

# Causality and Generative Models

---

*This week's paper is [1]*

*Author: Anton Zhitomirsky*

## Contents

<b>1 Causality</b>	<b>2</b>
1.1 What is Causlity?	2
1.2 Reichenbach's Common Cuase Principle	2
1.3 Simpson's Paradox	3
1.4 Predictive Modelling	4
1.4.1 Example	5
1.5 Ladder of Causation	6
1.6 Structural Causal Models	6
1.7 Distribution Math	8
1.7.1 Computing Counterfactuals Example	10
1.8 Deep Structural Causal Models	10
1.8.1 Normalizing Flows	11
1.8.2 VAEs	13
1.9 Evaluating Counterfactual	20
1.9.1 Composition Axiom	21
1.9.2 Effectiveness Axiom	22
1.9.3 Reversibility Axiom	22
<b>Bibliography</b>	<b>24</b>

# 1 Causality

## 1.1 What is Causality?

### What is Causality?

- Scientific inquiry is often motivated by causal questions:
- I. How effective is a **treatment** in preventing a **disease**?
  - II. If I take this **pill**, will my **headache** be gone?
  - III. Would my **grades** have been better had I **studied** more?

Causality is the relationship between **cause** and **effect**

**Figure 1:** Correlation doesn't imply causation

## 1.2 Reichenbach's Common Cause Principle

### Reichenbach's Common Cause Principle



Hans Reichenbach

- I. **Chocolate consumption** causes **Nobel prize win**

$$C \rightarrow N$$

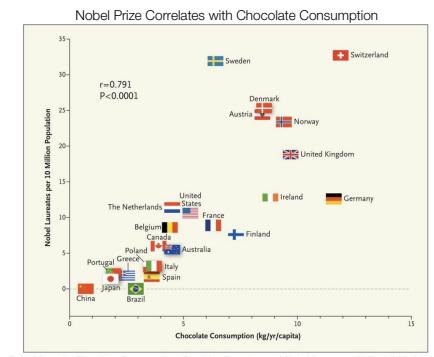
- II. **Nobel prize win** causes **chocolate consumption**

$$C \leftarrow N$$

- III. Both are caused by an unknown factor  $U$

$$C \leftarrow U \rightarrow N$$

Could  $U$  be GDP per capita or wealth per adult?

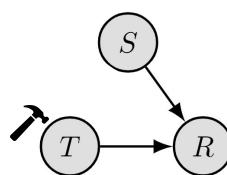


**Figure 2:** there is a case that countries that eat the most chocolate have the most Nobel laureates. So there are three possible explanations for the correlations. This is the **Reichenbach's Common Cause Principle**

### 1.3 Simpson's Paradox

## Simpson's Paradox

- Stone size ( $S$ ) is a **confounder**
- Make treatment ( $T$ ) independent of size:
  - ◆ What's the recovery ( $R$ ) rate if **all** subjects receive treatment **A** vs **B**?



Edward Simpson

Kidney Stone Size	Treatment A	Treatment B
Small	93% (81/87)	87% (234/270)
Large	73% (192/263)	69% (55/80)
Both	78% (273/350)	83% (289/350)

??

Patients with **large stones** received the **better treatment (A)**, and those with **small stones** received the **inferior treatment (B)**

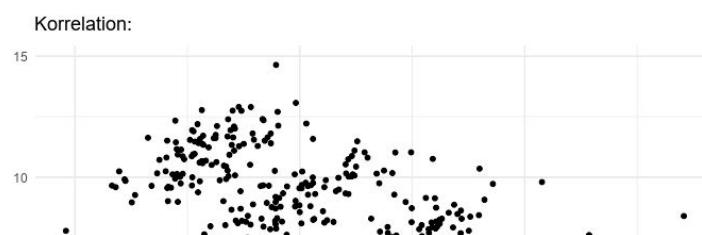
Charig CR, et al. Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. Br Med J (Clin Res Ed). 1986 Mar 29.

**Figure 3:** Another paradox includes two treatments for different sizes of kidneys. Treatment A has better recovery rates, but Treatment B has an average better recovery rate. This is the **Simpson's Paradox**. The reason is that the stone size is a confounding variable. What happened was that the doctors were prescribing the 'better' treatments for the patients with the worst conditions, which caused a bias.

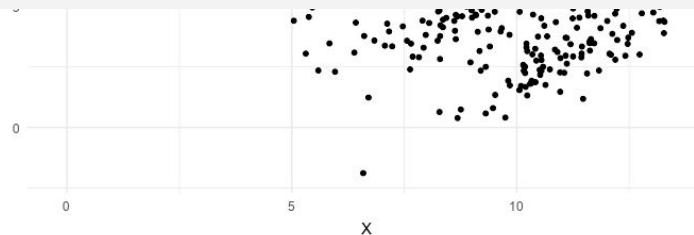
## Simpson's Paradox



Edward Simpson



Correlations may **reverse** depending on how we aggregate or filter data and its subpopulations



From: [https://upload.wikimedia.org/wikipedia/commons/f/fb/Simpsons\\_paradox\\_-\\_animation.gif](https://upload.wikimedia.org/wikipedia/commons/f/fb/Simpsons_paradox_-_animation.gif)

**Figure 4:** depending on the way you aggregate on your populations, the correlations may reverse (e.g. a downward slope vs many upwards sloping lines)

## 1.4 Predictive Modelling

### Predictive Modelling

Given an image  $X$ , train a model to predict some label  $Y$

$$P(Y|X)$$

→ Assumptions:

- ◆ Sufficient training data ( $X, Y$ ) is available
- ◆ Training and test data come from the same distribution

Figure 5: The goal is to estimate  $P(Y|X)$  given some assumptions.

### A Causal Perspective: Predictive Modelling

What is the causal relationship between image  $X$  and label  $Y$ ?

$$P(Y|X)$$

$$X \rightarrow Y$$

**causal**

(predict effect from cause)

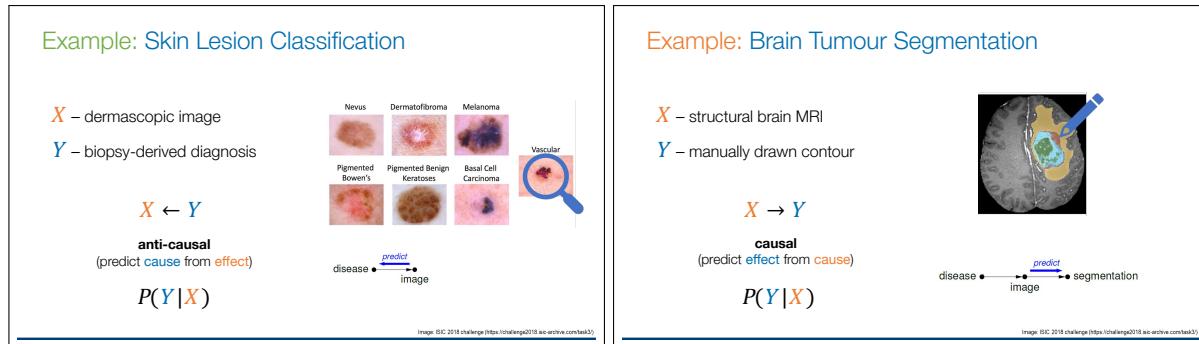
$$Y \rightarrow X$$

**anti-causal**

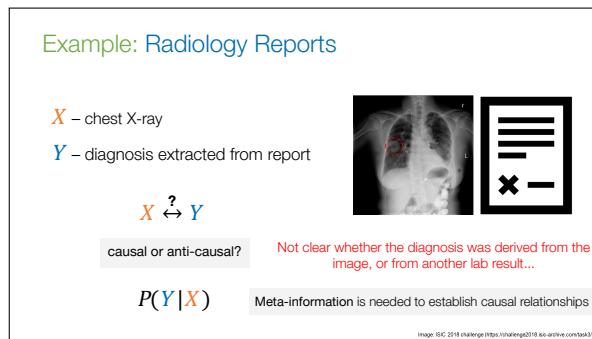
(predict cause from effect)

Figure 6: Firstly we are predicting the effect from the cause, and secondly we are predicting the cause from the effect.

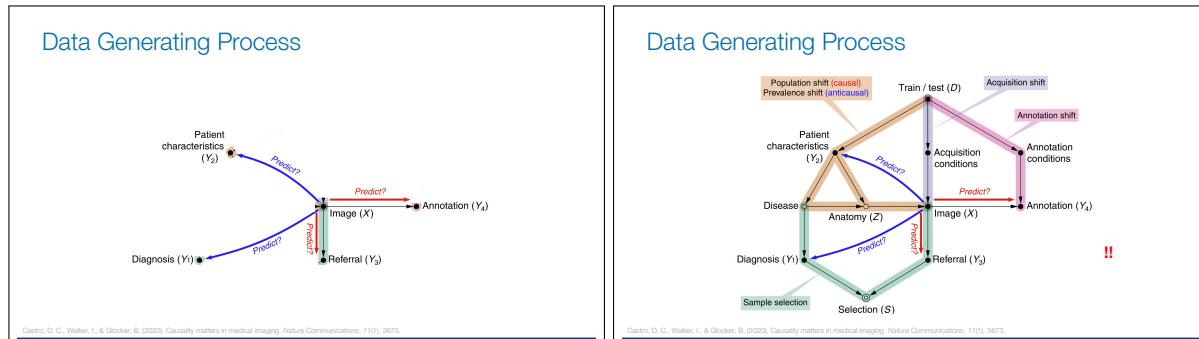
### 1.4.1 Example



- (a) the disease makes the image look a certain way, however, the thing we are predicting comes from a biopsy, not the image.
- (b) the segmentation is fully determined by the image. Therefore, the image causes the segmentation map.

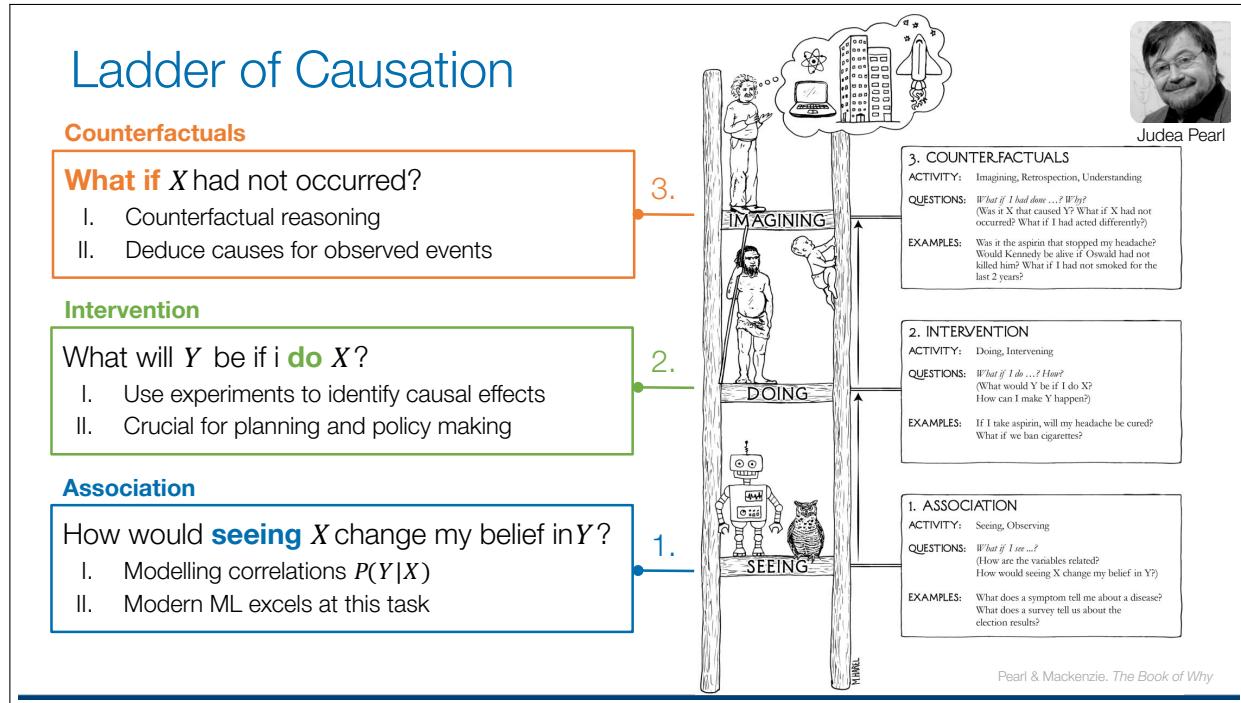


- (c) sometimes you need more information to determine



- (d) we need to be careful about our data generating process
- (e) complications may arise when you annotate the data, because we get annotation shifts. When we acquire the data there can be acquisition shifts and so on. E.g. an annotator might annotate inconsistently to the annotations.

## 1.5 Ladder of Causation



**Figure 7:** “the first rung deals with association, so how would seeing  $X$  change my belief in  $Y$ ? E.g. in deep learning models excel at this task if we have enough data and a big enough neural network. The second rung deals with intervention. So this is about asking questions like what will  $Y$  be if i do  $X$ ? The final does counterfactuals, these are hypothetical questions like what would happen if  $X$  had not occurred.

## 1.6 Structural Causal Models

### Structural Causal Models

→ A Structural Causal Model (SCM) is a triple:  $\mathcal{M} := \langle X, U, F \rangle$

- Two sets of variables:

$$X = \{x_1, \dots, x_N\} \quad U = \{u_1, \dots, u_N\}$$

- A set of functions known as **causal mechanisms**:

$$F = \{f_1, \dots, f_N\}$$

- The value of each variable is a function of its **parents** (direct causes):

$$x_k := f_k(\mathbf{pa}_k, u_k), \quad k = 1, \dots, N$$

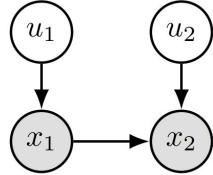
**Figure 8:**  $X$  is the set of observations, and  $U$  are unobserved conditions (e.g. background, unaware or things you have no control over.) Lastly, the idea is that each  $X$  variable is a function of its direct causes via these causal mechanisms.

For example:

## Structural Causal Models

→ Example: A simple SCM:

- ◆  $x_1, x_2$  are **endogenous** whereas  $u_1, u_2$  are **exogenous**



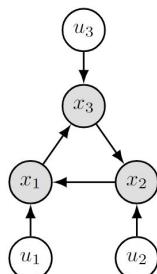
$$x_1 := f_1(u_1), \quad u_1 \sim \mathcal{N}(0, 1)$$

$$x_2 := f_2(x_1, u_2), \quad u_2 \sim \mathcal{N}(0, 1)$$

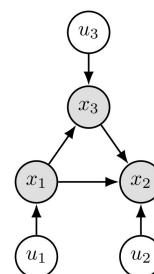
**Figure 9:** **endogenous:** caused by any variable in the model, **exogenous:** caused by factors external to the model or independent of the model.

## Structural Causal Models

→ Acyclic SCMs can be represented by a **Directed Acyclic Graphs (DAGs)**, with edges pointing from **causes** to **effects**



(a) Cyclic

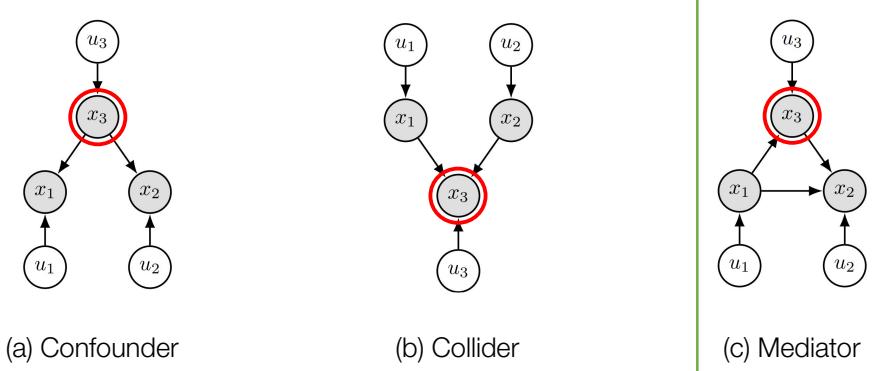


(b) Acyclic

**Figure 10:** We can represent Acyclic SCMs as a DAGs

## Structural Causal Models

- Acyclic SCMs can be represented by a Directed Acyclic Graphs (DAGs), with edges pointing from **causes** to **effects**



**Figure 11:** a) both the kidney and chocolate example, b) when you consider on  $x_3$  you would also experience correlation between  $x_1$  and  $x_2$ . c)  $x_3$  mediates the causal effect of  $x_1$  on  $x_2$ .

## 1.7 Distribution Math

### Observational Distribution (rung 1.)

- SCMs with jointly independent exogenous noises are called **Markovian**:

$$P(u_1, \dots, u_N) = \prod_{k=1}^N P(u_k)$$

- Markovian SCMs induce a unique joint **observational distribution** over the endogenous variables:

$$P_{\mathcal{M}}(x_1, \dots, x_N) = \prod_{k=1}^N P_{\mathcal{M}}(x_k \mid \text{pa}_k)$$

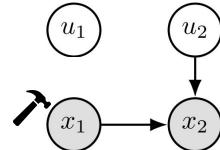
- Each variable is **independent** of its non-descendants given its direct causes (*causal Markov condition*)

**Figure 12:** 1) distribution of the noise terms factorize across every node in the graph. 2) these also factorize over every node in the graph.

## Interventional Distribution

(rung 2.)

- SCMs predict the **causal effects** of actions via interventions
- Interventions answer causal questions like:
  - ◆ E.g. what would  $x_2$  be if we set  $x_1 := c$ ?
- Interventions replace one or more of the structural assignments and are denoted with the *do* operator:  $do(x_k := c)$
- This induces a submodel  $\mathcal{M}_c$  and its entailed distribution is known as the **interventional distribution**:  $P_{\mathcal{M}_c}(X | do(c))$



**Figure 13:** we intervene with the *do* operator. This produces a submodel. This is generally different to the observational distribution we start out with before intervening on anything.

## Counterfactuals

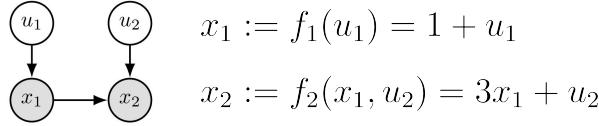
(rung 3.)

- SCMs can consider **hypothetical** scenarios:
  - ◆ Given that we observed  $(x_1, x_2)$ , what would  $x_2$  have been had  $x_1$  been  $c$ ?
  - ◆ All else being equal, would I have been **late** had I not **missed the train**?
- Counterfactual inference involves three steps:
  1. Abduction: Update  $P(U)$  given observed evidence, i.e. infer posterior  $P(U | X)$
  2. Action: Perform an intervention e.g.  $do(\tilde{x}_k := c)$  and obtain the submodel  $\mathcal{M}_c$
  3. Prediction: Use the modified model  $\langle \mathcal{M}_c, P(U | X) \rangle$  to compute counterfactuals

**Figure 14:** a way to solve this is counterfactual inference.

### 1.7.1 Computing Counterfactuals Example

#### Example: Computing Counterfactuals

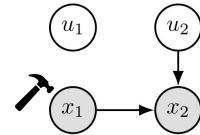


**Q:** Given we observed  $\{x_1=2, x_2=4\}$ , what would  $x_2$  have been had  $x_1$  been 5?

1. Abduction:  $u_1 = f_1^{-1}(x_1) = x_1 - 1 \implies u_1 = 1$

$$u_2 = f_2^{-1}(x_1, x_2) = x_2 - 3x_1 \implies u_2 = -2$$

2. Action:  $\tilde{x}_1 := 5$



3. Prediction:  $\tilde{x}_2 = 3\tilde{x}_1 + u_2 = 3 \cdot 5 - 2 = 13$

## 1.8 Deep Structural Causal Models

#### Deep Structural Causal Models<sup>1</sup>

- Leverage deep generative models to learn SCM mechanisms:

$$x_k := f_k(\mathbf{pa}_k, u_k)$$

- Tractably estimate causal effects of interventions and perform counterfactual inference, i.e. answer “what if...?” type questions

- Abduction is challenging in complex problems, e.g. medical imaging

**Research Question:** Can we generate plausible high-fidelity image counterfactuals of real-world data, and if so, how do we evaluate them?

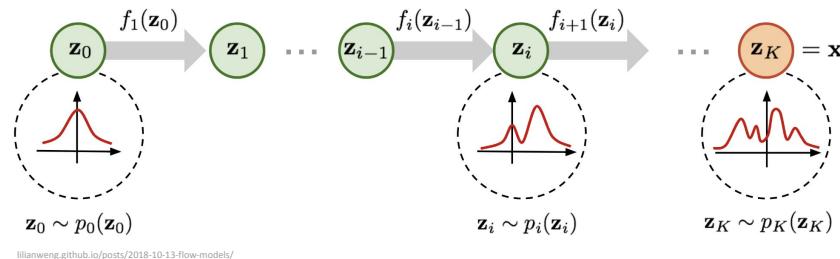
<sup>1</sup> Pawlowski, Castro, Glocker. Deep Structural Causal Models for Tractable Counterfactual Inference. NeurIPS 2020

**Figure 15:** Can we use DGMs to learn structural causal mechanisms? Abduction is challenging in medicine.

### 1.8.1 Normalizing Flows

## Deep Mechanisms: Normalizing Flows

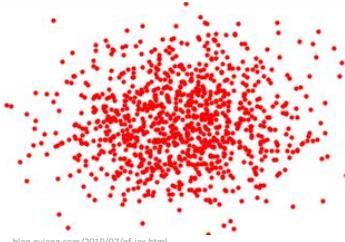
- Normalizing Flows (NFs) build complex probability distributions via successive (invertible) transformations of simple distributions



Kobyzev, Prince, Brubaker. Normalizing Flows: An Introduction and Review of Current Methods. *PAMI* 2020  
 Papamakarios, Nalisnick, Rezende, Mohamed, Lakshminarayanan. Normalizing Flows for Probabilistic Modeling and Inference. *JMLR* 2021

## Deep Mechanisms: Normalizing Flows

- Normalizing Flows (NFs) build complex probability distributions via successive (invertible) transformations of simple distributions



- **TLDR:** NFs enable invertible mechanisms and deterministic abduction:

$$x = f_\theta(\mathbf{pa}_x, u_x) \quad u_x = f_\theta^{-1}(\mathbf{pa}_x, x)$$

**Figure 16:** Here the dataset is transforming into a spiral after multiple transformations. We attempt to maximise the likelihood of observing the data. The benefit is that they allow us to learn invertible mechanisms and deterministic abduction then we can invert it (since we constrained it to be invertible) and find the exogenous noise term  $u_x$  given the  $x$

## Intuition: Normalizing Flows

- Change-of-variables formula:

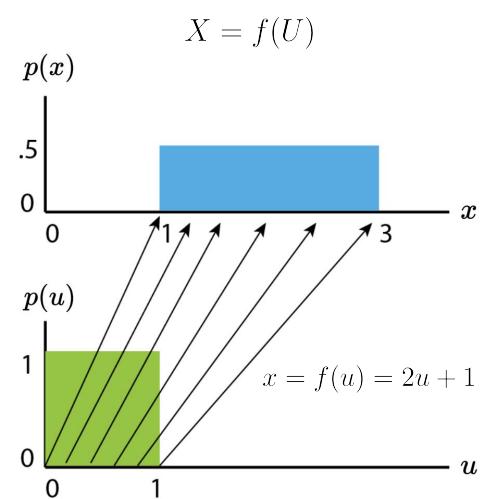
$$p(x) = p(u) \left| \frac{du}{dx} \right|$$

volume correction

$$u = f^{-1}(x) = \frac{x - 1}{2}$$

$$\frac{du}{dx} = \frac{df^{-1}(x)}{dx} = \frac{d}{dx} \left( \frac{x - 1}{2} \right) = \frac{1}{2}$$

$$p(u) = p_U(f^{-1}(x)) = \frac{1}{1 - 0} = 1$$



PDF of  $\mathcal{U}[a, b]$  is  $\frac{1}{b - a}$

**Figure 17:** “Let’s say that we have two random variables which are related by a deterministic mapping. Let’s say this function is  $f(u) = 2u + 1 = x$ . The change in variables says that we can evaluate the density of  $X$  as the density of  $u$  times the volume correction term which tries to preserve density when we apply this function. Therefore, we invert the function and differentiate it. After this we can see that for any value of  $u$  the density of  $x$  is just one half because we have  $1 \times \frac{1}{2}$ .”

## Training Objective: Normalizing Flows

- Maximum (log) likelihood training objective:

$$\log p(x) = \log p(f_\theta^{-1}(x)) + \log \left| \frac{df_\theta^{-1}(x)}{dx} \right|$$

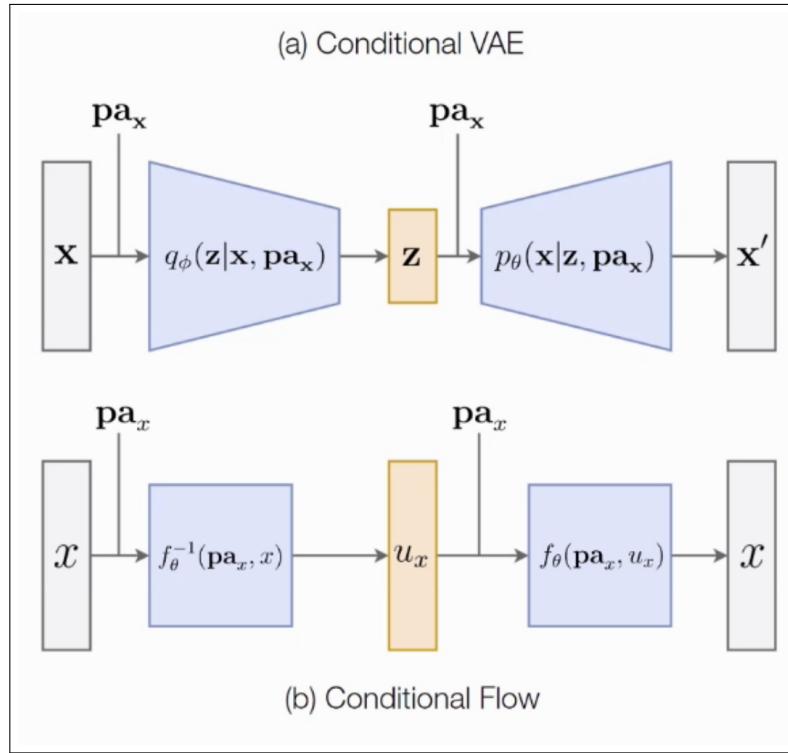
- Can condition on parents via a learned parameterised function:

$$\log p(x | \mathbf{pa}_x) = \log p(f_\theta^{-1}(\mathbf{pa}_x, x)) + \log \left| \det \left( \frac{\partial f_\theta^{-1}(\mathbf{pa}_x, x)}{\partial x} \right) \right|$$

- In multivariate settings, we need to compute the determinant of a Jacobian matrix, which can become pretty expensive!

**Figure 18:** This concept is what we use to train the normalizing clause. We take the log of it and use that as the training objective.

### 1.8.2 VAEs

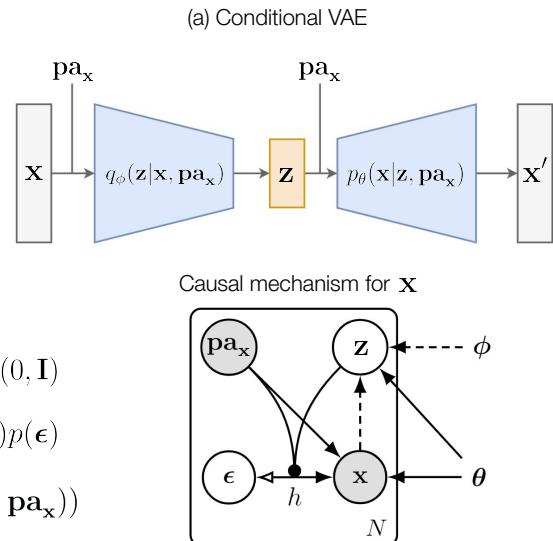


**Figure 19:** “conditional vae and the conditional flow look the same. The catch here, is that the attention step is now non-deterministic because the encoder produces a distribution of all these encodings and the decoder also produces a distribution over pixel values.”

## Deep Mechanisms: Variational Autoencoders

- Well suited for modelling structured variable mechanisms, e.g. for images
- Caveat: **abduction** is non-deterministic
- Causal mechanism:  

$$\begin{aligned} \mathbf{x} &:= f_\theta(\mathbf{pa}_x, \mathbf{u}_x) = h(\epsilon; g_\theta(\mathbf{z}, \mathbf{pa}_x)) \\ &= \mu(\mathbf{z}, \mathbf{pa}_x) + \sigma(\mathbf{z}, \mathbf{pa}_x) \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \end{aligned}$$
- Factored exogenous noise:  $p(\mathbf{u}_x) = p_\theta(\mathbf{z})p(\epsilon)$
- $$p(\mathbf{u}_x|\mathbf{x}, \mathbf{pa}_x) \approx q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{pa}_x)\delta(\epsilon - h^{-1}(\mathbf{x}; g_\theta(\mathbf{z}, \mathbf{pa}_x)))$$



**Figure 20:** “We can define the causal mechanism for an image  $X$ . The parents of  $X$  here are meta information like age etc.” We have  $g$  that is decoder model that predicts the per pixel mean and variances. and the  $H$  mechanism is how we sample from this distribution. The encoder is approximately invertible. Therefore, we have a factorized exogenous noise decomposition. To calculate the posterior it is approximated by the encoder mapping and the inversion of the deterministic sampling procedure

## Deep Mechanisms: Variational Autoencoders

→ Example: computing counterfactuals:

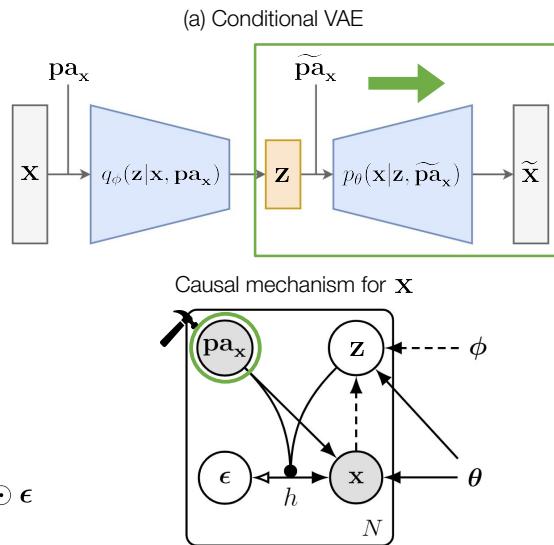
1. **Abduction:**  $z \sim q_\phi(z | x, pa_x)$

$$\epsilon = h^{-1}(x; g_\theta(z, pa_x)) = \frac{x - \mu(z, pa_x)}{\sigma(z, pa_x)}$$

2. **Action:**  $do(pa_x := \tilde{pa}_x)$

3. **Prediction:**  $\tilde{x} \sim p_\theta(\tilde{x} | z, \tilde{pa}_x)$

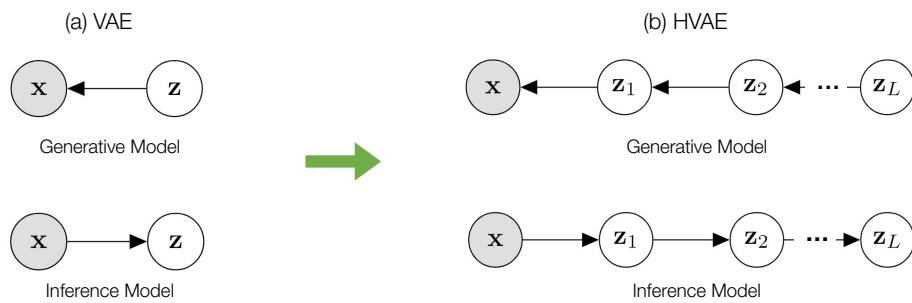
$$\tilde{x} := h(\epsilon; g_\theta(z, \tilde{pa}_x)) = \mu(z, \tilde{pa}_x) + \sigma(z, \tilde{pa}_x) \odot \epsilon$$



**Figure 21:** 1) encoding: pass it through encoder and receive  $z$ . Then decode. after you invert the sampling in the observation space so given its an image  $X$  we subtract the predicted mean and divide by the predicted variance per pixel. 2) next step is to perform some intervention upstream. e.g. if we have a causal graph with various dependencies between the parents, if we intervene on one of them and compute the value of the new parents, (the  $\tilde{z}$  values) and set these to 0 values. 3) in the final step we pass the new values through the decoder and plug values of epsilon and  $z$  that we abducted in the abduction step as well as the new values of the parents to condition with the decoder. *This applies to, if we want to see what the image would look like if we had a different age.*

## Deep Mechanisms: Hierarchical VAEs

- To produce high-fidelity image counterfactuals, we require a powerful generative model that is amenable to principled abduction
- Hierarchical VAEs extend VAEs to multiple layers of latent variables:



**Figure 22:** This is not scalable to high resolutions, so we need a more powerful generative model that is still amenable to principal deduction ut can also scale up to higher resolutions. A natural solution is the hierarchical vae; here we have  $L$  latent variables instead of just one.

## Deep Mechanisms: Hierarchical VAEs

- To produce high-fidelity **image counterfactuals**, we require a powerful generative model that is amenable to principled abduction
- Hierarchical VAEs extend VAEs to multiple layers of latent variables:

$$p(\mathbf{x}, \mathbf{z}_{1:L}) = p(\mathbf{x} \mid \mathbf{z}_{1:L}) p(\mathbf{z}_L) \prod_{i=1}^{L-1} p(\mathbf{z}_i \mid \mathbf{z}_{>i})$$

Generative Model

- **Goal:** Optimize our model  $p$  to be close to a given data distribution  $p_{\text{data}}$

$$\log p(\mathbf{x}) \geq \mathbb{E}_{q(\mathbf{z}_{1:L} \mid \mathbf{x})} [\log p(\mathbf{x} \mid \mathbf{z}_{1:L})] - D_{\text{KL}}(q(\mathbf{z}_{1:L} \mid \mathbf{x}) \parallel p(\mathbf{z}_{1:L})) =: \text{ELBO}(\mathbf{x})$$

**Figure 23:** The goal is similar; optimize model to be close to a given data distribution. The only difference from the VAE objective is that we now have a multitude of latent variables and factorize across the  $L$  layers.

## Literature: Deep

### Ladder Va

## Improved Variational Inference with Inverse Autoregressive Flow

### NVAE: A Deep Hiera

Ar  
farahd

Normalizing flows, autoregressive deep energy-based models are for deep generative learning, and have recently outperformed by other methods. However, there are challenges, we explore the techniques for hierarchical VAEs and how they can be used for batch normalization. NVAE is a state-of-the-art model which is NVAE achieves state-of-the-art models on the MNIST, CIFAR-10, and CelebA datasets. The state-of-the-art from 2.98 to 2.95. The figures on CelebA VAE show the qualitative results of VAEs. The source code is available at [here](#).



Casper Kaae Sønderby\*  
casperkaae@gmail.com

Diederik P. Kingma  
dpkingma@openai.com

Tim Salimans  
tim@openai.com

Rafal Jozefowicz  
rafal@openai.com

ZE AUTOREGRESSIVE  
M THEM ON IMAGES

Søren Kaae Sønd  
skaaesonderby@gmail.com

Ilya Sutskever  
ilya@openai.com

Max Welling\*  
M.Welling@uva.nl

ine, generates samples quickly  
all natural image benchmarks,  
usually represent autoregressive  
t, when made sufficiently deep,  
they perform well in log-  
by training a VAE with IAF, then  
using it CIFAR-10, ImageNet,  
very deep VAEs achieve higher  
stochastic losses for both  
Qualitative studies suggest that  
real representations. We release  
[ib.com/openai/vdvae](#).

### Abstract

Variational autoencoders are p  
deep models with several lay  
train which limits the improve  
We propose a new inference  
recursively corrects the gener  
likelihood in a process resen  
show that this model provides  
log-likelihood lower bound co  
Variational Autoencoders an  
analysis of the learned hierar  
inference model is qualitativ  
hierarchy of latent variables.  
deterministic warm-up (gradually turning on the KL-term) are crucial for training  
variational models with many stochastic layers.

The framework of normalizing flows provides a general strategy for flexible variational inference of posteriors over latent variables. We propose a new type of normalizing flow, inverse autoregressive flow (IAF), that, in contrast to earlier published flows, scales well to high-dimensional latent spaces. The proposed flow consists of a chain of invertible transformations, where each transformation is based on an autoregressive neural network. In experiments, we show that IAF significantly improves upon diagonal Gaussian approximate posteriors. In addition, we demonstrate that a novel type of variational autoencoder, coupled with IAF, is competitive with neural autoregressive models in terms of attained log-likelihood on natural images, while allowing significantly faster synthesis.



Low resolution → High resolution

**Figure 24:** caption

## Quick Detour: Connection to Diffusion Models

→ A Diffusion Model is a type of Hierarchical VAE where:

- I. The encoder  $q(\mathbf{z}_{1:T} \mid \mathbf{x})$  is fixed rather than learned
- II. Each latent variable  $\mathbf{z}_t$  has the same dimensionality as  $\mathbf{x}$
- III. A single (denoising) model is shared amongst all layers in the hierarchy

$$-\text{ELBO}(\mathbf{x}) = \mathbb{E}_{q(\mathbf{z}_{1:T} \mid \mathbf{x})} [-\log p(\mathbf{x} \mid \mathbf{z}_{1:T})] + D_{\text{KL}}(q(\mathbf{z}_{1:T} \mid \mathbf{x}) \parallel p(\mathbf{z}_{1:T}))$$

$$= \mathbb{E}_{q(\mathbf{z}_1 \mid \mathbf{x})} [-\log p(\mathbf{x} \mid \mathbf{z}_1)] + D_{\text{KL}}(q(\mathbf{z}_T \mid \mathbf{x}) \parallel p(\mathbf{z}_T)) + \boxed{\mathcal{L}_T(\mathbf{x})} \text{ diffusion loss}$$

$$\mathcal{L}_T(\mathbf{x}) = \sum_{t=2}^T \mathbb{E}_{q(\mathbf{z}_t \mid \mathbf{x})} [D_{\text{KL}}(q(\mathbf{z}_{t-1} \mid \mathbf{z}_t, \mathbf{x}) \parallel p(\mathbf{z}_{t-1} \mid \mathbf{z}_t))] = \boxed{\frac{T}{2} \mathbb{E}_{\epsilon \sim \mathcal{N}(0, I), t \sim U\{1, T\}} [\mathbf{w}(t) \parallel \epsilon - \hat{\epsilon}_\theta(\mathbf{z}_t; t)]_2^2}$$

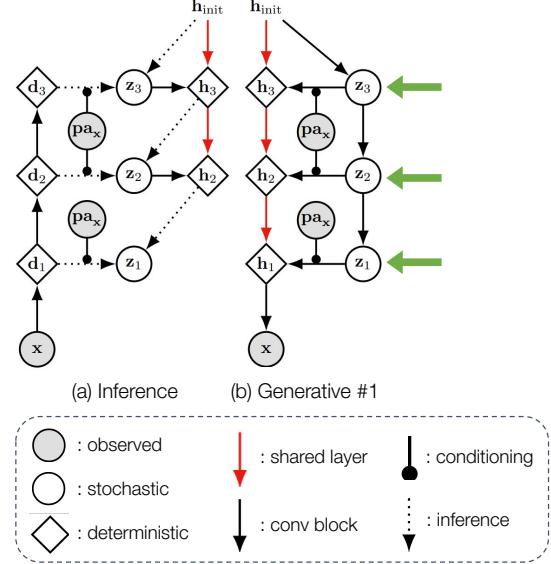
Kingma, Diederik, et al. Variational diffusion models. NeurIPS 2021; De Sousa Ribeiro, F. & Glocker, B. (2024). Demystifying Variational Diffusion Models. arXiv preprint arXiv:2401.06281.

**Figure 25:** boxed term is MSE loss. point being here is that it is a hierarchical VAE with a different inference model.

## Deep Mechanisms: Conditional HVAEs

→ Generative model structures:

1. Exogenous Prior:  $p_\theta(\mathbf{z}_{1:L})$

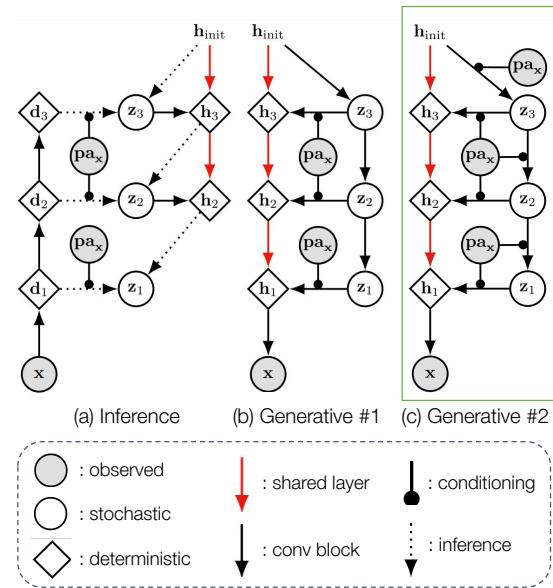


**Figure 26:** “We had to figure out a couple of ways to condition these VAEs. The first, is to retrain the role of the latent variable as being part of the exogenous noise for  $X$ . in this case, the latent variables are independent of the values of the parents. So when inferring these latent variables they won’t get access to the values of the parents.’

## Deep Mechanisms: Conditional HVAEs

→ Generative model structures:

1. Exogenous Prior:  $p_\theta(\mathbf{z}_{1:L})$
2. Conditional Prior:  $p_\theta(\mathbf{z}_{1:L} | \mathbf{pa}_x)$



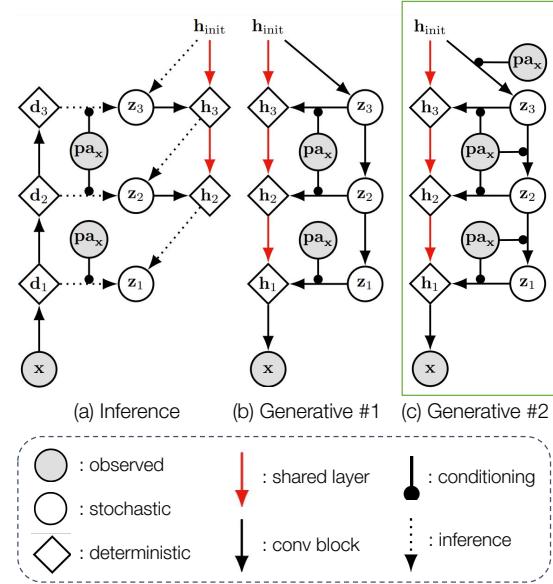
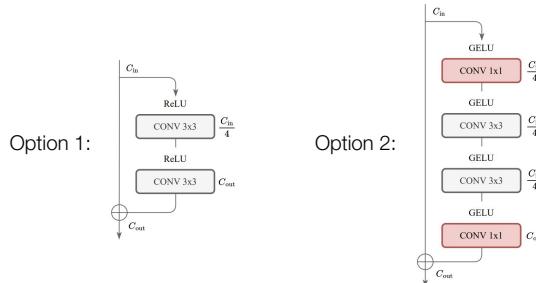
**Figure 27:** The second option is to include a conditional prior, so the latent variables do see the values of the parents.

## Deep Mechanisms: Conditional HVAEs

→ Generative model structures:

1. Exogenous Prior:  $p_\theta(\mathbf{z}_{1:L})$
2. Conditional Prior:  $p_\theta(\mathbf{z}_{1:L} | \mathbf{pa}_x)$

→ Standard residual blocks:

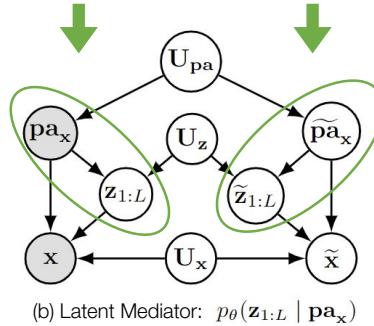
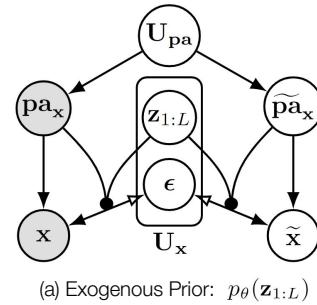


**Figure 28:** The way you construct this is by stacking residual blueing blocks into a u-net style architecture. These should then be trained to predict the mean and variabce of the latent variables.

## Latent Mediator Model

- The conditional prior  $p_\theta(\mathbf{z}_{1:L} | \mathbf{pa}_x)$  induces a latent mediator, as  $\mathbf{z}_{1:L}$  is no longer exogenous
- Nonetheless, the underlying SCM has a Markovian interpretation:

$$p(\mathbf{U}) = p(\mathbf{U}_x) \left( \prod_{k=1}^K p(\mathbf{U}_{\mathbf{pa}_k}) \right) \left( \prod_{i=1}^L p(\mathbf{U}_{\mathbf{z}_i}) \right)$$



**Figure 29:** using this conditional prior as an implication induces a latent mediator. These latent variables are no longer part of the exogenous noise, this is because they are caused by things within the model (there is a directed arrow from the parents to the latent variables now). This is still Markovian because the way you give rise to these latent variables still depends on some external independent noise when you sample them.

In these diagrams, the grey nodes are observed variables, the middle are exogenous noise  $u$  terms and on the right we have unobserved counterfactual values. The key thing is that the exogenous noise is shared between the factual observed and counterfactual branches.

## Latent Mediator Model

### Causal Mediation Analysis

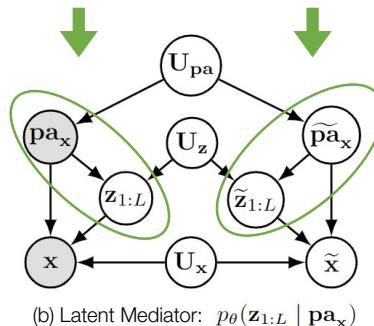
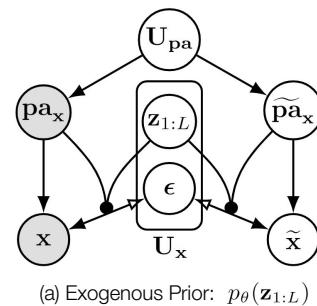
The study of how a treatment effect is mediated by another variable, to help explain why or how an individual may respond to certain stimulus.

- Enables estimation of Direct (DE), Indirect (IE) and Total (TE) causal effects:

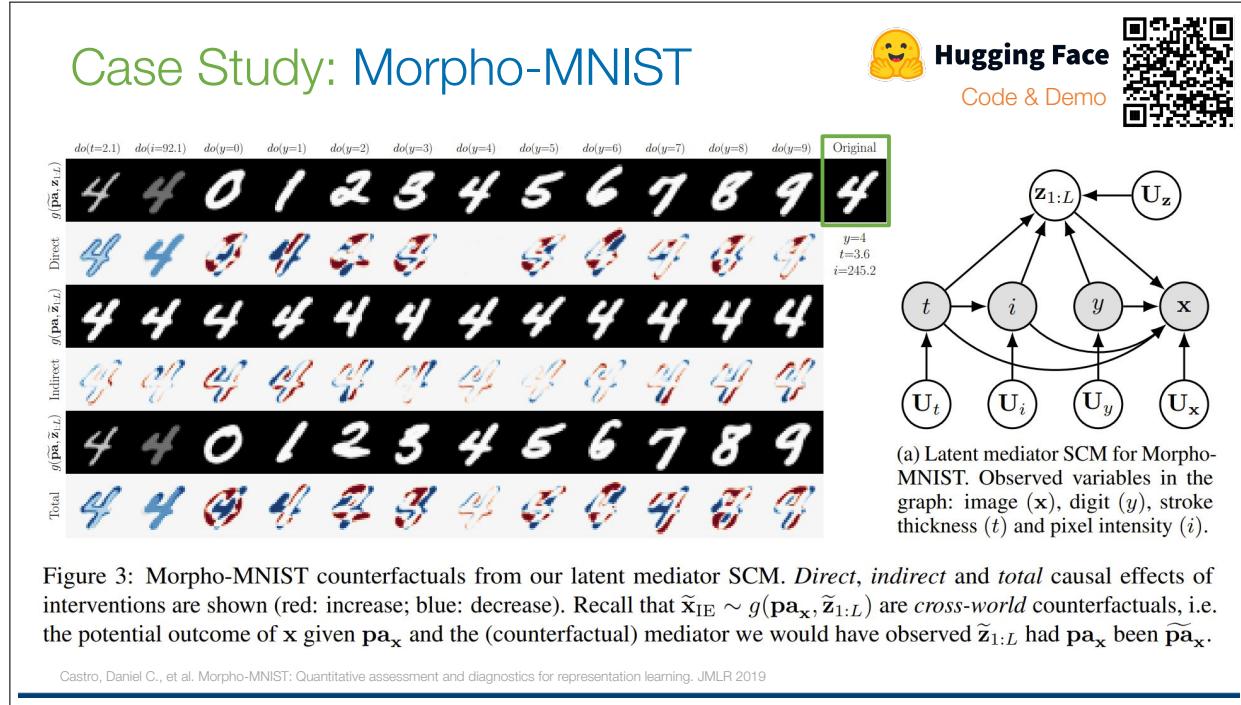
$$DE_x(\tilde{\mathbf{pa}}_x) = \mathbb{E}[g_\theta(\tilde{\mathbf{pa}}_x, \mathbf{z}_{1:L}) - g_\theta(\mathbf{pa}_x, \mathbf{z}_{1:L})]$$

$$IE_x(\tilde{\mathbf{z}}_{1:L}) = \mathbb{E}[g_\theta(\mathbf{pa}_x, \tilde{\mathbf{z}}_{1:L}) - g_\theta(\mathbf{pa}_x, \mathbf{z}_{1:L})]$$

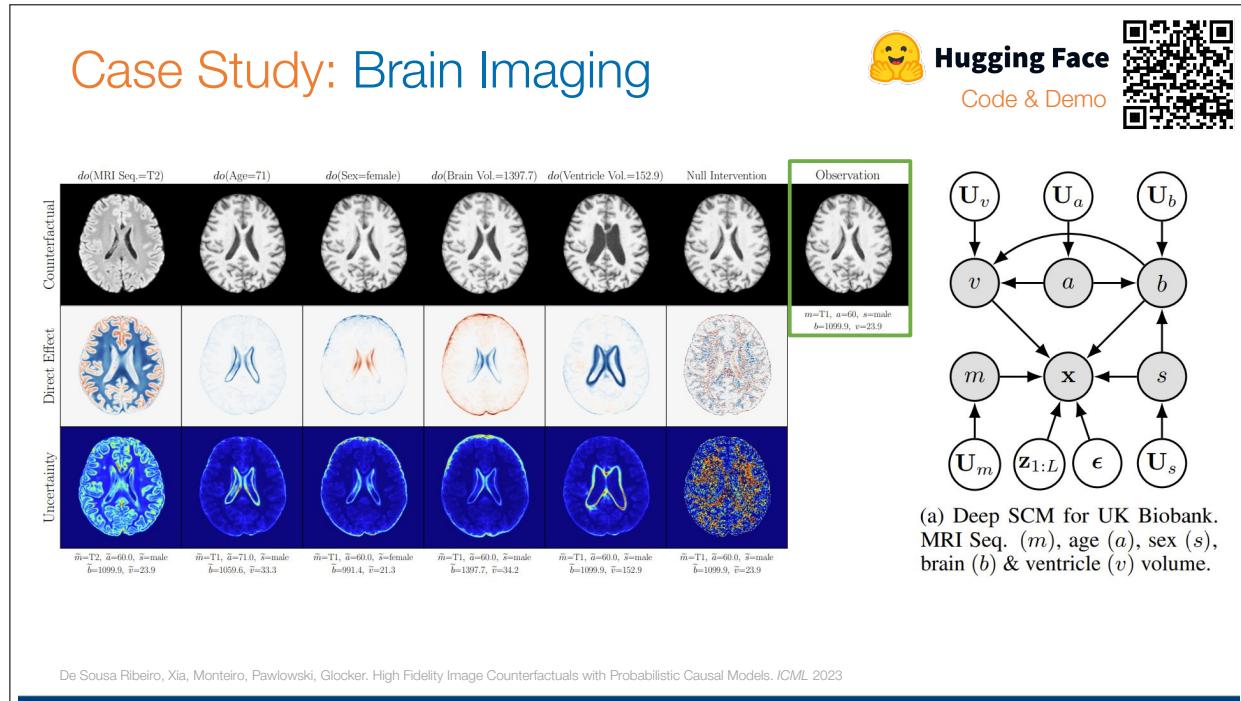
$$TE_x(\tilde{\mathbf{pa}}_x, \tilde{\mathbf{z}}_{1:L}) = \mathbb{E}[g_\theta(\tilde{\mathbf{pa}}_x, \tilde{\mathbf{z}}_{1:L}) - g_\theta(\mathbf{pa}_x, \mathbf{z}_{1:L})]$$



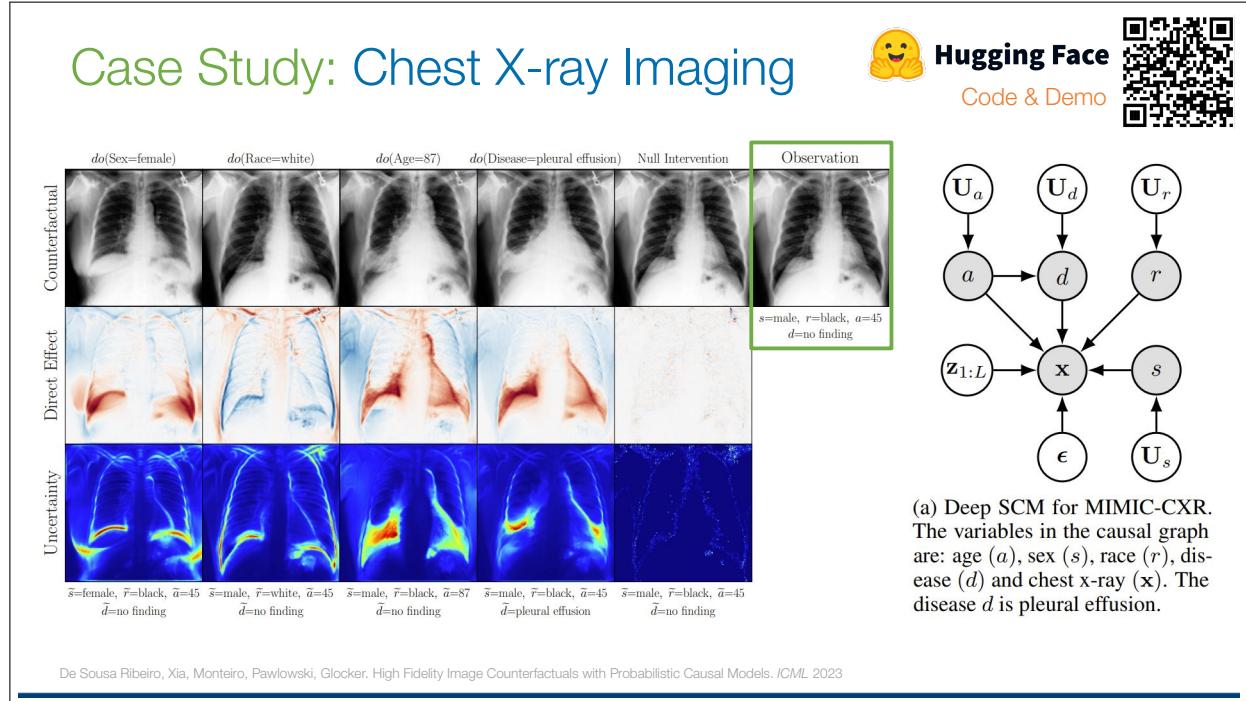
**Figure 30:** this allows you to apply causal mediation analysis defintion above. the main takeaway, is how the parents affect the image vs how the latent variables affect the image. We can visualise this and understand what our interventions are doing when we're chanign things and confusing counterfactuals.



**Figure 31:** Here we have 4 observed variables and a predefined causal graph. Each column is a different intervention; so we can change the digits, the thickness and the intensity. We can see that when we change the thickness, it obeys the causal graph (thickness causes intensity but the same is not true the other way around).



**Figure 32:** Similar example, we start with a medically assumed causal graph with 6 observed variables.  $X$  is one of the variables being the image, we also have ventricle volume, age, brain volume, MRI sequencing  $M$ , sex of patient  $S$ . Each column here is an intervention on a different attribute. If we change nothing we get a reconstruction back out. Idea here is we're not generating different brains, we generating this particular brain with different attributes.



**Figure 33:** Here we have a different assumed causal graph, but once again it was medically informed.

## 1.9 Evaluating Counterfactual

### Evaluating Counterfactuals: Axiomatic Properties

- The **soundness theorem** states that the properties of **composition**, **effectiveness**, and **reversibility** are necessary in all causal models (Galles & Pearl, 1998).
- The **completeness theorem** states that these properties are sufficient (Halpern, 1998).
- We can measure **counterfactual soundness** using these axiomatic properties

Published as a conference paper at ICLR 2023

#### MEASURING AXIOMATIC SOUNDNESS OF COUNTERFACTUAL IMAGE MODELS

Miguel Monteiro<sup>1†</sup>, Fabio De Sousa Ribeiro<sup>1†</sup>, Nick Pawlowski<sup>2</sup>, Daniel C. Castro<sup>1,2</sup>, Ben Glocker<sup>1</sup>  
<sup>1</sup>Imperial College London, <sup>2</sup>Microsoft Research Cambridge. <sup>†</sup>Joint first authors  
(miguel.monteiro, f.de-sousa-ribeiro, b.glocker)@imperial.ac.uk

#### ABSTRACT

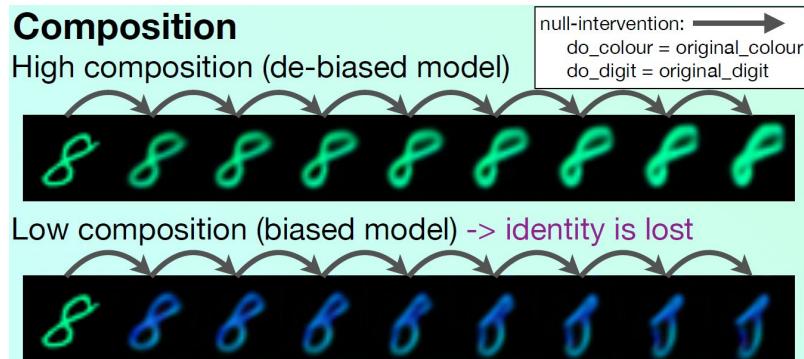
We present a general framework for evaluating image counterfactuals. The power and flexibility of deep generative models make them valuable tools for learning mechanisms in structural causal models. However, their flexibility makes counterfactual identifiability impossible in the general case. Motivated by these issues, we revisit Pearl's axiomatic definition of counterfactuals to determine the necessary constraints on any counterfactual inference model: *composition*, *reversibility*, and *effectiveness*. We frame counterfactuals as functions of an input variable, its parents, and a function of the causal graph. This allows us to approximate these functions that could represent the counterfactual, thus deriving distance metrics between the approximate and ideal functions. We demonstrate how these metrics can be used to compare and choose between different approximate counterfactual inference models and to provide insight into a model's shortcomings and trade-offs.

David Galles and Judea Pearl. An axiomatic characterization of causal counterfactuals. Foundations of Science, 3:151–182, 1998.  
Joseph Y. Halpern. Axiomatizing causal reasoning. In Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, UAI'98, San Francisco, CA, USA, 1998.

**Figure 34:** How do we evaluate these counterfactuals? There are a couple theorems above.

### 1.9.1 Composition Axiom

#### Example: Composition Axiom



**Definition.** Intervening on a variable to have the value it would otherwise have without the intervention will not affect other variables.

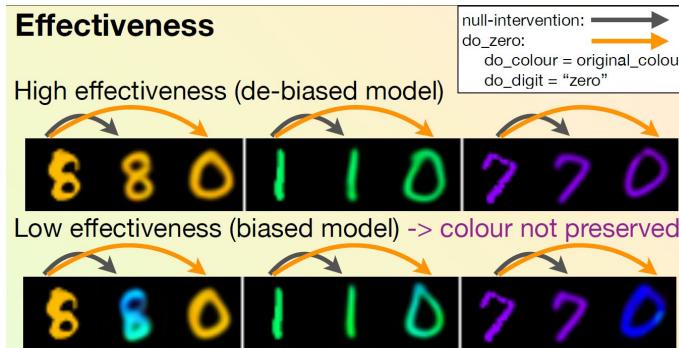
→ Implies the existence of a null-intervention

**Figure 35:** Definition above, we can see an example above with high composition; a de-biased model. If we continuously apply the null intervention iteratively and try to reconstruct the image, if the identities are well preserved across these iterated applications mechanisms, then we know that the model has high composition.

Conversely, if it has low composition the identity is lost.

### 1.9.2 Effectiveness Axiom

#### Example: Effectiveness Axiom



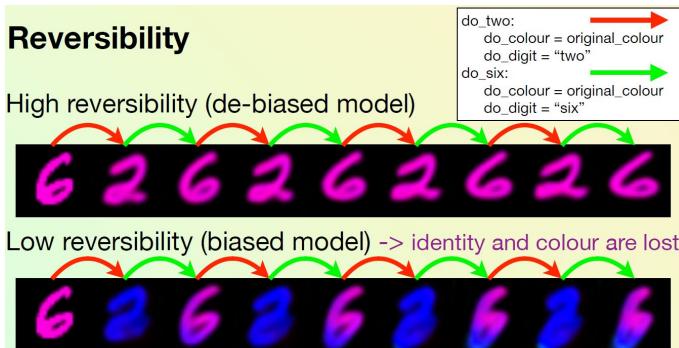
**Definition.** Intervening on a variable to have a specific value will cause the variable to take on that value.

→ Caveat: often relies on a **pseudo-oracle**, e.g. a classifier/regressor

**Figure 36:** Definition above. An example with high effectiveness, if we want to change digit identity it will preserve colour, if we want to change colour then it should preserve digit identity. So a model with low effectiveness will be biased.

### 1.9.3 Reversibility Axiom

#### Example: Reversibility Axiom



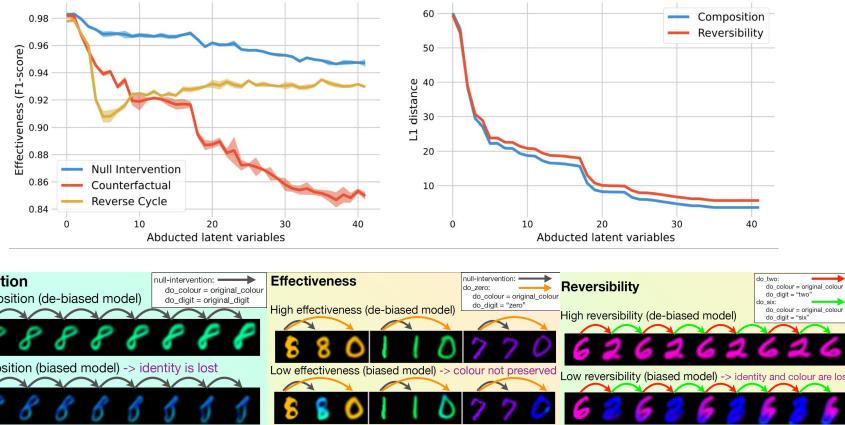
**Definition.** Precludes multiple solutions due to feedback loops. If setting a variable  $X$  to a value  $x$  results in a value  $y$  for a variable  $Y$ , and setting  $Y$  to a value  $y$  results in value  $x$  for  $X$ , then  $X$  and  $Y$  will naturally take on the values  $x$  and  $y$ .

→ Follows directly from **composition** in recursive systems such as DAGs

**Figure 37:** It has to do with cyclic consistency. A model with good reversibility will be able to do this, otherwise its color information will be lost. The reversibility follows from the composition axiom if you use a directed acyclic graph.

## Evaluating Counterfactuals: Axiomatic Properties

→ **TLDR:** We identify an inherent trade-off



Monteiro, De Sousa Ribeiro, Pawłowski, Castro, Glocker. Measuring Axiomatic Soundness of Counterfactual Image Models. ICLR 2023  
De Sousa Ribeiro, Xia, Monteiro, Pawłowski, Glocker. High Fidelity Image Counterfactuals with Probabilistic Causal Models. ICML 2023

**Figure 38:** We identified an inherent trade-off between composition and effectiveness – when you train counterfactual inference models you want to be good at all of these things, but there seems to be a trade off between composition and effectiveness specifically.

## Conclusion & Outlook

- Deep SCMs can generate plausible high-fidelity **counterfactuals** of medical images as measured by **axiomatic soundness** of counterfactuals
- Tractable estimation of **direct**, **indirect** and **total** causal effects for high-dimensional structured variables
- **Limitations:**
  - ◆ Only consider Markovian SCMs; although Markovianity is a common assumption in causality literature, it is strong in most cases
  - ◆ Measuring counterfactual effectiveness relies on separately trained classifiers

**Figure 39:** caption

## Conclusion & Outlook

- Deep SCMs can generate plausible high-fidelity **counterfactuals** of medical images as measured by **axiomatic soundness** of counterfactuals

### Future Work

- I. Targeted data augmentation to improved robustness, sample efficiency & fairness
- II. Providing principled **causal explanations**, e.g. through mediation analysis
- III. Improving **counterfactual soundness**, via ML techniques and/or medical guidance
- IV. Provide theoretical guarantees of identifiability under plausible causal assumptions
- V. Extensions to Semi-Markovian and Non-Markovian settings

**Figure 40:** caption

## References

- [1] Daniel C Castro, Ian Walker, and Ben Glocker. “Causality matters in medical imaging”. en. In: *Nat Commun* 11.1 (July 2020), p. 3673.