Imperial College London

BioMedIA

# Measuring axiomatic soundness of counterfactual image models

Miguel Monteiro, Fabio De Sousa Ribeiro, Nick Pawlowski, Daniel C. Castro, Ben Glocker

project mira ERC StG 2017 grant No 757173

erc European Research Council

## Motivation

- Counterfactuals are useful for explainability, interpretability, fairness and data-augmentation;
- For images, deep generative models are essential to estimate mechanisms;
- In the general case, model identifiability is impossible for deep models;
- Many approaches have been proposed for approximate counterfactual inference. Less work has been done on evaluation.

## Axiomatic characterisation

- The **soundness theorem** states that the properties of **composition, effectiveness and reversibility** are necessary in all causal models. The **completeness theorem** states that these properties are sufficient.

These properties **can be measured** for approximate counterfactual models in order to **compare and evaluate** them.

## Defining and measuring soundness:

1. **Composition**: 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**. Measured using distance metrics (*e.g.* **l2** distance);
2. **Effectiveness**: intervening on a variable to have a specific value will cause the variable to take that value. Measured using auxiliary classifiers or regressors;
3. **Reversibility**: Informally, it prohibits the existence of feedback loops. Measured using distance metrics.

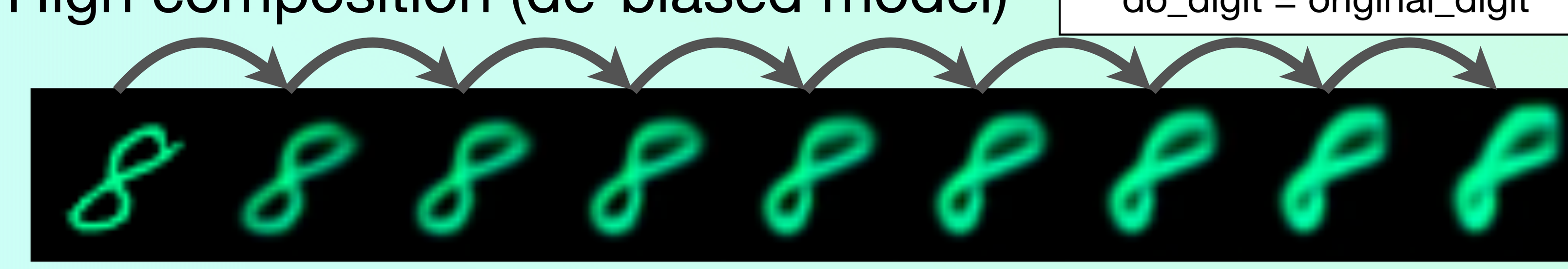
## Example: VAE trained on colour MNSIT

- Digit and colour are confounded in the training set;
- One model is trained using a simulated intervention to de-bias the data, one is not;

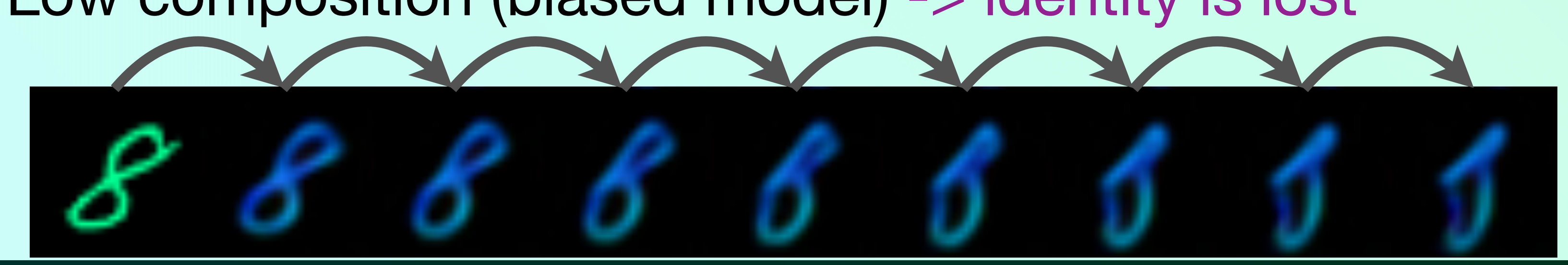
### Composition

null-intervention:
do_colour = original_colour
do_digit = original_digit

High composition (de-biased model)



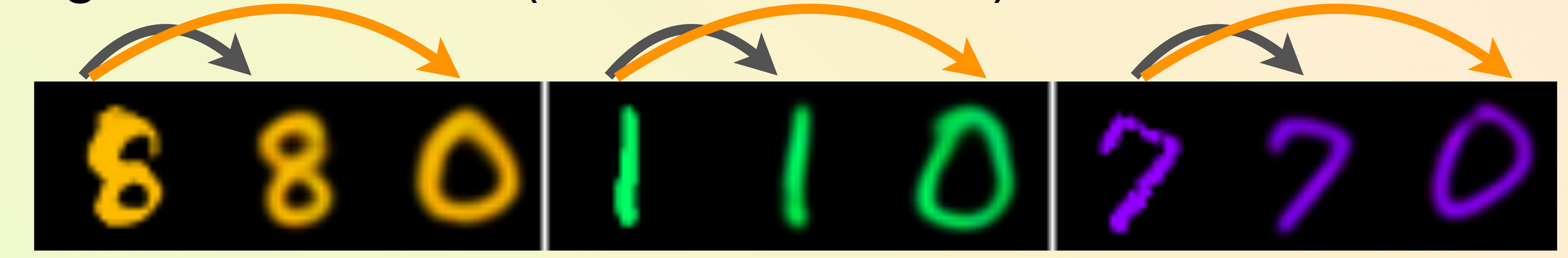
Low composition (biased model) -> identity is lost



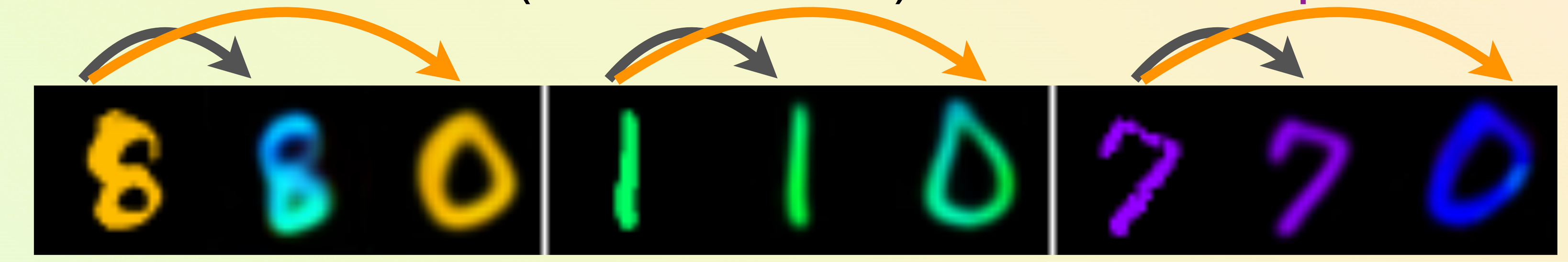
### Effectiveness

null-intervention:
do_zero:
do_colour = original_colour
do_digit = “zero”

High effectiveness (de-biased model)



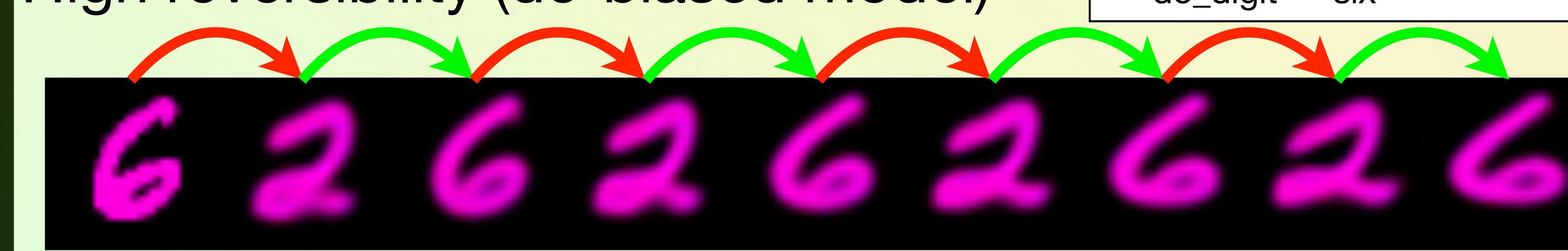
Low effectiveness (biased model) -> colour not preserved



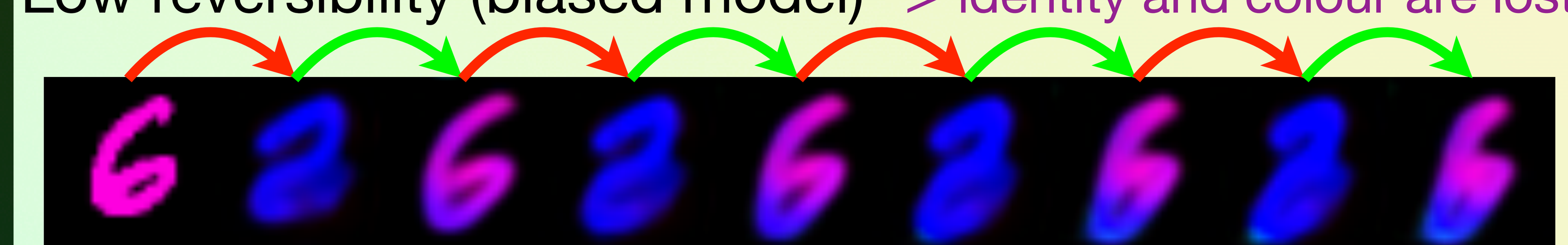
### Reversibility

do_two:
do_colour = original_colour
do_digit = “two”
do_six:
do_colour = original_colour
do_digit = “six”

High reversibility (de-biased model)



Low reversibility (biased model) -> identity and colour are lost



## CELEB-A HQ with hierarchical VAE

- Counterfactuals performed by abducting all the latent variables or only a subset;
- Reveals trade-off between composition and effectiveness: preserving identity vs performing the intervention.

null-intervention:
invert-smiling: