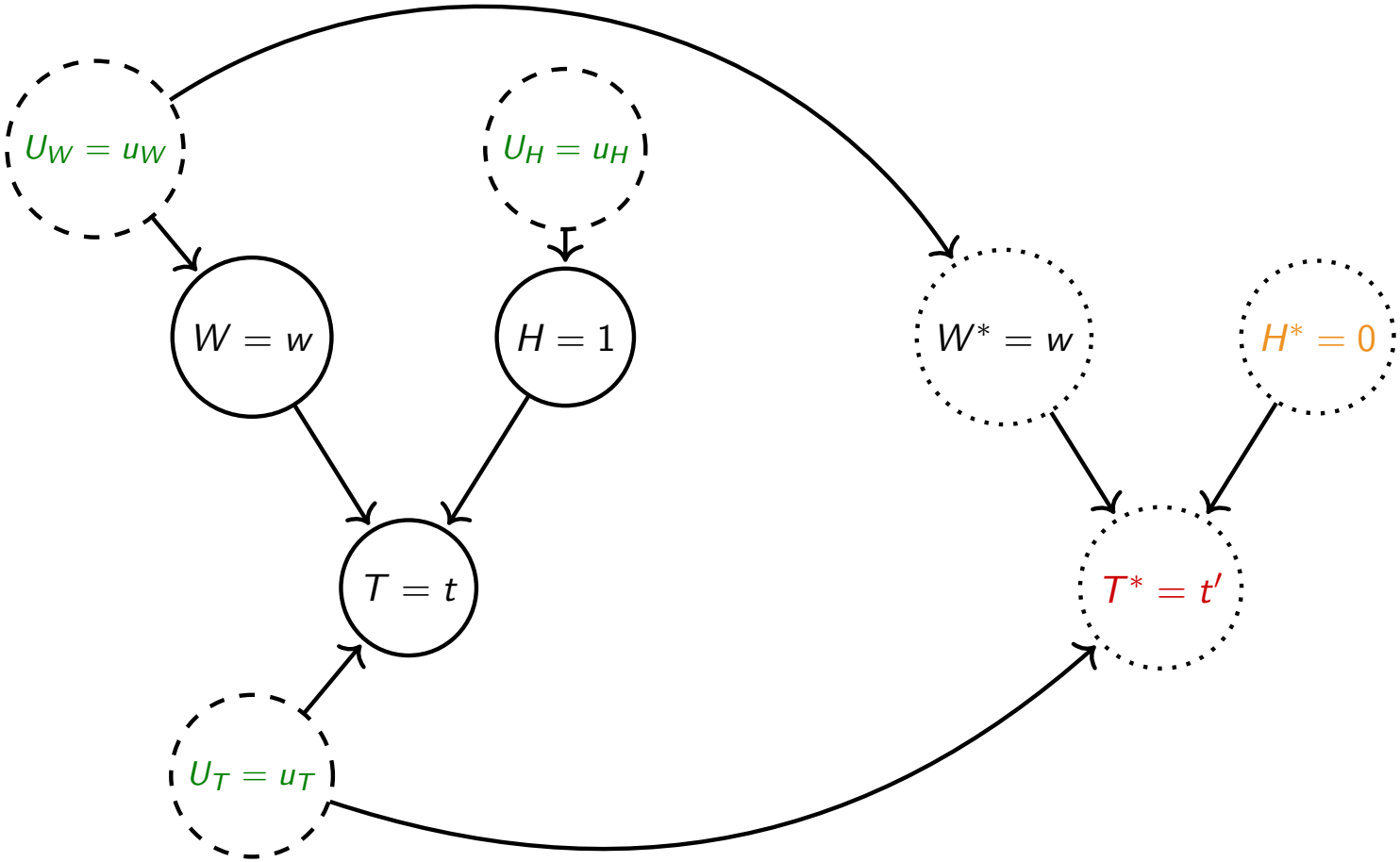


$$P\left(T_{do(H=0)} | H = 1\right)$$

1. Observe the real-world system
2. Abduce the exogenous values
3. Perform the counterfactual intervention
4. Observe the counterfactual outcome

Twin Models

We can study counterfactuals via **twin-worlds** [3, 4]:



connecting a *factual world* (left) to a *counterfactual world* (right)

3. Intentional Interventions

Intentional Intervention

We want to model intentions as **operators** akin to standard intervention, generating and relating different SCMs.

Enter **intentional interventions**.

Key assumptions

In order to define intentional interventions, we need the following *assumptions*:

- **Perfect knowledge:** no misspecification in agent's SCM;
- **High-frequency agent:** agent's action have higher frequency than exogenous variables.

(These assumptions are *essential* in our modelling.)

Other simplifications

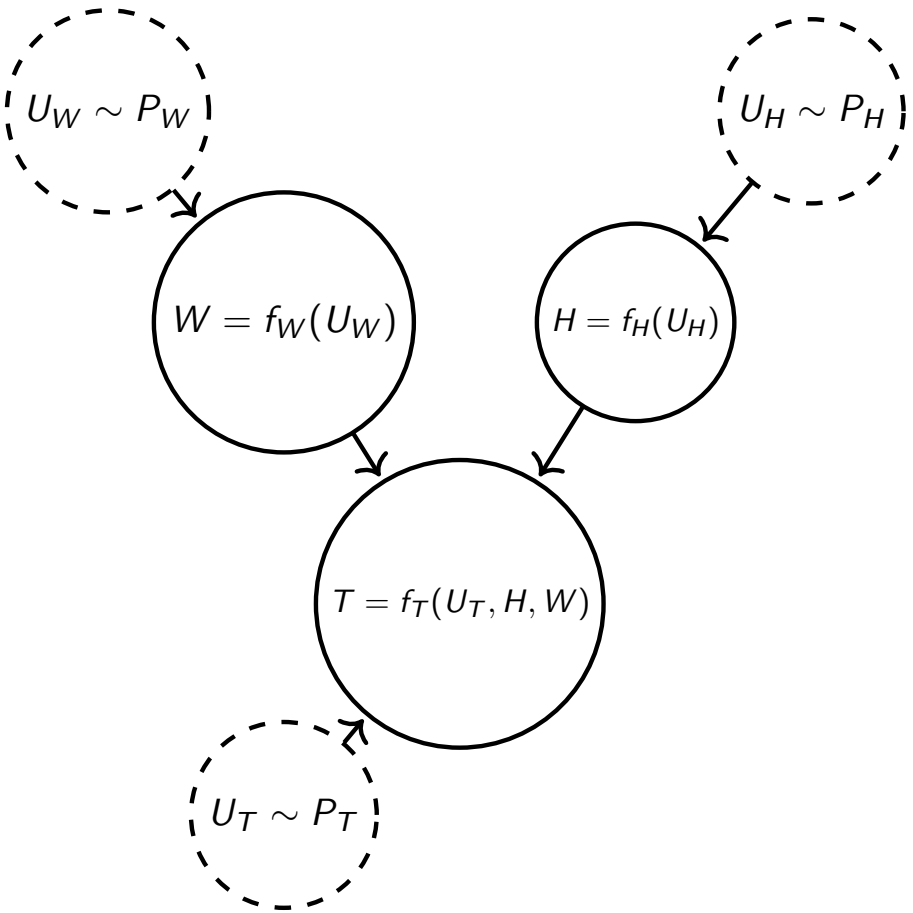
For conceptual ease and clarity, we also adopt the following *simplifications*:

- *Markovianity:* no unobserved confounders;
- *Twin graphs:* other constructions are possible [8];
- *Atomic intentional intervention:* the agent acts on a single variable;
- *Single-variable context:* the agent observes a single variable.

(These simplification are *just for convenience* in our modelling.)

Intentional Intervention

We model an **intentional intervention** on a causal model:

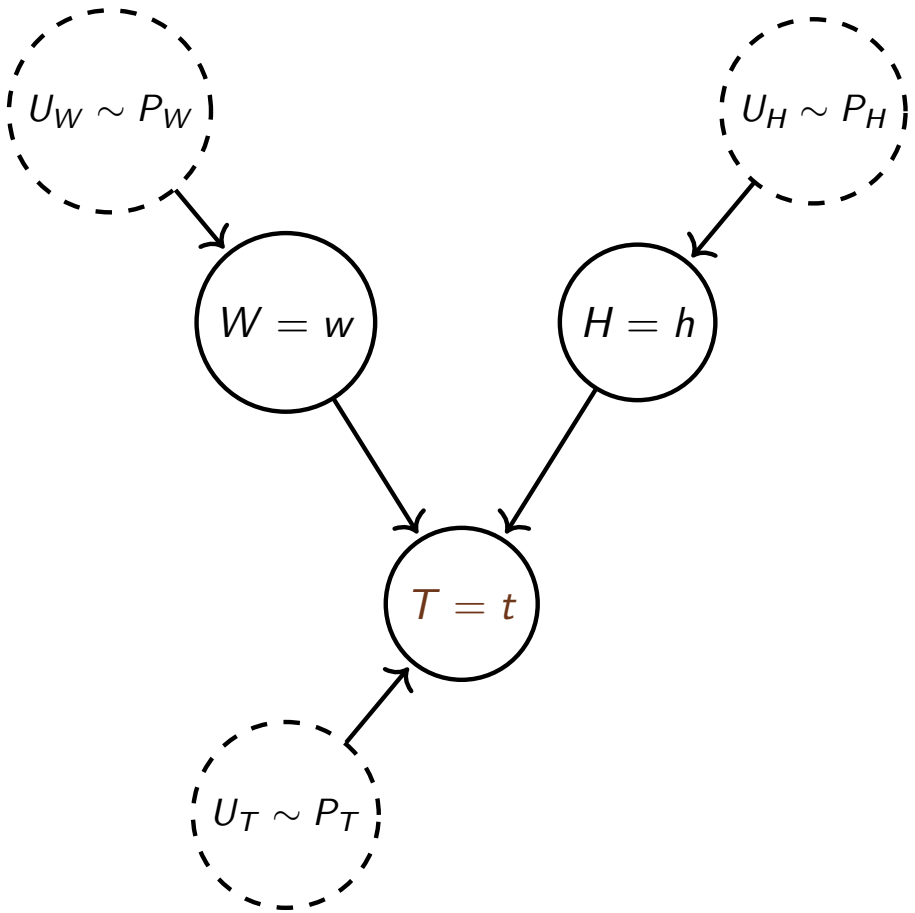


$do^*\,(H = f(t))$

1.
2.
3.
4.

Intentional Intervention

We model an **intentional intervention** on a causal model:

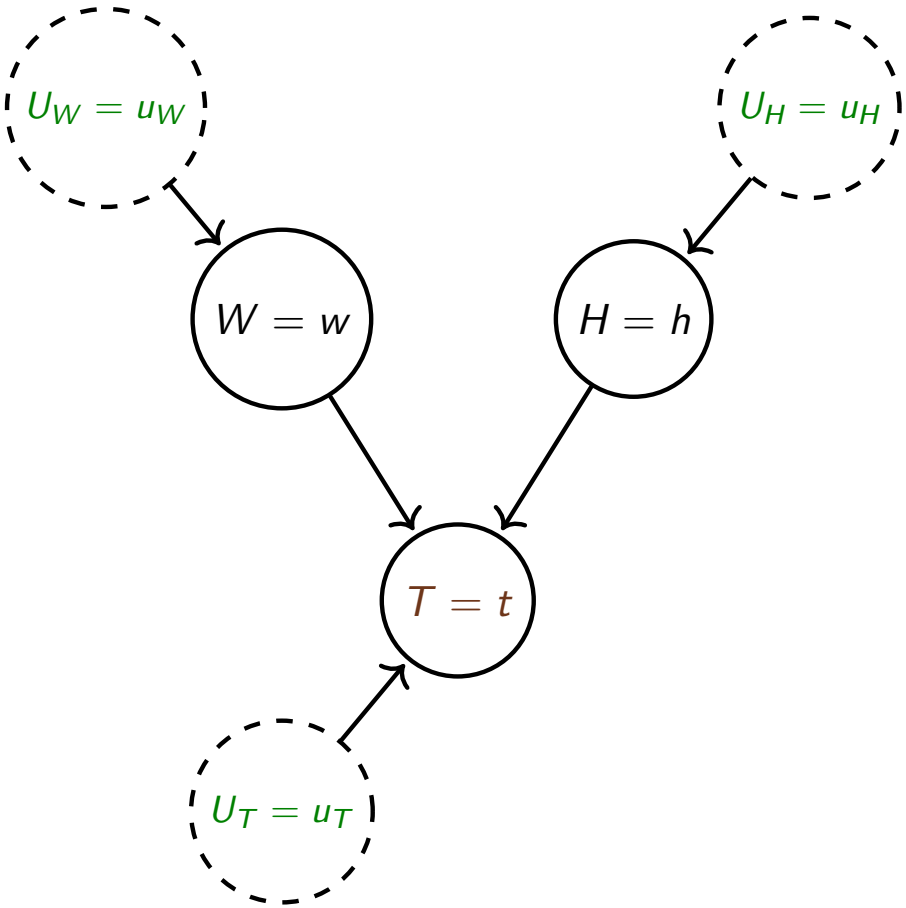


$do^*(H = f(t))$

1. Simulate/observe the system and get the context
2.
3.
4.

Intentional Intervention

We model an **intentional intervention** on a causal model:

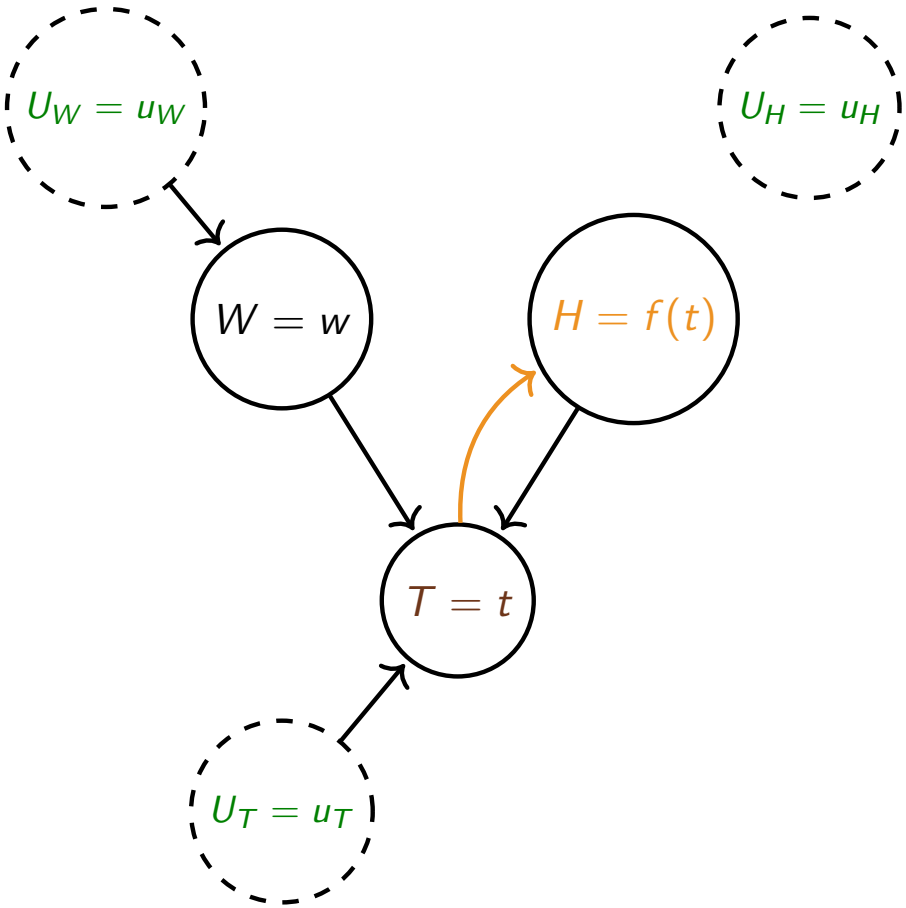


$do^*(H = f(t))$

1. Simulate/observe the system and get the context
2. Abduce the exogenous values
3.
4.

Intentional Intervention

We model an **intentional intervention** on a causal model:

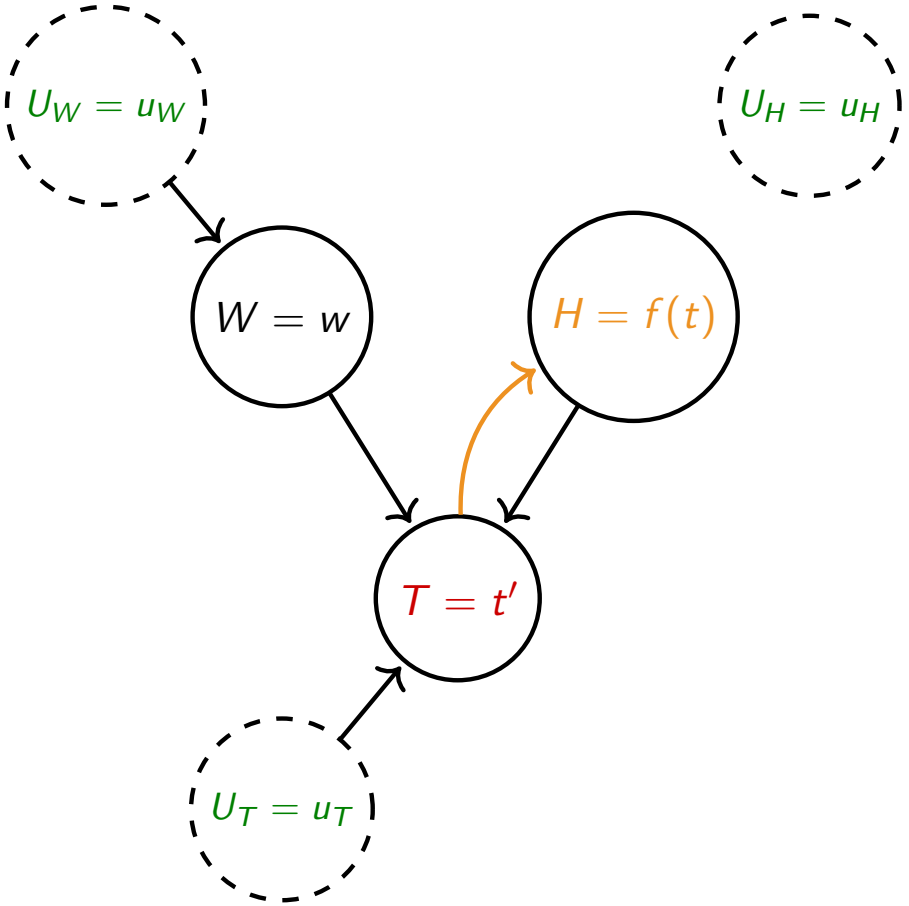


$do^*(H = f(t))$

1. Simulate/observe the system and get the context
2. Abduce the exogenous values
3. Perform the intentional intervention
4.

Intentional Intervention

We model an **intentional intervention** on a causal model:



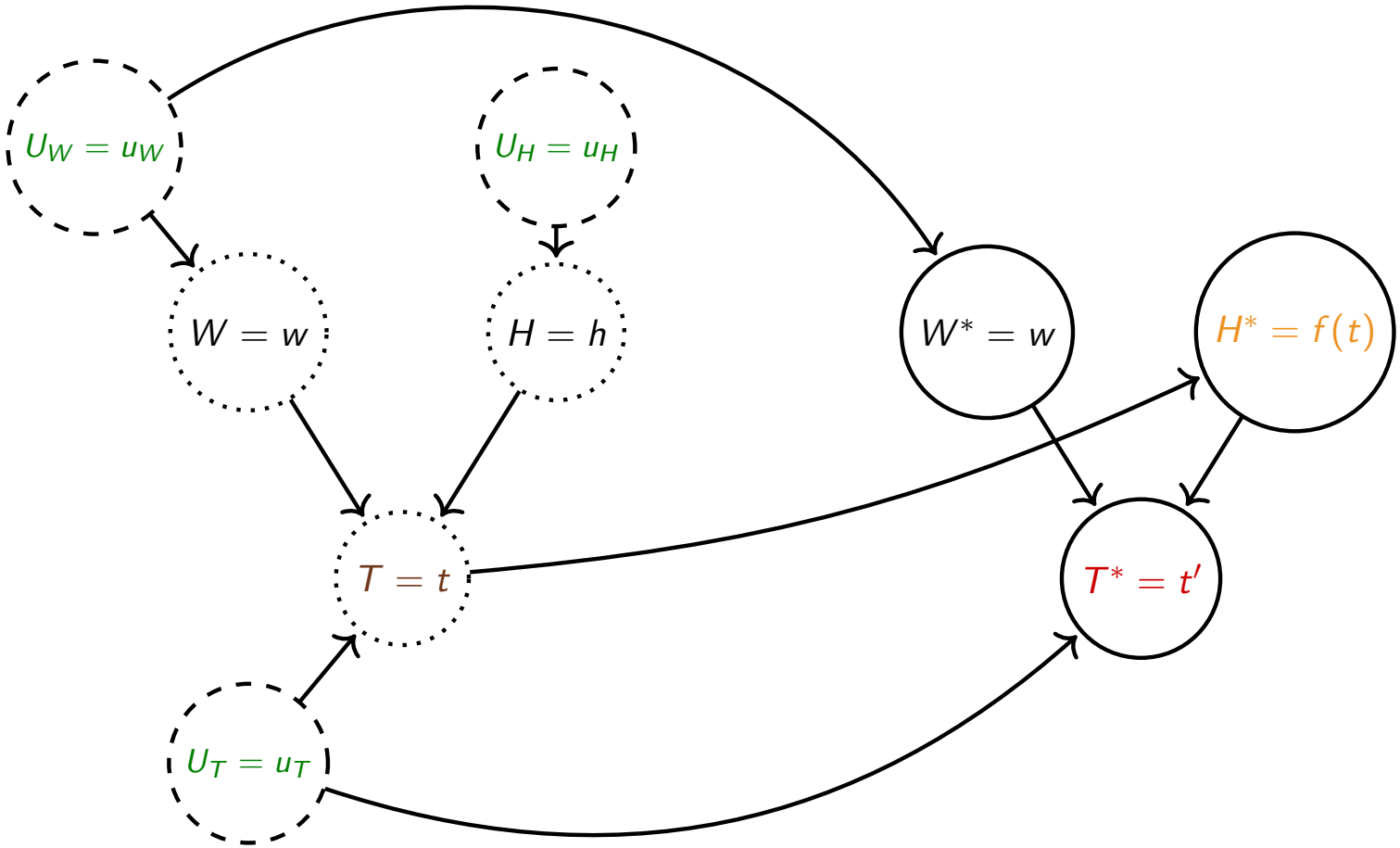
$do^*(H = f(t))$

1. Simulate/observe the system and get the context
2. Abduce the exogenous values
3. Perform the intentional intervention
4. Observe intentional outcome

Do we have a cycle???

Final Models

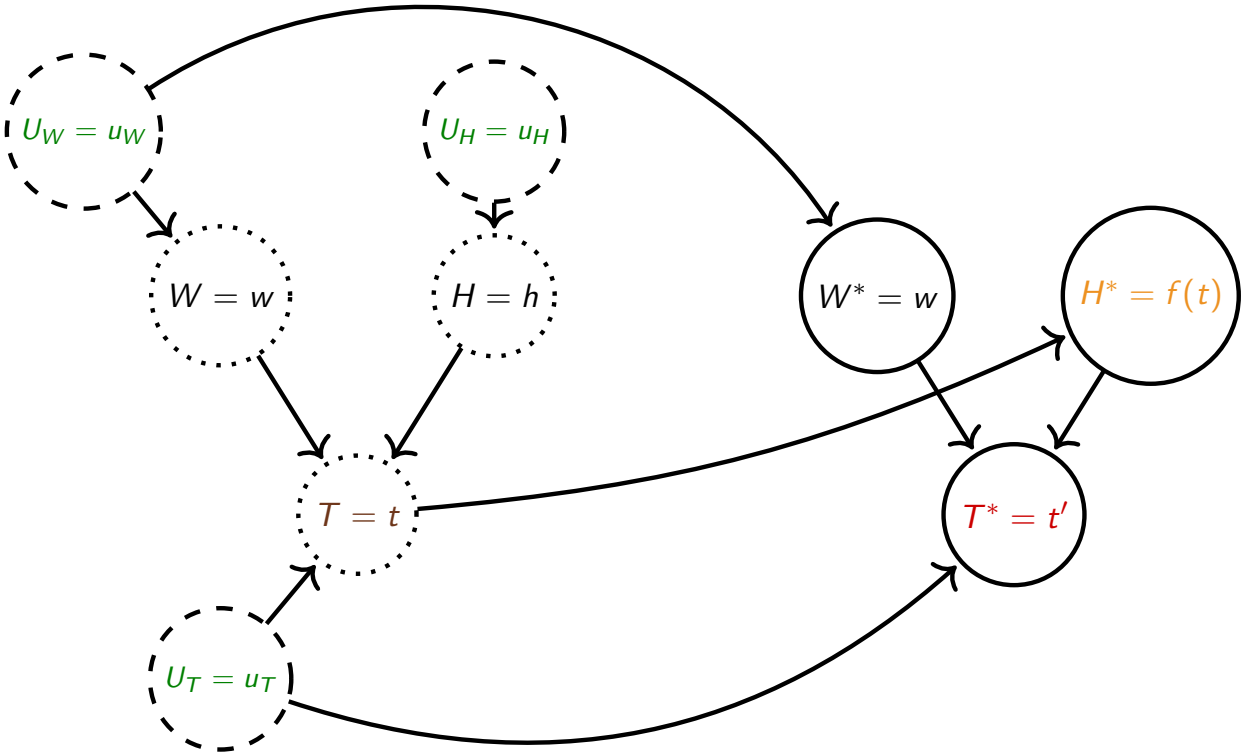
Twinning solves the apparent loop in a **final model** (FM) [7]:



connecting an *imagined world* (left) to a *agent-intervened world* (right)

Final Models

A **FM** is a sort of "*reverse/anticipating counterfactual*":



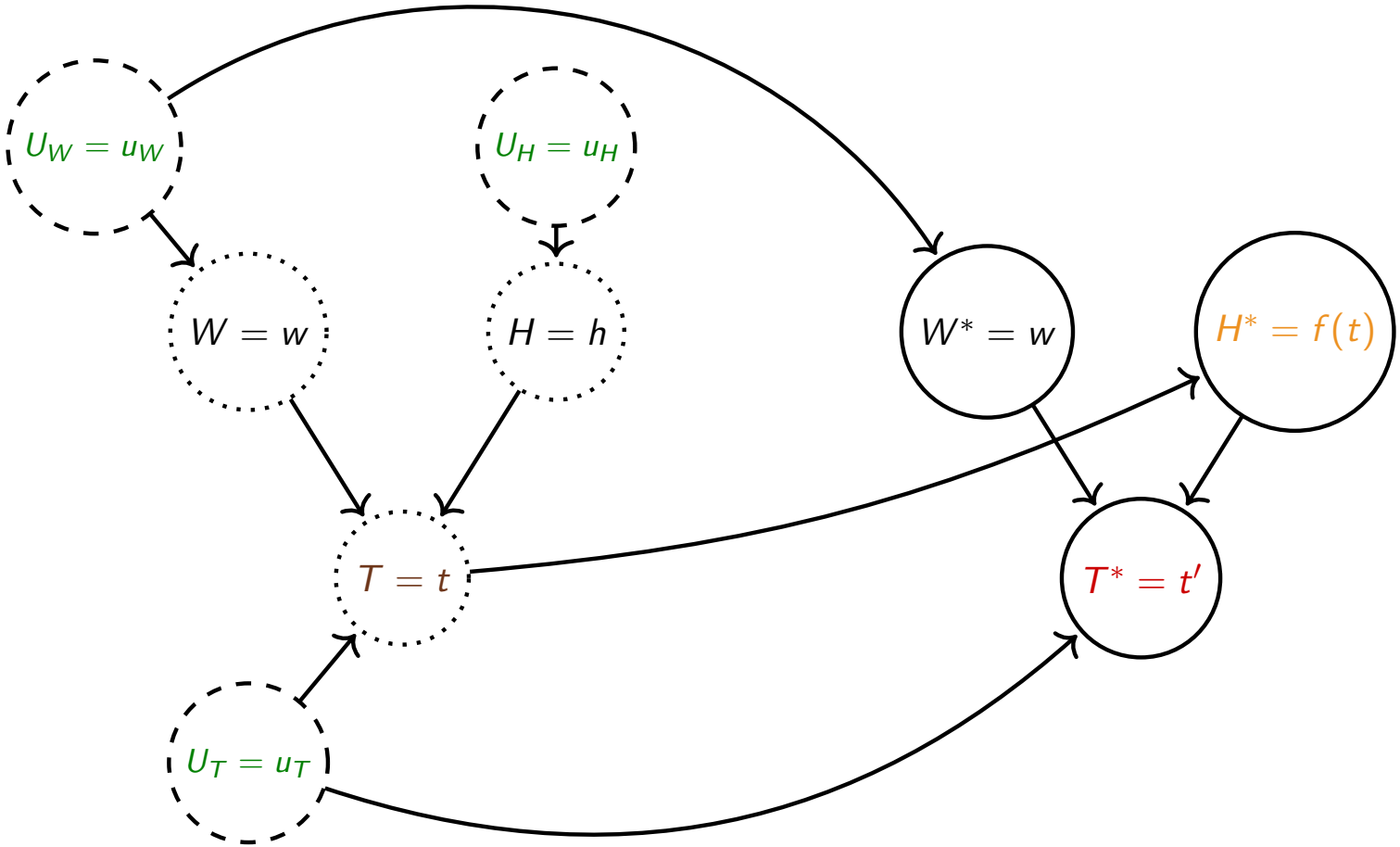
Two interpretations:

- The agent *observes* a state of the world and intentionally modifies it (e.g.: temperature in a room);
- The agent *imagines* an outcome and intentionally correct for it (e.g.: steering a car from a cliff)

4. Teleological Reasoning

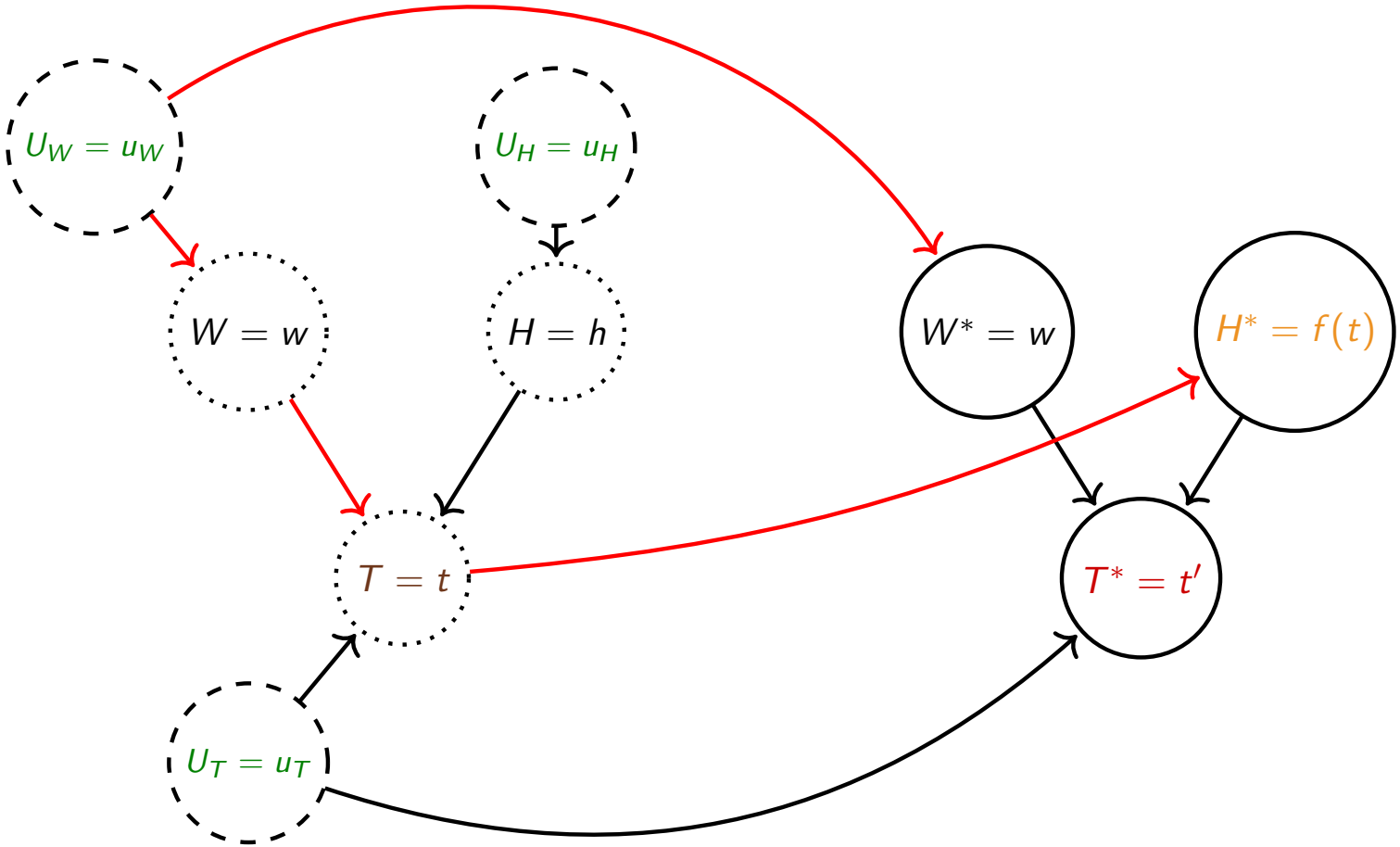
Detection of Intentions

Given a causal DAG $W \rightarrow T \leftarrow H$ and data $\mathcal{D} = \{W_i^*, T_i^*, H_i^*, \}$, is it possible to decide *whether an intentional intervention happened*?



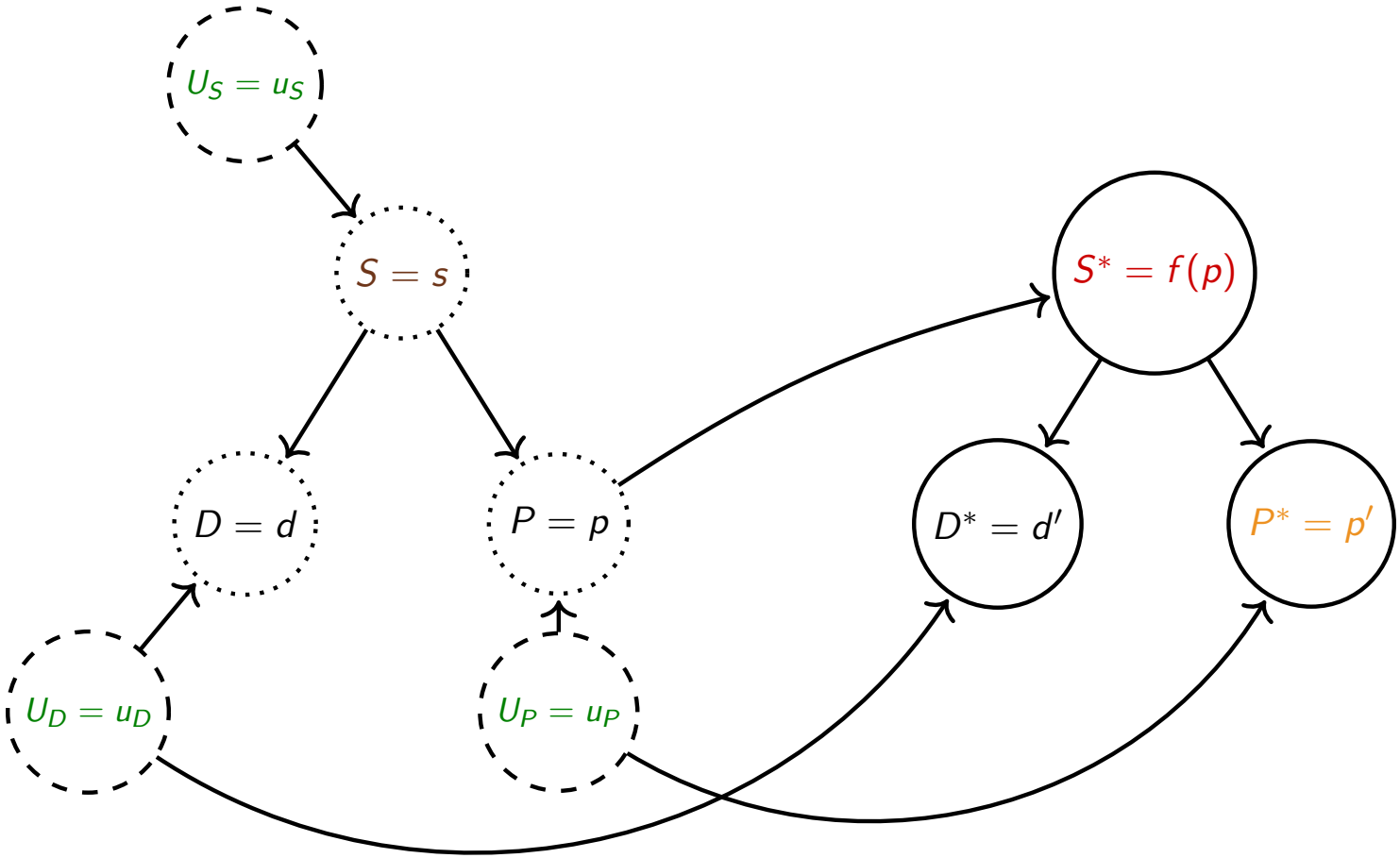
Detection of Intentions

In colliders, an intentional intervention $do^*(H^* = f(t))$ creates a *dependency* $W^* \not\perp H^*$ via a backdoor path.



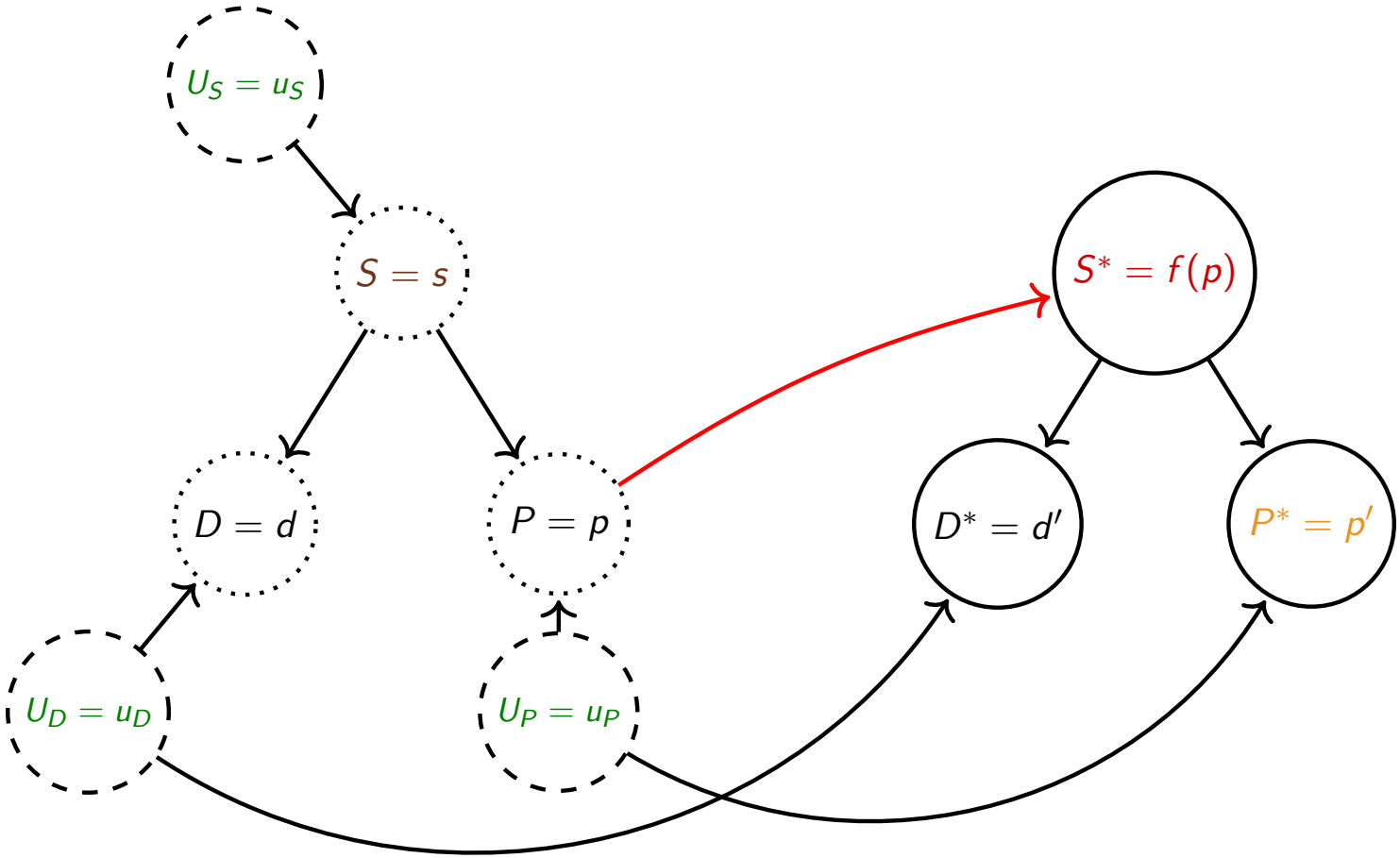
Discrimination of Intentions

Given an SCM and data $\mathcal{D} = \{S_i^*, D_i^*, P_i^*, \}$, is it possible to *identify the argument* x of an agent' intentional intervention $do^*(S^* = f(x))$?



Discrimination of Intentions

In forks, interventions $do(D)$ and $do(P)$ *affect differently the distribution* $P(S^*)$ wrt to the available data.



5. Conclusion

Conclusion

- ✓ FMs are **quasi-counterfactual** (consistently with the *high-frequency assumption*)
- ✓ FMs satisfy **all our desiderata**:
  - (D1) Intentional interventions induce relations among FMs
  - (D2) State-awareness encoded in argument of f
  - (D3) Goal-directedness encoded in form of f
  - (D4) Detection of intentions in teleological reasoning (*preliminary*)
  - (D5) Discrimination of intentions in teleological reasoning (*preliminary*)
  - (D6) Causal syntax respected (no cycles, no time-dimension)

Thanks!

Thank you for listening!

References I

[1] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1, 2004.

[2] G. Elisabeth M. Anscombe. *Intention*. Blackwell, 1957.

[3] Alexander Balke and Judea Pearl. Probabilistic evaluation of counterfactual queries. 1994.

[4] Stephan Bongers, Patrick Forré, Jonas Peters, and Joris M Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915, 2021.

[5] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

References II

[6] Philip R Cohen and Hector J Levesque. Intention is choice with commitment. *Artificial intelligence*, 42(2-3):213–261, 1990.

[7] Dario Compagno. Final models: A finalistic interpretation of statistical correlation. *arXiv preprint arXiv:2310.02272*, 2023.

[8] Juan D Correa and Elias Bareinboim. Counterfactual graphical models: Constraints and inference. In *Forty-second International Conference on Machine Learning*, 2025.

[9] Peter H Farquhar. State of the art—utility assessment methods. *Management science*, 30(11):1283–1300, 1984.

[10] Meir Friedenberg and Joseph Y Halpern. Intents in actions. 2025.

[11] Susan A Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65(2):331–355, 2003.

[12] Judea Pearl. *Causality*. Cambridge University Press, 2009.

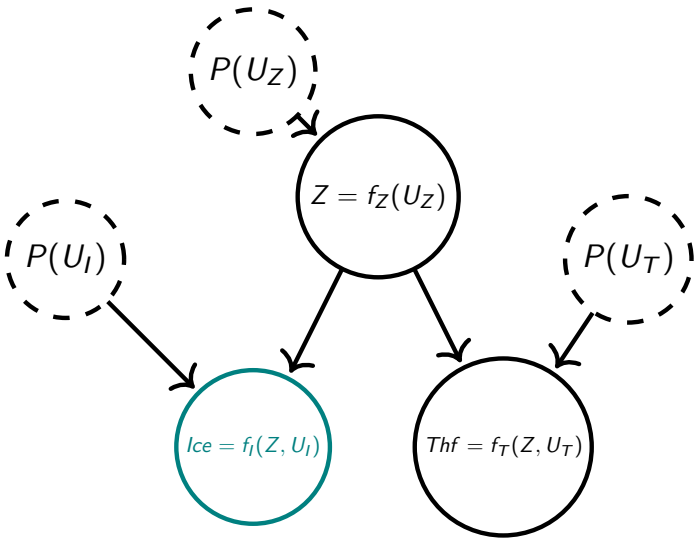
References III

[13] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: Foundations and learning algorithms*. MIT Press, 2017.

[14] Francis Rhys Ward, Matt MacDermott, Francesco Belardinelli, Francesca Toni, and Tom Everitt. The reasons that agents act: Intention and instrumental goals. *arXiv preprint arXiv:2402.07221*, 2024.

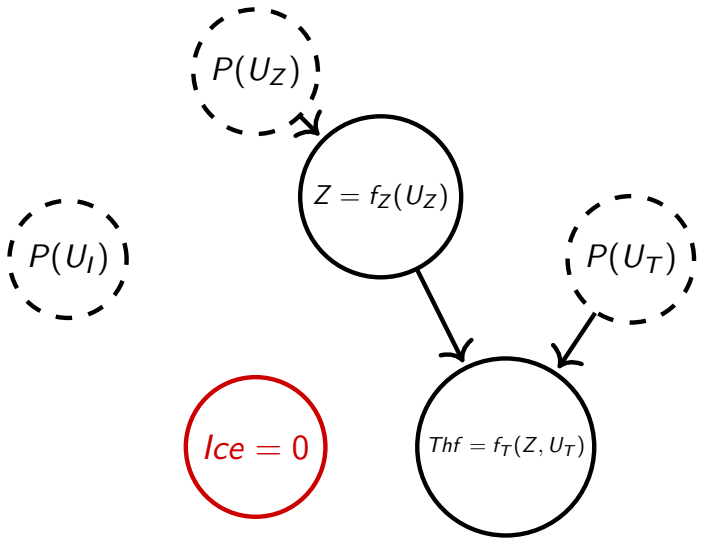
L2: Conditioning and Intervening

Conditioning $\neq$ Intervention



$$P_{\mathcal{M}}(Thf|Ice = 0)$$

Knowledge of $Ice = 0$ allows inference on distribution of Z and then Thf.

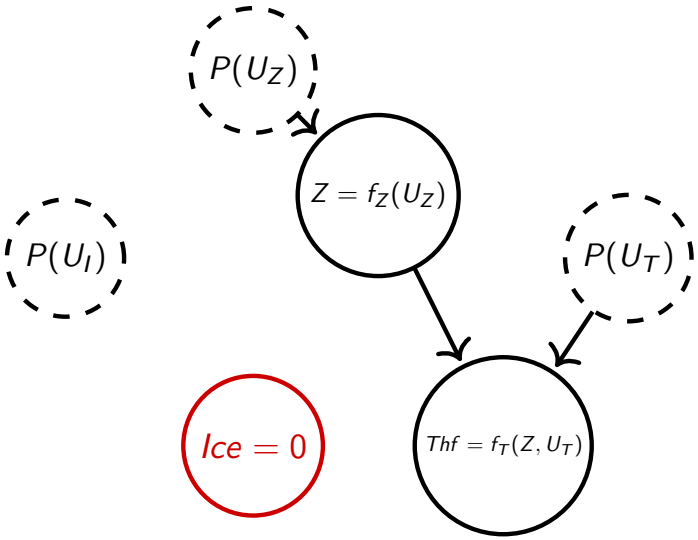


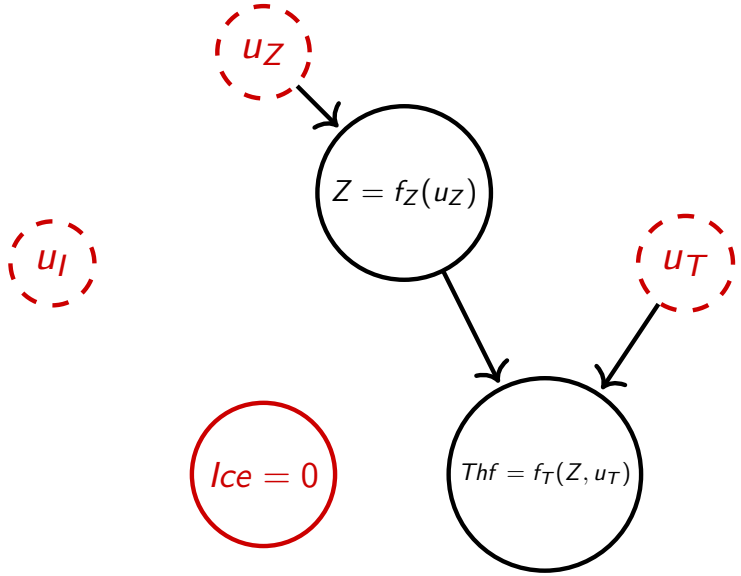
$$P_{\mathcal{M}}(Thf|do(Ice = 0))$$
$$P_{\mathcal{M}_{do}}(Thf|Ice = 0)$$

Knowledge of $do(Ice = 0)$ does not affect the distribution of Z.

L3: Intervening and Counterfactuals

Intervention $\neq$ Counterfactuals





$$P_{\mathcal{M}}(Thf|do(Ice = 0))$$
$$P_{\mathcal{M}_{do}}(Thf|Ice = 0)$$

The value of *Thf* is independent of *Ice*, determined by a random samples of Z.

$$P_{\mathcal{M}}(Thf_{Ice=42}|do(Ice = 0))$$
$$P_{\mathcal{M}_{Ice=42do}}(Thf|Ice = 0)$$

The value of *Thf* is always independent of *Ice*, but the sample from Z is conditioned to the value that produced $Ice = 42$.