Causal Models and Machine Learning

Fabio Massimo Zennaro
fabiomz@ifi.uio.no

University of Oslo

Oslo Machine Learning Meetup
December 2nd, 2019

Outline

1. *Introduction*: what is this theory of causality?
2. *A Motivating Examples*: why do we care about causality? - an example.
3. *Beyond the Limits of the Language*: how do we express causality?
    4. *Causal Effects*
    5. *Counterfactuals*
6. *Machine Learning*: how does all of this relate to ML?
7. *Conclusions*

1. Introduction

What do we mean by causality?

Theoretically, hard to define (Aristotle, Aquinas, Hume; Kirchoff, Russell, Rohrlich [7])

Yet, we have an **operational** intuition of what it means:

- It implies a *relationship* between things/variables.
- It has a *counterfactual* aspect: *ceteribus paribus*, the presence or absence of a variable determines the outcome.
- We can probe causality by *interventions*.
- We can learn from observations (*averaging*).

Why to consider causality?

Theoretically:

- It is the foundation of our understanding of the world.
- It is at the core of the scientific endeavour.

Practically:

- It allows us to differentiate association and causation.
- It allows us to model non-static settings.
- It allows us to define interventions and policies.
- It allows us to learn robust models.

Approaches to Causality [8][11][15]

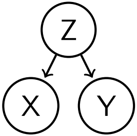
Potential outcomes

$Y_0(1)$

Jerzy Neyman, Donald Rubin

Statistics, epidemiology...

Structural causal model (SCM)



Tryggve Haavelmo, Judea Pearl

Economics, computer science...

The two formalisms are equivalent.
Here we will follow the SCM approch.

2. A Motivating Example

Ice Creams and Thefts [10]

Assume we monitored the number of *ice-creams sold* (Ice) and the number of *thefts* (Thf) in our town:

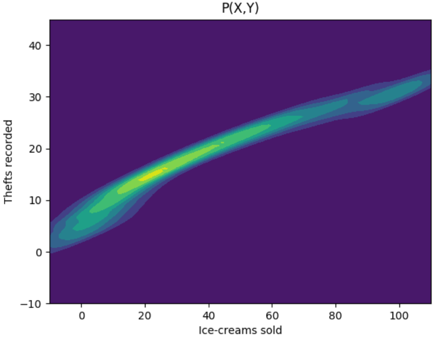
Ice	Thf
36	20
35	18
101	31
17	12
50	23
65	25
...	...

What can we infer from this data?

The Ideal Statistician

- ✓ We learn the *joint distribution* of the variables: $P(Ice, Thf)$
- ✓ We can *marginalize* and *condition*: $P(Thf), P(Thf|Ice)$

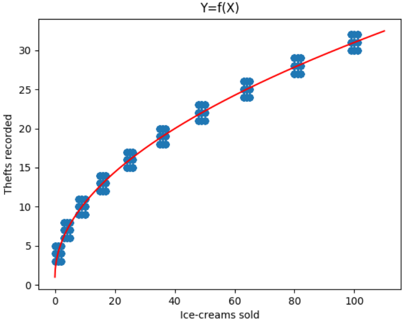
Ice	Thf
36	20
35	18
101	31
17	12
50	23
...	...



The Ideal Machine Learner

- ✓ We can learn how the variables are *correlated*: $Ice \uparrow$, $Thf \uparrow$
- ✓ We can *predict* a variable from another: $Thf = f(Ice)$, $Ice = f(Thf)$

Ice	Thf
36	20
35	18
101	31
17	12
50	23
...	...



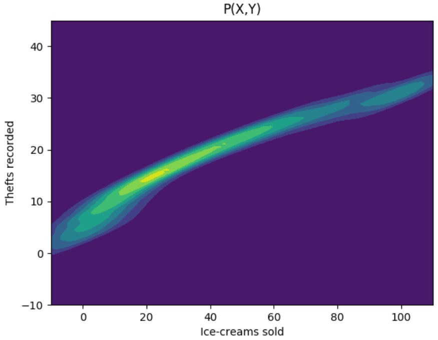
$$\boxed{Thf = 3 * \sqrt{Ice} + 1}$$

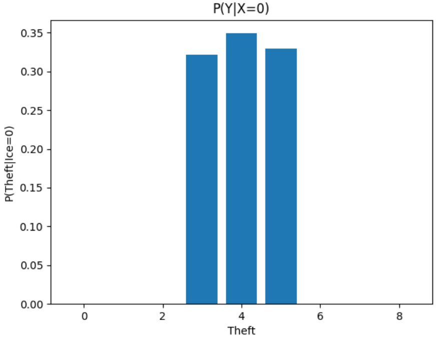
Let's Intervene!

So, what if we stop the sale of ice-creams?

The Naive Statistician

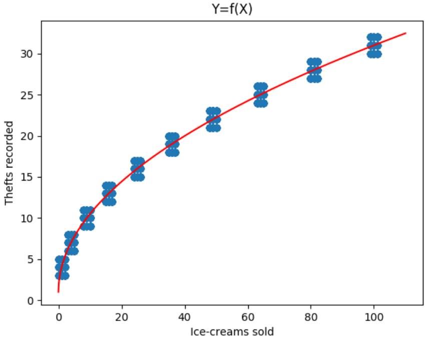
Let's compute the conditional for $Ice = 0$.





The Naive Machine Learner

Let's use our model to compute $Ice = 0$.



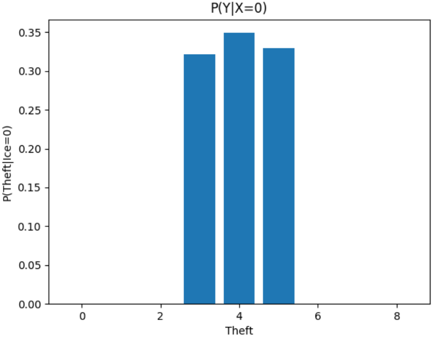
$$\boxed{Thf = 3 * \sqrt{Ice} + 1}$$

$$Thf = 3 * \sqrt{0} + 1$$

$$Thf = 1$$

Clashing with Reality

Let's collect data now.

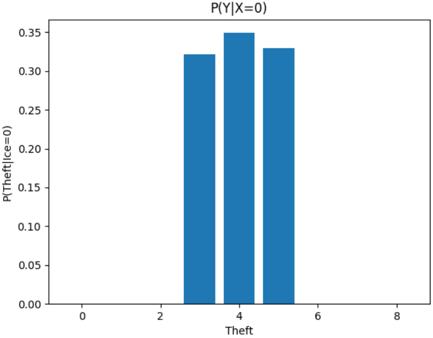


$$\boxed{Thf = 1}$$

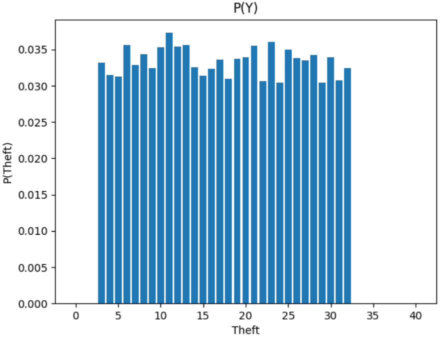
Ice	Thf
0	6
0	29
0	9
0	10
0	17
0	12
0	14
...	...

Clashing with Reality

Let's collect data now.



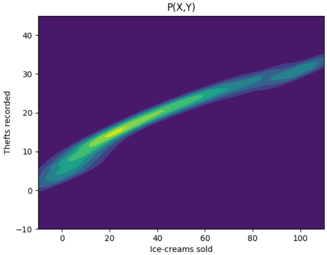
$$\boxed{Thf = 1}$$



$$E\left[Thf\right] = 17.628$$

What's the Problem in What We Did?

From the point of view of the *data model*:



- Changing *Ice* means changing the joint distribution.
- Samples are not from the same distribution anymore.

What's the Problem in What We Did?

From the point of view of the *learned model*:

$$\boxed{Thf = 3 * \sqrt{Ice} + 1}$$

- The input-output relation is not causal.
- We learned to predict a correlation, not a causal mechanism.

How to Account for Intervening?

- ✓ We want to learn a causal mechanism:

$$\text{Effect} = f\left(\text{Cause}\right)$$

$$P\left(\text{Effect} \middle| \text{Cause}\right)$$

- ✓ We need an idea of *directionality* between variables:

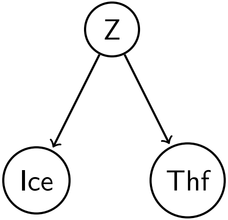
$$\text{Ice} \longrightarrow \text{Thf} \qquad\qquad \text{Ice} \longleftarrow \text{Thf}$$

- ✓ We need to understand how correlated variables can be causally related.

Reichenbach's Principle

Two correlated variables X and Y can be causally related in only three ways[1]: $X \rightarrow Y$, $X \leftarrow Y$, $X \leftarrow Z \rightarrow Y$.

There likely is a *common cause* (Z) between the variables, such as the temperature:



We have a **confounder** between *Ice* and *Thf*.

[1] Excluding colliders and coincidences.

Bottom Line of Our Example

Causal reasoning is not necessary if:

- We want to model/predict in a static setting.

However, causal modelling may allow us (among other things) to:

- Distinguish and learn *actual causal mechanisms*;
- Deal with settings changing under *interventions*.

3. Beyond the Limits of the Language

Concepts We Can Not Express [11, 6]

There are ideas we can not express in statistical/ML language.

Statistics	**Causality**
Association	Cause
Correlation	Causation
Non-directionality	Directionality
Prediction	Action
Observation	Intervention

There is a **chasm between statistics and causality**.

Questions We Can Not Express

There are questions we can not express in statistical/ML language!

Causality	**3. Counterfactuals**	What would have Y been, had X been x' when instead it was x? $P\left(Y_{do(X=x')} \middle\| Y=y, X=x\right)$ Structural causal models
	2. Causal Effects	What is the effect of X on Y? $P\left(Y \| do\left(X=x\right)\right)$ Causal Bayesian networks
Stat/ML	**1. Associative Relationships**	How does Y relate to X? $P\left(Y \| X\right)$ Bayesian networks

This constitutes the **Pearl's Causality Ladder** [12, 13, 20, 15]

A New Language

We want a **language** that allow us to express the ideas and questions we care about (*cause, directionality, intervention, counterfactual*).

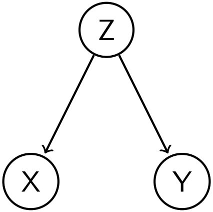
We want a **theory** that bridges the gap with statistics.

$$\begin{array}{c}\text{causality}\\ \text{formalism}\end{array} \longrightarrow \begin{array}{c}\text{statistical}\\ \text{formalism}\end{array}$$

$$\begin{array}{c}\text{interventional}\\ \text{domain}\end{array} \longrightarrow \begin{array}{c}\text{observational}\\ \text{domain}\end{array}$$

DAG Language

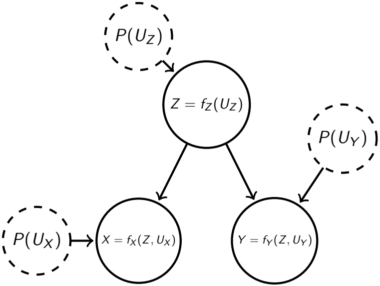
Directed acyclic graphs are a natural and intuitive way to express causal relatioships and their directionality.



A DAG is a purely *mathematical structure* which we endow with *causal meaning*.

SCM Language

Structural causal models provide a way to deal with interventions and counterfactuals.



We have a *probabilistic model* expressed via a *reparametrization trick*.

Assumptions

A SCM expresses and encodes statistical and causal **assumptions**:

- *Acyclicity*: no loops in the graph.
- *Causal Markov assumption*: a node is independent of its non-effects given its direct causes.
- *Zero influence*: missing arrow means no causal relationship.
- *Common cause completeness*: all common causes are modeled.
- *Autonomous functions*: changing a function does not affect other functions.
- ...

No causes in, no causes out.

Structural causal models [18]

$$P(X, Y)$$

$$X \longrightarrow Y \qquad\qquad X \longleftarrow Y$$

$$P(Y|X)P(X) \qquad\qquad P(X|Y)P(Y)$$

- Statistics works with the *joint*; factorizations are instrumental.
- Causality makes the *assumption* that one of the factorizations is the *true causal model*.

A causal model contains more information than a statistical one.

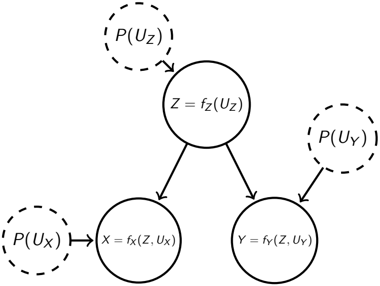
4. Causal Effects

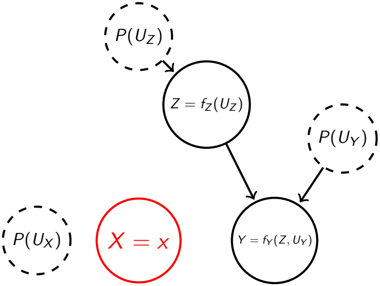
Level 2

Causality	**3. Counterfactuals**	What would have Y been, had X been x' when instead it was x? $P\left(Y_{do(X=x')} \middle\| Y=y, X=x\right)$ Structural causal models
	2. Causal Effects	What is the effect of X on Y? $P\left(Y \middle\| do\left(X=x\right)\right)$ Causal Bayesian networks
Stat/ML	**1. Associative Relationships**	How does Y relate to X? $P\left(Y \middle\| X\right)$ Bayesian networks

Interventions

An **intervention** is a new operation $\mathrm{do}(X = x)$ by which a variable is set to a fixed value.

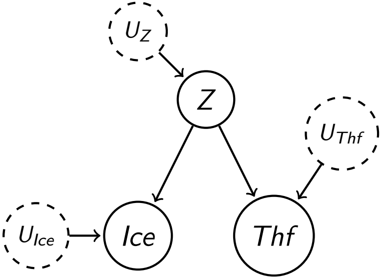




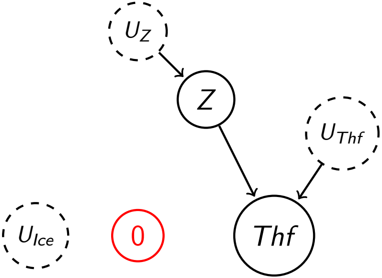
We obtained the new *intervened* (or *post-intervention*) model.

Back to Our Example

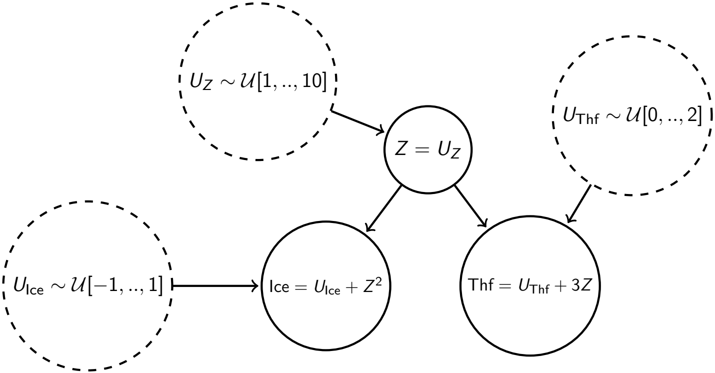
We learned in an *observational* environment:



We deployed in this *interventional* environment:

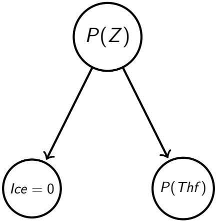


(Behind the Scene: The Actual SCM in Our Example)



Interventions are not Conditioning

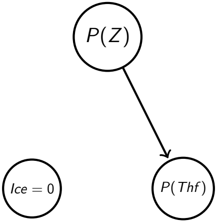
Conditioning $\neq$ Intervention



$$P\,(Thf\,|\,Ice = 0)$$

Distribution of Thf when observing $Ice = 0$.

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



$$P\,(Thf\,|\mathrm{do}\,(X = 0))$$

Distribution of Thf when intervening to do $Ice = 0$.

Knowledge of $\mathrm{do}\,(Ice = 0)$ does not affect the distribution of Z.

Causal Inference

Most of our data are statistical/observational data:

$$\begin{array}{c} \text{causality} \\ \text{formalism} \end{array} \longrightarrow \begin{array}{c} \text{statistical} \\ \text{formalism} \end{array}$$

$$\begin{array}{c} \text{interventional} \\ \text{domain} \end{array} \longrightarrow \begin{array}{c} \text{observational} \\ \text{domain} \end{array}$$

Causal inference provides theory (*do-calculus*) and algorithms (*ID algorithm*) to decide whether an interventional question can be reduced to an observational question, and techniques (*backdoor adjustment, inverse probability weighting, propensity score*) to compute these estimates.

Intervention $\rightsquigarrow$ **Conditioning**

5. Counterfactuals

Level 3

Causality	**3. Counterfactuals**	What would have Y been, had X been x' when instead it was x? $P\left(Y_{do(X=x')} \middle| Y=y, X=x\right)$ Structural causal models	
	2. Causal Effects	What is the effect of X on Y? $P\left(Y \middle| do\left(X=x\right)\right)$ Causal Bayesian networks	
Stat/ML	**1. Associative Relationships**	How does Y relate to X? $P\left(Y \middle| X\right)$ Bayesian networks	

Counterfactuals

A **counterfactual** is an operation by which we compute a quantity of interest in an alternate world in which we perform an intervention.

$$P\left(Y_{do(X=x')} | Y = y, X = x\right)$$

This reflects the *counterfactual question*: assuming we observed $Y = y$ and $X = x$, what would have Y been, had we acted on $do\,(X = x')$?

Counterfactuals

Interventions $\neq$ Counterfactuals

$$P\left(Bet = Coin | \mathrm{do}\left(Bet = head\right)\right)$$

Probability of winning if we force the bet to head.

The outcome of the coin toss is still random, and the chance of winning half.

$$P\left(Bet = Coin_{do(Bet=head)} | Coin = head, Bet = tail\right)$$

Probability of winning if we had forced the bet to head, having observed the outcome head and the bet tail.

We know with certainty the result of the bet.

Topics We Did Not Address

- Approaches to *causal inference*
- *Causal discovery*
- Causal modelling in *time-varying settings*
- *Mediation analysis*
- Inference with *missing data*
- Inference with *partially specified models*
- Inference with *hidden variables*
- ...

6. Machine Learning

Relation to Machine Learning

A double relation: ML can use causality theory to improve learning, and causaliy theory can use ML to improve causal inference.

We consider four sample applications:

- Causal and anti-causal learning
- Invariance learning
- Reinforcement learning
- Counterfactual fairness

Causal and Anti-Causal Learning [19]

Causal Learning

Given samples *(cause, effect)* we learn:

$$\text{Effect} = f\left(\text{Cause}\right)$$

$$P\left(\text{Effect}|\text{Cause}\right)$$

e.g.: predicting structure of proteins.

Anti-Causal Learning

Given samples *(effect, cause)* we learn:

$$\text{Cause} = f\left(\text{Effect}\right)$$

$$P\left(\text{Cause}|\text{Effect}\right)$$

e.g.: classifying images.

$$P\left(\text{Effect}|\text{Cause}\right) \perp P\left(\text{Cause}\right)$$

Semi-supervised Learning [19]

$$P\left(\text{Effect}|\text{Cause}\right) \perp P\left(\text{Cause}\right)$$

Causal Learning

In SSL, we receive more samples *(cause)*, and we aim to learn:

$$P\left(\text{Effect}|\text{Cause}\right)$$

Learning more on how the cause distributes do not provide information on how the effect mechanism behaves. (But it may help reducing the risk!)

Anti-Causal Learning

In SSL, we receive more samples *(effect)*, and we aim to learn:

$$P\left(\text{Cause}|\text{Effect}\right)$$

Learning more on how the effect distributes may help us infer more about the cause mechanism under standard SSL assumptions (*smoothness*, *clustering*).

Covariate Shift [19]

$$P\,(\text{Effect}|\text{Cause}) \perp P\,(\text{Cause})$$

Causal Learning

In CS, we receive test samples from $P'(\text{Cause})$, and we aim to compute:

$$P\,(\text{Effect}|\text{Cause})$$

The effect mechanism is not affected by shifts in the distribution of the causes. (But risk may require adjustment!)

Anti-Causal Learning

In CS, we receive test samples from $P'(\text{Effect})$, and we aim to compute:

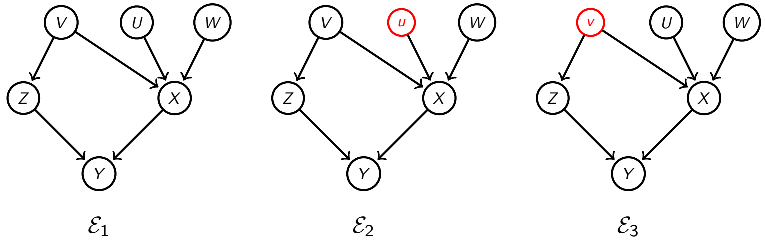
$$P\,(\text{Cause}|\text{Effect})$$

A change in the effect mechanism affects the conditional distribution of causes.

Invariance Learning

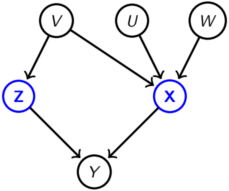
In absence of a model, we may try to learn a *local structure*.

Suppose we are given data from different *environment* (= *interventional domains*)



Invariance Learning

From data in different environments $\mathcal{E}_i$ we can learn the sets of variables that is *invariant* in all the settings (= under all interventions).



$\mathcal{E}_i$

The set of invariant variables are the (true) *direct causes* of the variable of interest.

Prediction of invariance [14, 16] and learning with invariant risk minimization [1] allow for learning *robust model* (= *transfer learning*).

Reinforcement Learning

Reinforcement learning deals with an **interventional setting**.

Performing actions, an agent probes the distribution of an environment under intervention:

$$P(E|do(A = a))$$

Bandit problems and *reinforcement learning* may be expressed in causal terms [3].

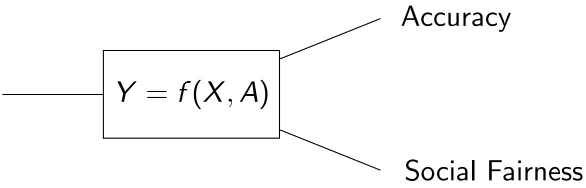
Reinforcement Learning

Reinforcement learning works without structural models and causal formalization is still debated.
There are promising point of contacts:

- *Counterfactual reasoning with structure* in ad placement problems [3]
- Relation between *offline policy evaluation* and *inverse probability weighting* [3, 21]
- *Counterfactually-guided policy search* [4]

Fairness

Fairness is concerned with deciding if learned systems are socially fair.



Important in applications such as job recruiting, loan decisions, police deployment.

How to Measure Fairness?

There are several approaches to guarantee fairness [2]:

- *Fairness through unawareness*: $\hat{Y} = f(X)$
- *Demographic parity*: $P(\hat{Y}|A=0) = P(\hat{Y}|A=1)$
- *Equality of opportunity*: $P(\hat{Y}|A=0, Y=1) = P(\hat{Y}|A=1, Y=1)$

These measures are either insufficient [9] or conflicting [5].

Counterfactual Fairness [9, 17]

We can enforce an individual-level fairness in *counterfactual* terms:

$$P\left(\hat{Y}|X=x, A=a\right) = P\left(\hat{Y}_{do(A=a')}|X=x, A=a\right)$$

For instance:

$$P(\text{accepted}|X=x, A=\text{female}) \\ = \\ P\left(\text{accepted}_{do(A=\text{male})}|X=x, A=\text{female}\right)$$

7. Conclusions

Schölkopf-Peters Table

	i.i.d. prediction	Prediction under distributional shift or intervention	Counter-factual Question	Physical insight	Learn from data
Mechanistic-physical	✓	✓	✓	✓	?
Structural causal model	✓	✓	✓	?	?
Causal graphical model	✓	✓	×	?	?
Statistical model	✓	×	×	×	✓

Conclusions

The theory of causality empowers machine learning:

- Provides a formalism to reason causally (the SCM framework is general, it helps making assumptions explicit, and it eases reasoning via graphs).
- Allows to express causal statements.
- It will likely have an important role in learning robust and flexible models.
- It may spur us to move beyond deep learning.

It comes with a cost though:

- Assumptions/structures!

Thanks!

Thank you for listening!

References I

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[2] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: the state of the art. *arXiv preprint arXiv:1703.09207*, 2017.

[3] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research*, 14(1):3207–3260, 2013.

[4] Lars Buesing, Theophane Weber, Yori Zwols, Sebastien Racaniere, Arthur Guez, Jean-Baptiste Lespiau, and Nicolas Heess. Woulda, coulda, shoulda: Counterfactually-guided policy search. *arXiv preprint arXiv:1811.06272*, 2018.

References II

[5] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[6] A Philip Dawid. Statistical causality from a decision-theoretic perspective. *Annual Review of Statistics and Its Application*, 2:273–303, 2015.

[7] Mathias Frisch. *Causal reasoning in physics*. Cambridge University Press, 2014.

[8] Andrew Gelman. Causality and statistical learning. *American Journal of Sociology*, 117(3):955–966, 2011.

[9] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4069–4079, 2017.

[10] Judea Pearl. *Causality*. Cambridge university press, 2009.

References III

[11] Judea Pearl. An introduction to causal inference. *The international journal of biostatistics*, 2010.

[12] Judea Pearl. Theoretical impediments to machine learning with seven sparks from the causal revolution. *arXiv preprint arXiv:1801.04016*, 2018.

[13] Judea Pearl. Sufficient causes: Revisiting oxygen, matches, and fires. *Journal of Causal Inference*, 2019.

[14] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

[15] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

References IV

[16] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

[17] Chris Russell, Matt J Kusner, Joshua Loftus, and Ricardo Silva. When worlds collide: integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems*, pages 6417–6426, 2017.

[18] Bernhard Schlkopf. Causality for machine learning, 2019.

[19] Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*, 2012.

[20] Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9(Sep):1941–1979, 2008.

References V

[21] Guy Tennenholtz, Shie Mannor, and Uri Shalit. Off-policy evaluation in partially observable environments. *arXiv preprint arXiv:1909.03739*, 2019.