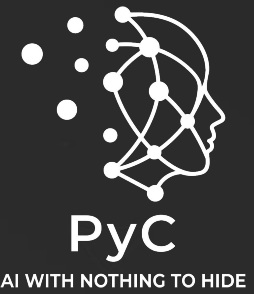
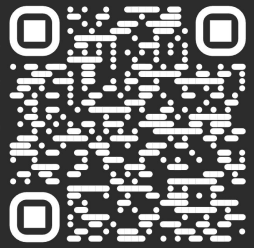
Foundations of Concept-Based Interpretable Deep Learning



Giuseppe Marra & Pietro Barbiero
giuseppe.marra@kuleuven.be pietro.barbiero@ibm.com

In collaboration with: Giovanni de Felice and Mateo Espinosa Zarlenga

ESSAI26



KU LEUVEN IBM Research

fwo Swiss National Science Foundation HASLERSTIFTUNG

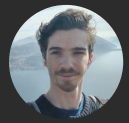
Who are we?

Giuseppe Marra
Assistant Professor
KU Leuven
giuseppe.marra@kuleuven.be

KU LEUVEN

Pietro Barbiero
Swiss Postdoctoral Fellow
IBM Research (Switzerland)
pietro.barbiero@ibm.com

IBM Research



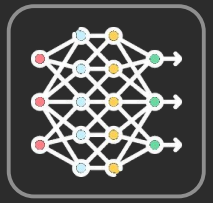
Giovanni De Felice

Università della Svizzera italiana

Mateo Espinosa Zarlenga

UNIVERSITY OF OXFORD

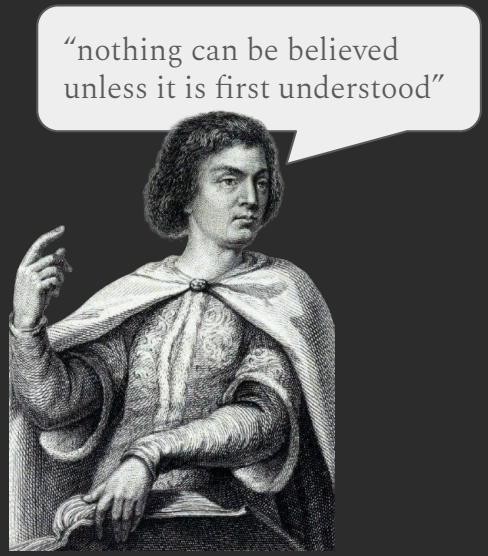
When is f interpretable for the **target user**?



Machine
learning model

Target
user

Approach



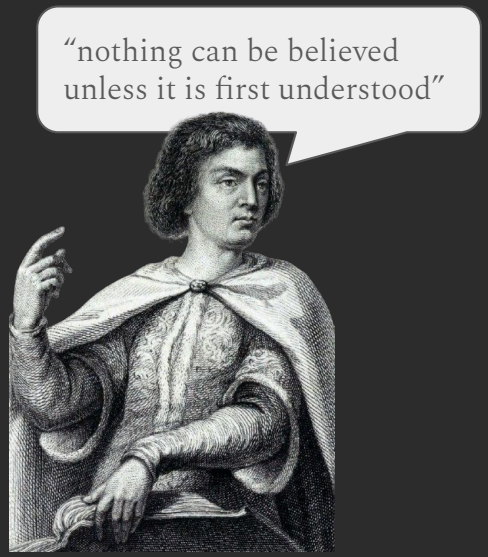
Theory

- **Definitions**
- **Research method**

1

[1] Pierre Abélard, Historia Calamitatum (1132).

Approach



Theory

- Definitions
- Research method

Code

PyC

- Model design
- Benchmarks

Methods

- Models
- Metrics

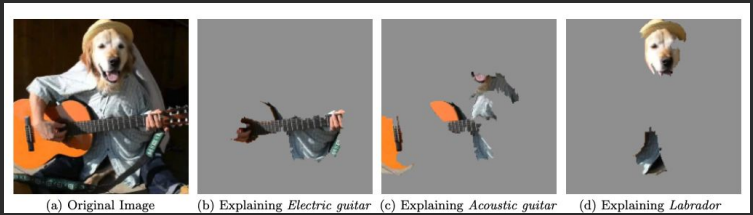
1 → 2 → 3

[1] Pierre Abélard, Historia Calamitatum (1132).

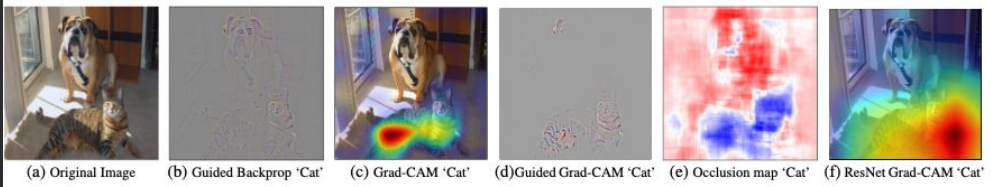
What is this course NOT about?

We will not have time to dive deep into:

1. "Traditional" explainable AI (XAI) methodologies



Example of LIME (taken from [1])



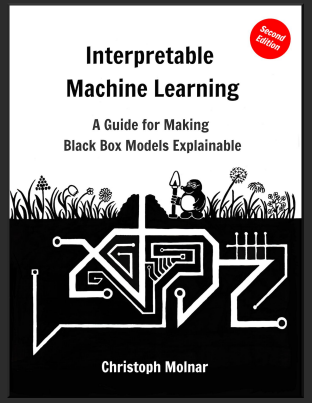
Example of GradCAM and other saliency methods (taken from [2])

[1] Ribeiro et al. "'Why should I trust you?' Explaining the predictions of any classifier." KDD (2016).
[2] Selvaraju et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." ICCV (2017).

What is this course NOT about?

We will not have time to dive deep into:

1. "Traditional" explainable AI (XAI) methodologies



"Interpretable Machine Learning"
Christoph Molnar

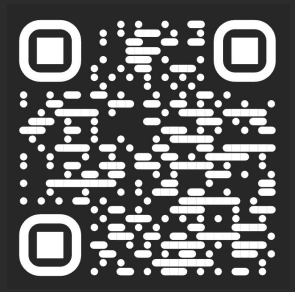
Link to Book

[1] Molnar. Interpretable machine learning. Lulu. com, 2020.

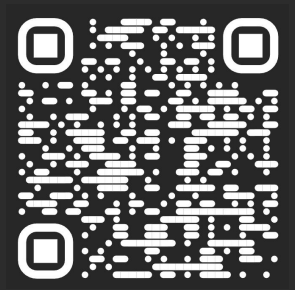
What is this course NOT about?

We will not have time to dive deep into:

1. "Traditional" explainable AI (XAI) methodologies
2. Deep philosophical aspects of explaining models



"The Mythos of Interpretability"
Lipton (2018) [1]



"To Explain or to Predict?"
Shmueli (2018) [2]

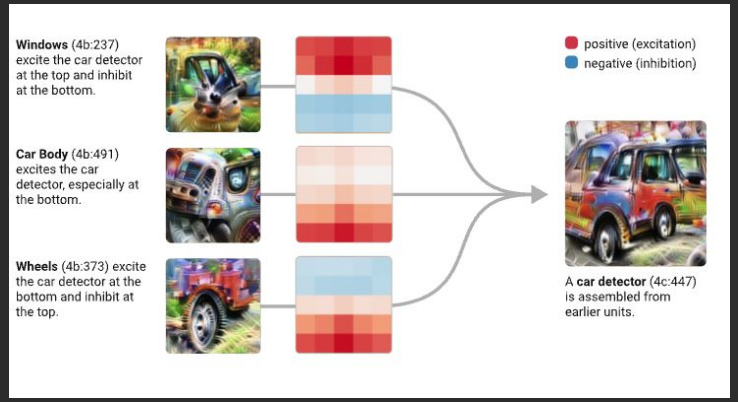
"Explanation Theory"
Bromberger (1992) [3]

[1] Lipton. "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." Queue (2018).
[2] Shmueli. "To Explain or to Predict?." Statist. Sci. 25 (3) 289 - 310, August 2010. https://doi.org/10.1214/10-STS330
[3] Bromberger. On what we know we don't know: Explanation, theory, linguistics, and how questions shape them. University of Chicago Press, 1992.

What is this course NOT about?

We will not have time to dive deep into:

1. "Traditional" explainable AI (XAI) methodologies
2. Deep philosophical aspects of explaining models
3. Post-hoc interpretability such as Mechanistic Interpretability



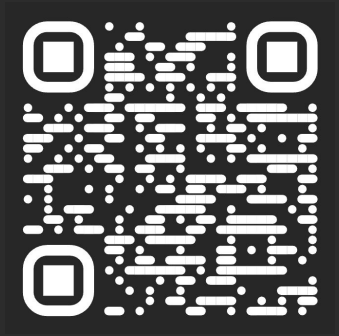
Taken from [1]

[1] Olah et al. "Zoom in: An introduction to circuits." Distill 5.3 (2020): e00024-001.

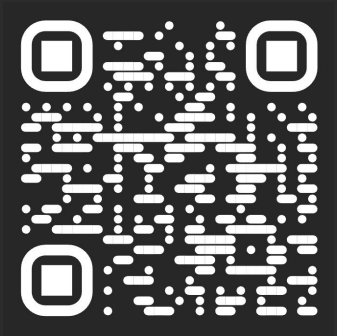
What is this course NOT about?

We will not have time to dive deep into:

1. "Traditional" explainable AI (XAI) methodologies
2. Deep philosophical aspects of explaining models
3. Post-hoc interpretability such as Mechanistic Interpretability



Distill Circuits Thread



Anthropic Circuits Thread

[1] Cammarata et al. "Thread: circuits." Distill 5.3 (2020): e24.
[2] Anthropic "Transformer Circuits Thread" found at https://transformer-circuits.pub/

Course Outline

I. Interpretability theory

II. Blueprint for interpretable models

+ interpretability in PyTorch

III. Concept representations

IV. Concept-based predictors

Mon 2:00 pm – 3:30 pm

Tue 2:00 pm – 3:30 pm

Course Outline

I. Interpretability theory

II. Blueprint for interpretable models

+ interpretability in PyTorch

III. Concept representations

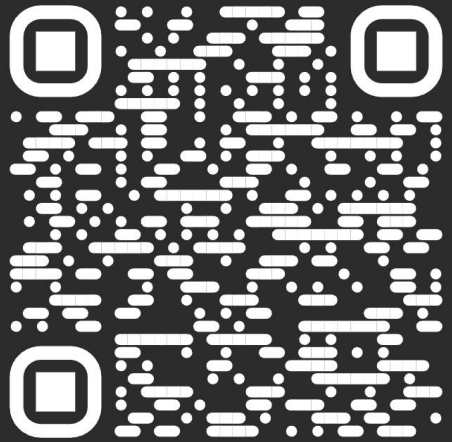
IV. Concept-based predictors

Wed 2:00 pm – 3:30 pm

Fri 2:00 pm – 3:30 pm

Course Website and Materials

This course's slides, schedule, and materials are all available at our website:

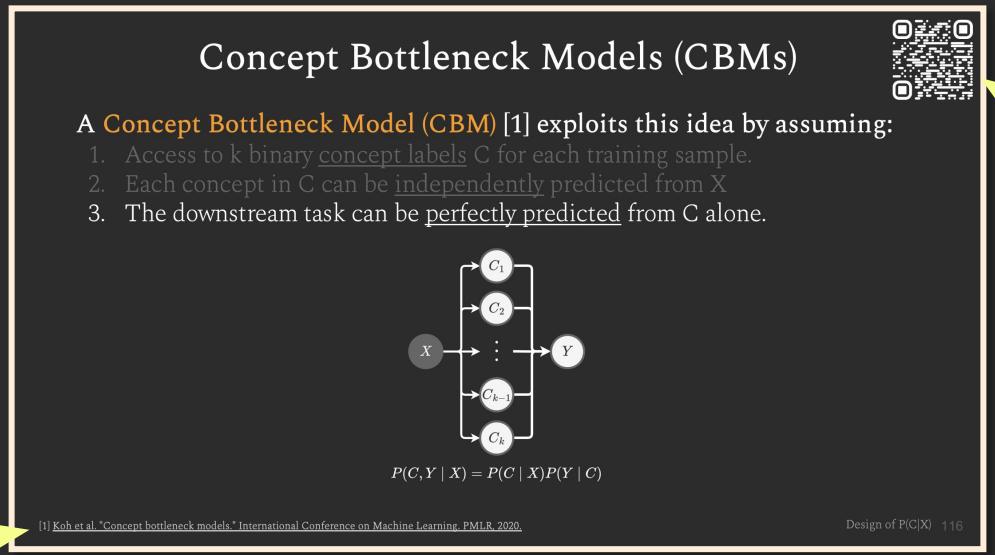


interpretabledeeplearning.github.io

Course Website and Materials

Throughout the course, watch for:

1. QR codes for relevant references and extra material



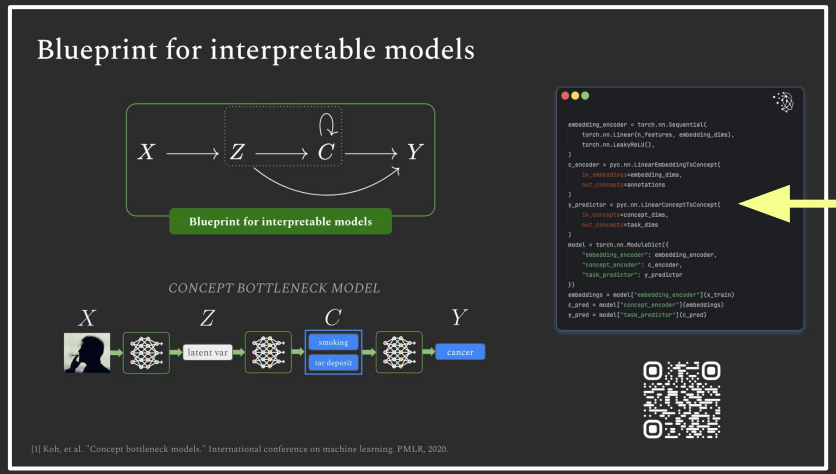
QR code to reference/extra material

Citation + Hyperlink (if you download slides)

Course Website and Materials

Throughout the course, watch for:

1. QR codes for relevant references and extra material
2. Code snippets showing example implementations in PyC



PyC Code Snippets

Course Website and Materials

Throughout the course, watch for:

1. QR codes for relevant references and extra material
2. Code snippets showing example implementations in PyC
3. Discord channel for discussions (or just come & talk to us!)

Interpretability Theory Outline

I. Introduction to interpretability (~15 min)
II. The Standard Interpretable Model (~10 min)
III. Interpretability premises (~5 min)
IV. Interpretability symmetries (~10 min)
V. Interpretability constraints (~10 min)

Q&A + 5 min break

VI. Lagrangian of interpretable machine learning models (~5 min)
V. Trajectories towards interpretability (~10 min)
VI. Interpretable architectures (~10 min)

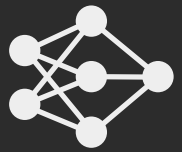
Final Q&A

Introduction to interpretability

Why do we need interpretability?

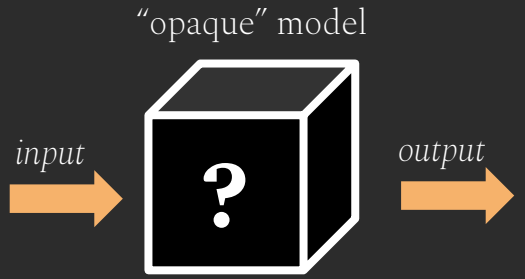
How is interpretability defined?

Why do we need interpretability in the first place?

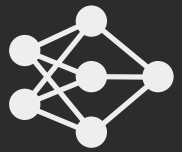


- Highly parametric
- Complicated inference steps
- Non-linearities
- Sensitive to initial states and update rules

≈

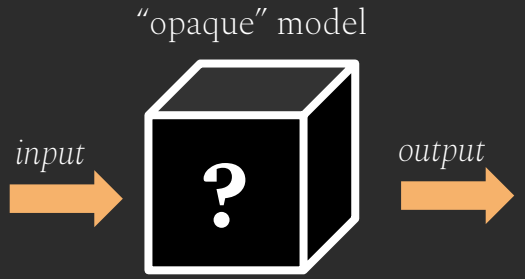


Why do we need interpretability in the first place?

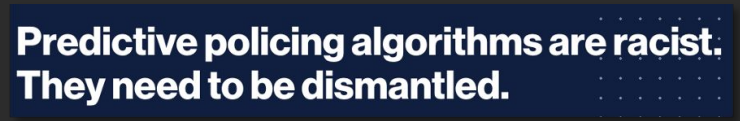


- Highly parametric
- Complicated inference steps
- Non-linearities
- Sensitive to initial states and update rules

≈



1. **Blindly using opaque models** can lead to all sorts of **problems**



[1] Kashmir Hill, "Wrongfully Accused by an Algorithm." The New York Times (2020).
[2] Rachel Goodman, "Why Amazon's Automated Hiring Tool Discriminated Against Women." ACLU (2018).
[3] Will Douglas Heaven, "Predictive policing algorithms are racist. They need to be dismantled." MIT Technology Review (2020).

ChatGPT 5.2

Create an image of a left-handed writer.

Creating image

Ask anything

Gemini

Create an image of a left-handed writer

Loading Nano Banana...

Ask Gemini 3

Fast

ChatGPT 5.2

Create an image of a left-handed writer.

Image created • Left-handed writing at a wooden desk

Ask anything

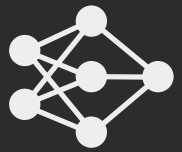
Gemini

Create an image of a left-handed writer

Ask Gemini 3

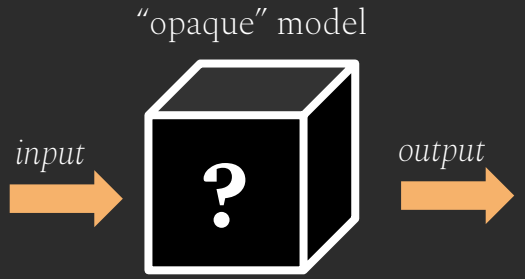
Fast

Why do we need interpretability in the first place?



- Highly parametric
- Complicated inference steps
- Non-linearities
- Sensitive to initial states and update rules

≈



1. **Blindly using opaque models** can lead to all sorts of **problems**
2. **Regulatory/legal** constraints:

General Data Protection Regulations (GDPR, 2016)

"The data subject shall have the right not to be subject to a **decision** based solely on **automated processing**, including profiling,…" (Art. 22)

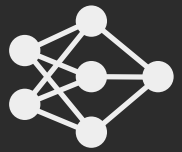
The data subject has the right to "**meaningful information** about the **logic** involved" in the decision. (Art. 13 and 15)

EU AI Act (2024):

"Any affected person subject to a **decision** which is taken by… a **high-risk AI system** … shall have the right to obtain from the deployer **clear and meaningful explanations** (Art. 86)

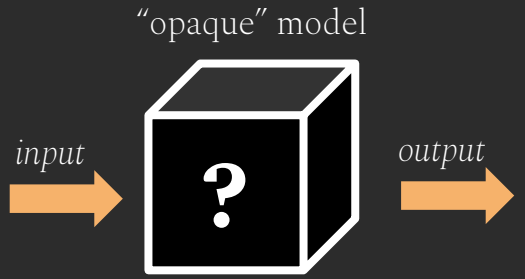
[1] GDPR, EU. "Automated individual decision-making, including profiling." (2022).
[2] Act, EU Artificial Intelligence. "The EU Artificial Intelligence Act." (2024).

Why do we need interpretability in the first place?



- Highly parametric
- Complicated inference steps
- Non-linearities
- Sensitive to initial states and update rules

≈



1. **Blindly using opaque models** can lead to all sorts of **problems**
2. **Regulatory/legal** constraints
3. **Human-AI interactions** are enabled by understanding how a model reasons

But what does interpretability really mean?

Welcome to the Wild West of interpretability terminology

Goebel et al. (2018)

Freiesleben et al. (2023)

Gilpin et al. (2018)

Vilone et al. (2021)

Confalonieri et al. (2020)

Rudin et al. (2019)

Barredo Arrieta et al. (2020)

*"A method is interpretable **if a user can** correctly and efficiently **predict** the method's **results**"* [1]

*"Systems are interpretable **if their operations can be understood** by a human"* [2]

*"Interpretability is the degree to which an observer can **understand the cause of a decision**"* [3]

*"There is **no universal, mathematical definition** of interpretability, and there never will be"* [4]

Informal

Not actionable

[1] Kim et al. "Examples are not enough, learn to criticize! criticism for interpretability" NeurIPS (2016).
[2] Biran et al. "Explanation and Justification in Machine Learning: A Survey" IJCAI workshop (2017).
[3] Miller. "Explanation in artificial intelligence: Insights from the social sciences" Artificial Intelligence (2019).
[4] Murphy. "Probabilistic machine learning: Advanced topics" MIT press (2023).

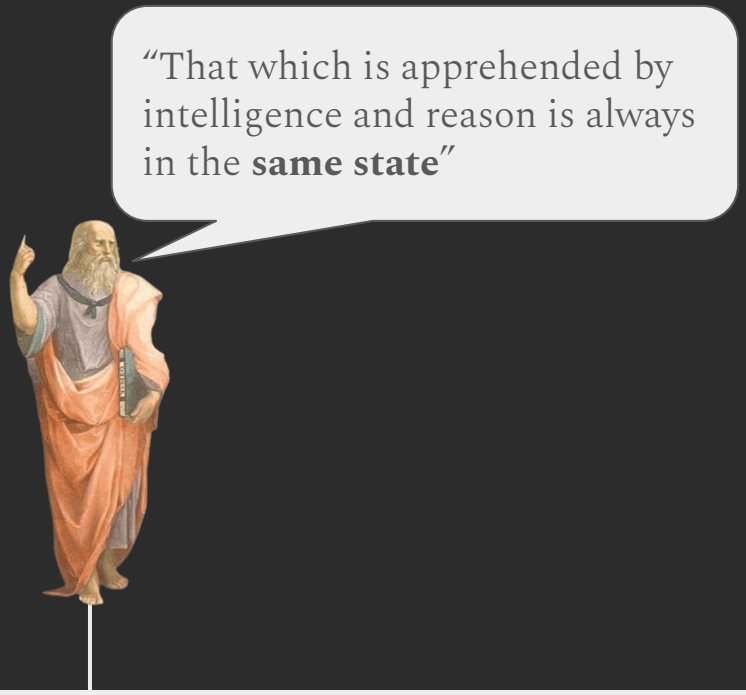
The ultimate question of this lecture, machine learning and everything:

The ultimate question of this lecture, machine learning and everything:

The ultimate question of this lecture, machine learning and everything:

The ultimate question of this lecture, machine learning and everything:

The ultimate question of this lecture, machine learning and everything:



~400 BC

philosophy

[1] Plato, Timaeus.

The ultimate question of this lecture, machine learning and everything:

"That which is apprehended by intelligence and reason is always in the **same state**"

Vergleichende Betrachtungen

über

neuere geometrische Forschungen

von

Dr. Felix Klein,

a. o. Professor der Mathematik an der Universität Erlangen

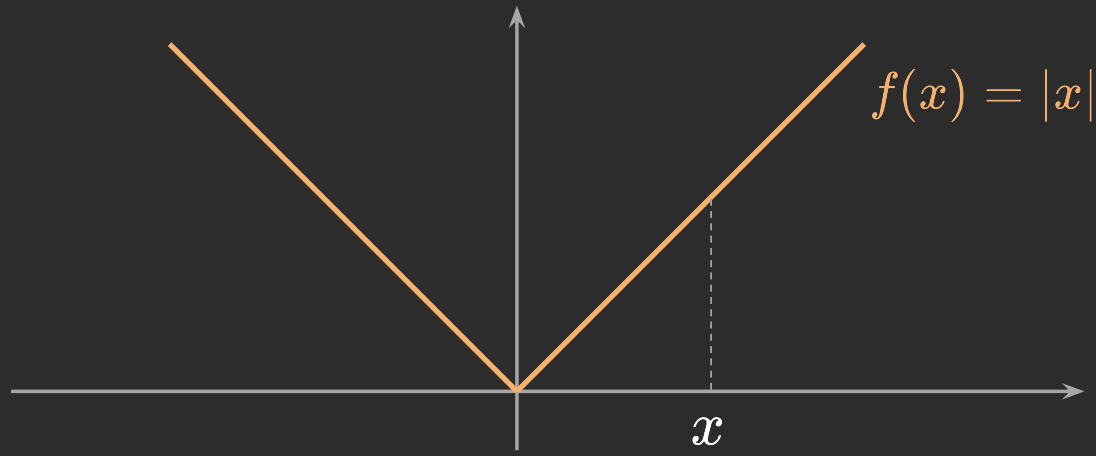
~400 BC

philosophy

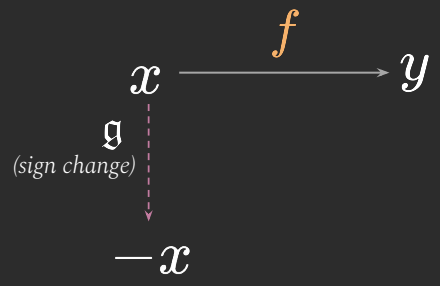
1872

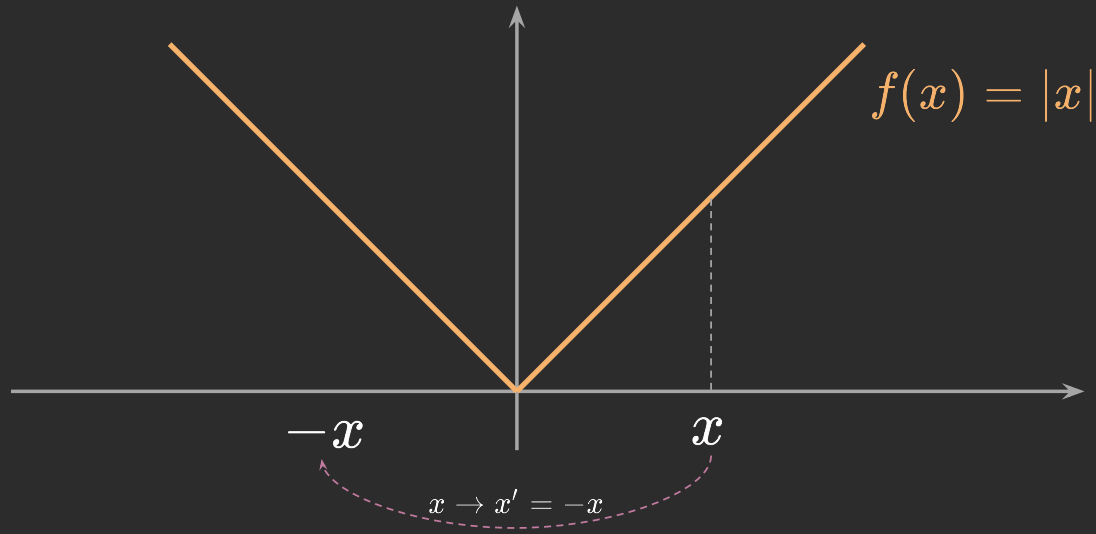
geometry

The ultimate question of this lecture, machine learning and everything:

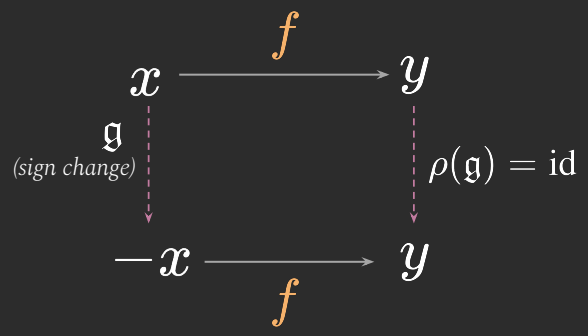


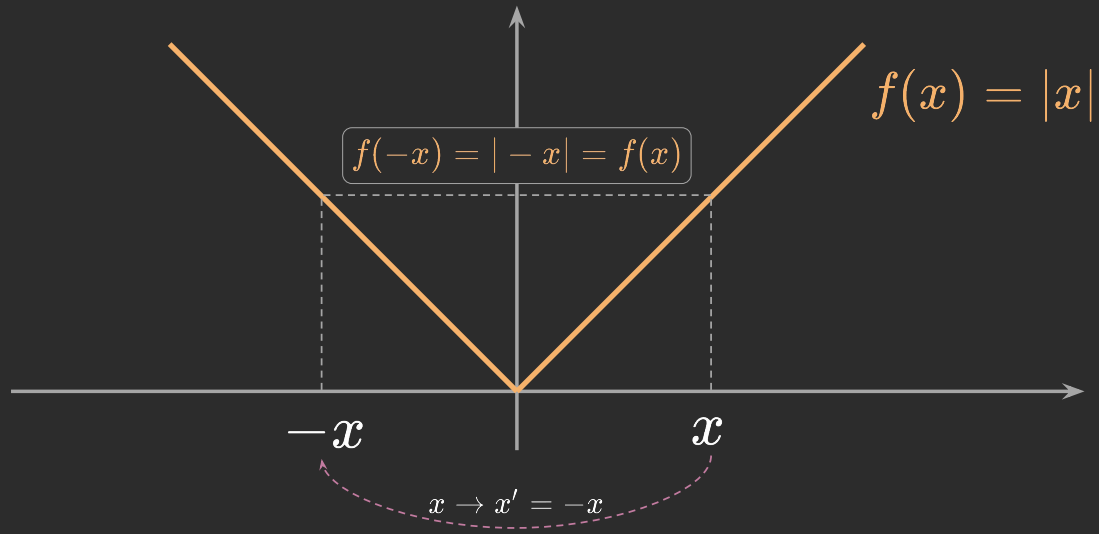
The ultimate question of this lecture, machine learning and everything:





The ultimate question of this lecture, machine learning and everything:





The ultimate question of this lecture, machine learning and everything:

"That which is apprehended by intelligence and reason is always in the **same state**"

Vergleichende Betrachtungen

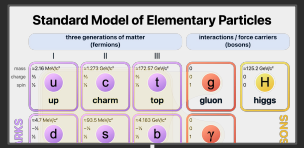
über

neuere geometrische Forschungen

von

Dr. Felix Klein,

a. o. Professor der Mathematik an der Universität Erlangen.



~400 BC
philosophy

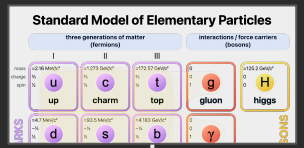
1872
geometry

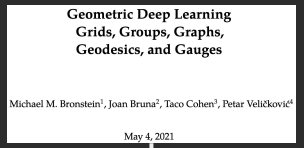
'900
physics

The ultimate question of this lecture, machine learning and everything:

"That which is apprehended by intelligence and reason is always in the **same state**"


Vergleichende Betrachtungen
über
neuere geometrische Forschungen
von
Dr. Felix Klein,
a. o. Professor der Mathematik an der Universität Erlangen




Geometric Deep Learning
Grids, Groups, Graphs,
Geodesics, and Gauges
Michael M. Bronstein[1], Joan Bruna[2], Taco Cohen[3], Petar Veličković[4]
May 4, 2021

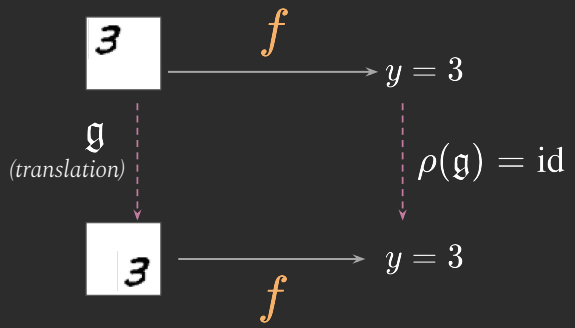
~400 BC
philosophy

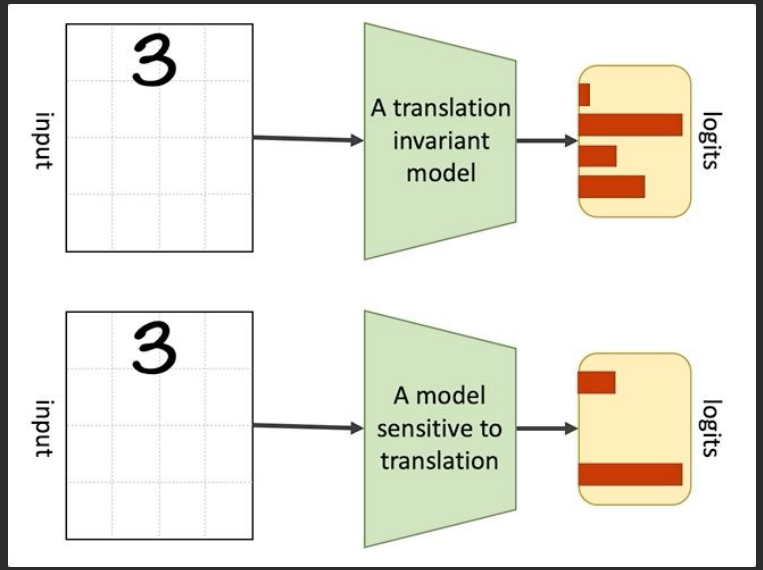
1872
geometry

'900
physics

2021
deep learning

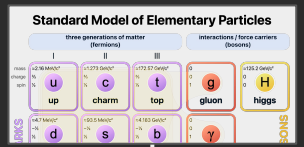
The ultimate question of this lecture, machine learning and everything:

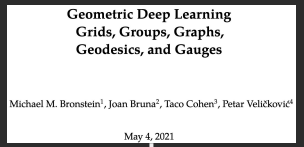




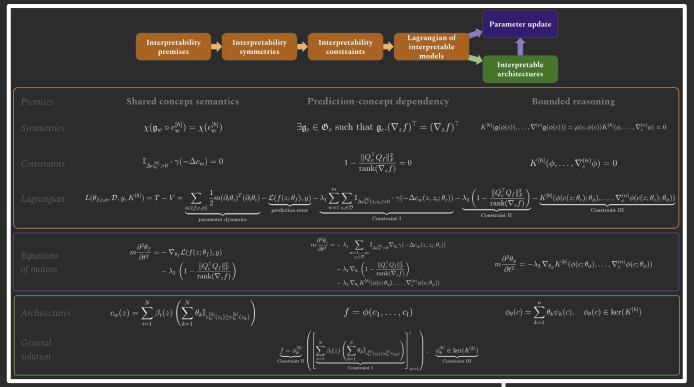
The ultimate question of this lecture, machine learning and everything:

"That which is apprehended by intelligence and reason is always in the **same state**"





Standard Interpretable Model



~400 BC *philosophy*

1872 *geometry*

'900 *physics*

2021 *deep learning*

2026 *interpretable ML*

Interpretability Theory Outline

I. Introduction to interpretability (~15 min)
II. The Standard Interpretable Model (~10 min)
III. Interpretability premises (~5 min)
IV. Interpretability symmetries (~10 min)
V. Interpretability constraints (~10 min)

Q&A + 5 min break

VI. Lagrangian of interpretable machine learning models (~5 min)
V. Trajectories towards interpretability (~10 min)
VI. Interpretable architectures (~10 min)

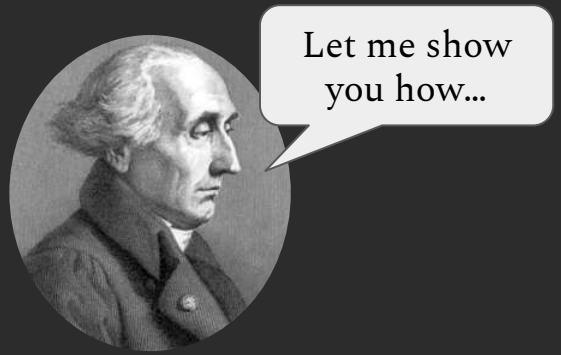
Final Q&A

The Standard Interpretable Model

What counts as interpretable for the user X?

The Standard Interpretable Model

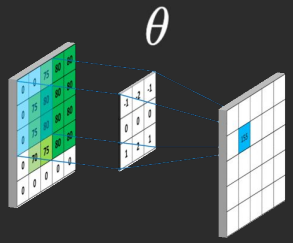
Interpretable & accurate model

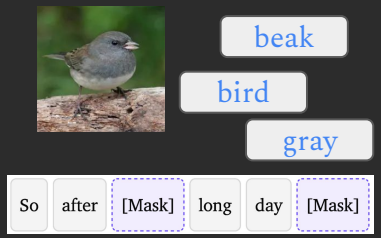


Giuseppe Luigi Lagrangia

[1] Lagrange. Mécanique analytique. (1788).

A Lagrangian perspective on Machine Learning





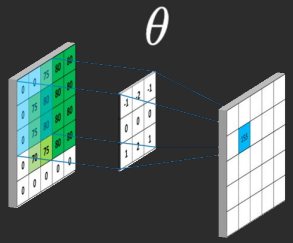
$$\|x^\top \theta - y\|^2$$

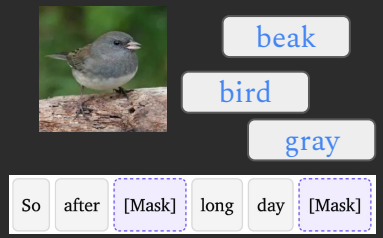
$$\frac{1}{2} m \partial_t \theta^\top \partial_t \theta$$

Parameters	(Training) Data	Objective	Dynamics
θ	$\mathcal{D}$	$V(\mathcal{D}, \theta)$	$T(\partial_t \theta)$

[1] Gori et al. Machine Learning: A constraint-based approach. Elsevier (2023).
[2] Guo et al. Physics of Learning: A Lagrangian perspective to different learning paradigms. arXiv preprint (2025).

A Lagrangian perspective on Machine Learning





$$\|x^\top \theta - y\|^2$$

$$\frac{1}{2} m \partial_t \theta^\top \partial_t \theta$$

Parameters	(Training) Data	Objective	Dynamics
θ	$\mathcal{D}$	$V(\mathcal{D}, \theta)$	$T(\partial_t \theta)$

$$L(\mathcal{D}, \theta) = \underbrace{T(\partial_t \theta)}_{\text{parameter dynamics}} - \underbrace{V(\mathcal{D}, \theta)}_{\text{objective function}}$$

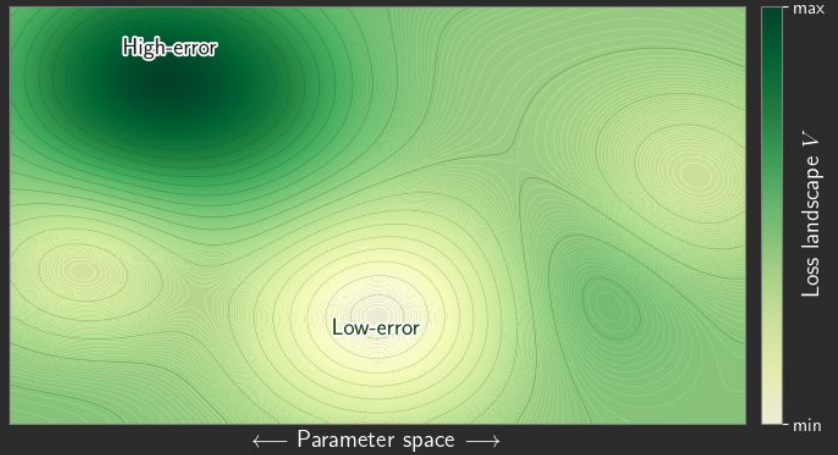
Lagrangian

[1] Gori et al. Machine Learning: A constraint-based approach. Elsevier (2023).
[2] Guo et al. Physics of Learning: A Lagrangian perspective to different learning paradigms. arXiv preprint (2025).

A Lagrangian perspective on Machine Learning

$$L(\mathcal{D}, \theta) = \underbrace{\|x^\top \theta - y\|^2}_{\text{objective function}}$$

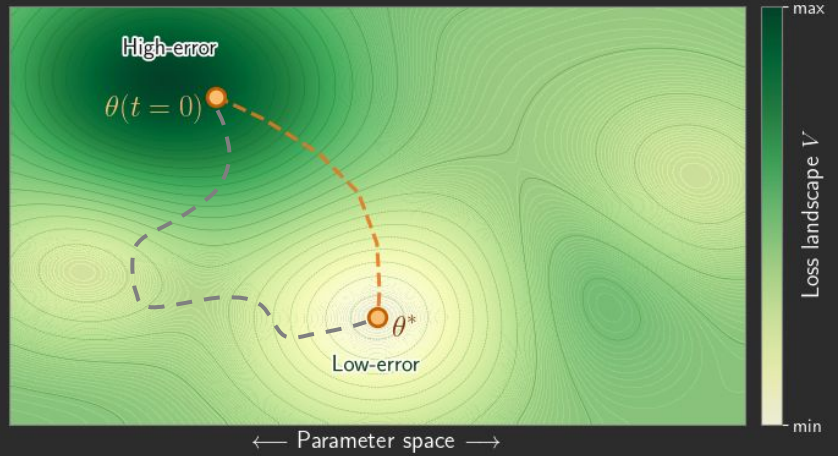
Lagrangian



A Lagrangian perspective on Machine Learning

$$L(\mathcal{D}, \theta) = \underbrace{\frac{1}{2} m \partial_t \theta^\top \partial_t \theta}_{\text{parameter dynamics}} - \underbrace{\|x^\top \theta - y\|^2}_{\text{objective function}}$$

Lagrangian



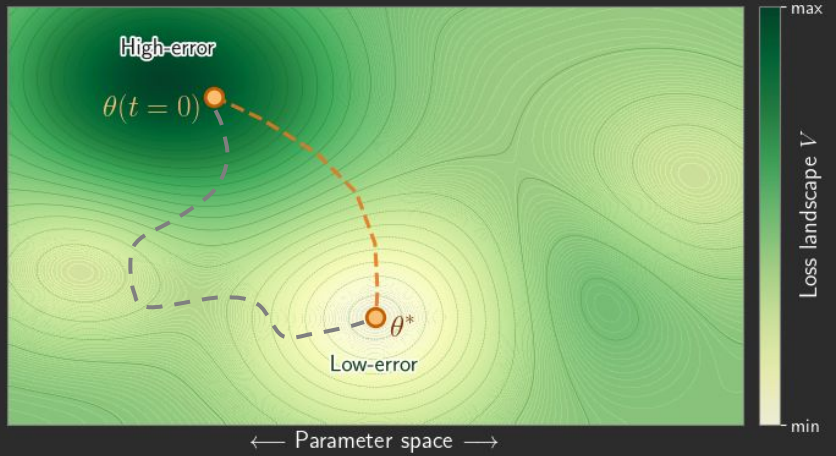
A Lagrangian perspective on Machine Learning

$$L(\mathcal{D}, \theta) = \underbrace{\frac{1}{2} m \partial_t \theta^\top \partial_t \theta}_{\text{parameter dynamics}} - \underbrace{\|x^\top \theta - y\|^2}_{\text{objective function}}$$

Lagrangian

$$S = \iint L(\mathcal{D}, \theta)\, \mathrm{d}x\, \mathrm{d}t$$

Action



A Lagrangian perspective on Machine Learning

$$L(\mathcal{D}, \theta) = \underbrace{\frac{1}{2} m \partial_t \theta^\top \partial_t \theta}_{\text{parameter dynamics}} - \underbrace{\|x^\top \theta - y\|^2}_{\text{objective function}}$$

Lagrangian

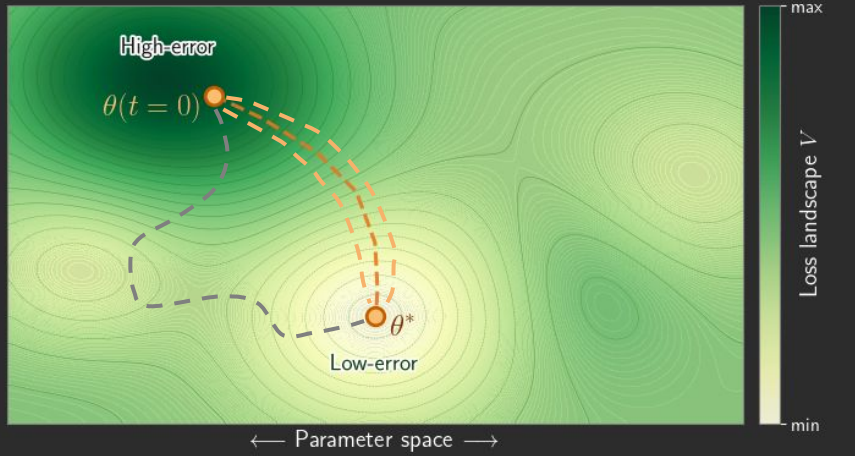
$$S = \iint L(\mathcal{D}, \theta)\, \mathrm{d}x\, \mathrm{d}t$$

Action

Principle of least action

$$\delta S = 0 \iff \frac{d}{dt}\left(\frac{\partial L}{\partial(\partial_t \theta)}\right) - \frac{\partial L}{\partial \theta} = 0$$

Euler-Lagrange equations



A Lagrangian perspective on Machine Learning

$$L(\mathcal{D}, \theta) = \underbrace{\frac{1}{2} m \partial_t \theta^\top \partial_t \theta}_{\text{parameter dynamics}} - \underbrace{\|x^\top \theta - y\|^2}_{\text{objective function}}$$

Lagrangian

$$S = \iint L(\mathcal{D}, \theta)\, \mathrm{d}x\, \mathrm{d}t$$

Action

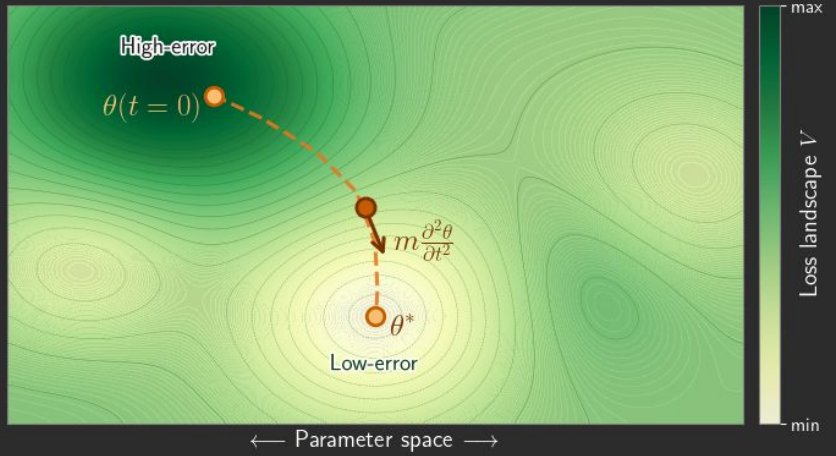
Principle of least action

$$\delta S = 0 \iff \frac{d}{dt}\left(\frac{\partial L}{\partial(\partial_t \theta)}\right) - \frac{\partial L}{\partial \theta} = 0$$

Euler-Lagrange equations

$$m \frac{\partial^2 \theta}{\partial t^2} = \nabla_\theta \|x^\top \theta - y\|^2$$

Equations of motion



A Lagrangian perspective on Machine Learning

$$L(\mathcal{D}, \theta) = \underbrace{\frac{1}{2} m \partial_t \theta^\top \partial_t \theta}_{\text{parameter dynamics}} - \underbrace{\|x^\top \theta - y\|^2}_{\text{objective function}}$$

Lagrangian

$$S = \iint L(\mathcal{D}, \theta)\, \mathrm{d}x\, \mathrm{d}t$$

Action

Principle of least action

$$\delta S = 0 \iff \frac{d}{dt}\left(\frac{\partial L}{\partial(\partial_t \theta)}\right) - \frac{\partial L}{\partial \theta} = 0$$

Euler-Lagrange equations

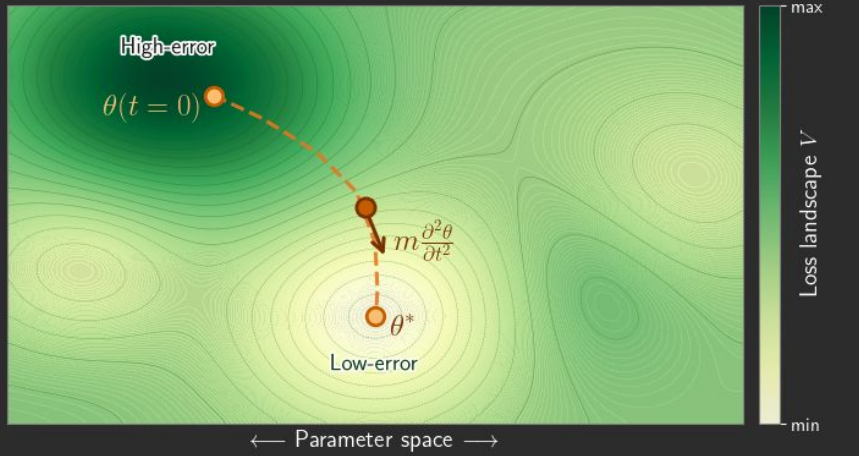
$$m \frac{\partial^2 \theta}{\partial t^2} = \nabla_\theta \|x^\top \theta - y\|^2$$

Equations of motion

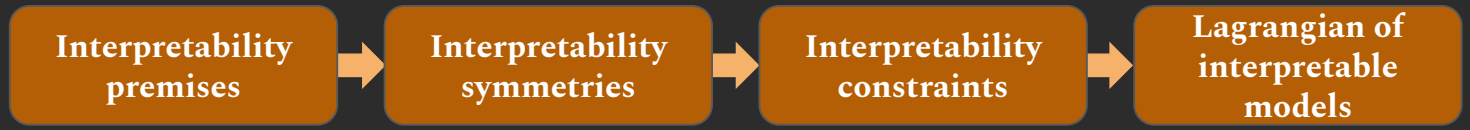
Discretization

$$\theta_{t+1} = \theta_t + (\theta_t - \theta_{t-1}) - \frac{(\Delta t)^2}{m} \nabla_\theta \|x^\top \theta - y\|^2$$

Parameter update



The Standard Interpretable Model



The Standard Interpretable Model

The Standard Interpretable Model

Interpretability Theory Outline

I. Introduction to interpretability (~15 min)

II. The Standard Interpretable Model (~10 min)

III. Interpretability premises (~5 min)

IV. Interpretability symmetries (~10 min)

V. Interpretability constraints (~10 min)

Q&A + 5 min break

VI. Lagrangian of interpretable machine learning models (~5 min)

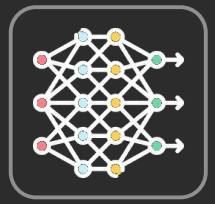
V. Trajectories towards interpretability (~10 min)

VI. Interpretable architectures (~10 min)

Final Q&A

Interpretability premises

When is f interpretable for the **target user**?



Machine
learning model

Bounded
formal user

Assumption: bounded and formal users

Bounded formal user

Alfred Tarski

Herbert Simon

Assumption: bounded and formal users

Bounded formal user

apple

red

gray

Fixed vocabulary of symbols (syntax)

$$c^{[h]}_{\texttt{red}}(\text{🍎}) = 0.9$$

$$c^{[h]}_{\texttt{red}}(\text{🍋}) = 0.2$$

Mechanism to assign meaning to symbols (semantics)

Alfred Tarski

Herbert Simon

Assumption: bounded and formal users

Bounded formal user

apple

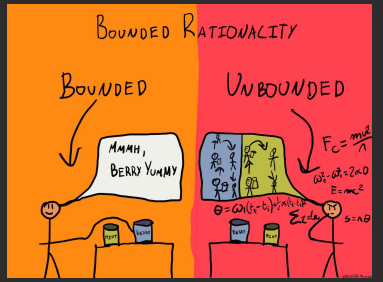
red

gray

Fixed vocabulary of symbols (syntax)

$$c^{[h]}_{\texttt{red}}(🍎) = 0.9$$
$$c^{[h]}_{\texttt{red}}(🍋) = 0.2$$

Mechanism to assign meaning to symbols (semantics)



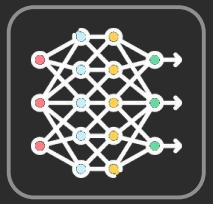
Time / computational limits (bounded rationality)

Alfred Tarski

Herbert Simon

Interpretability premises

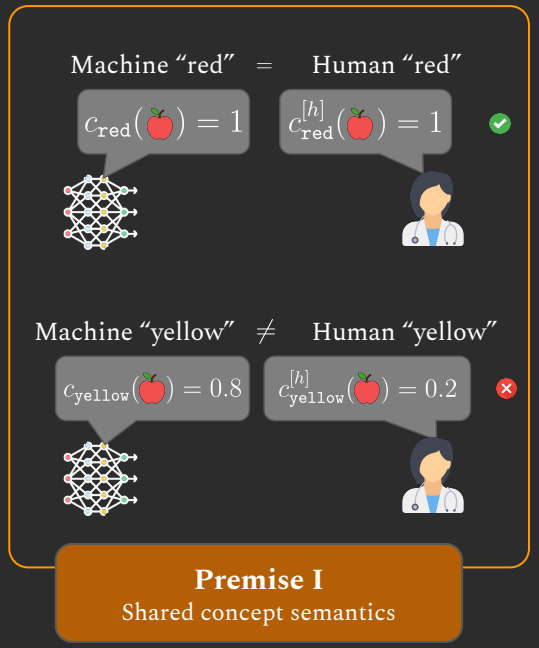
When is f interpretable for the **target user**?



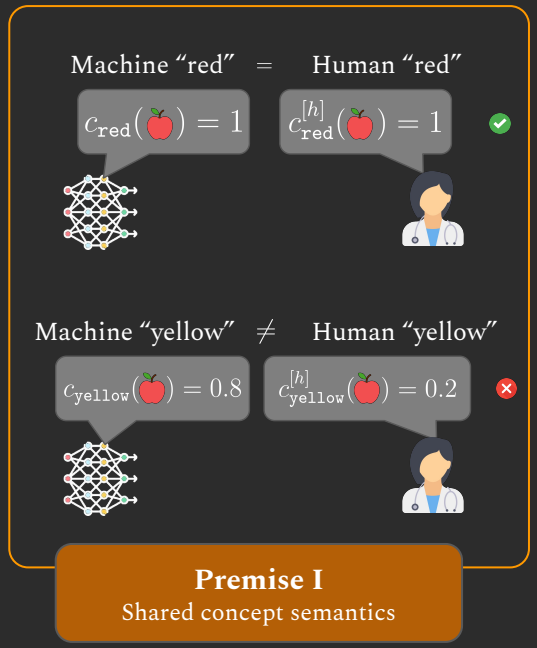
Machine
learning model

Bounded
formal user

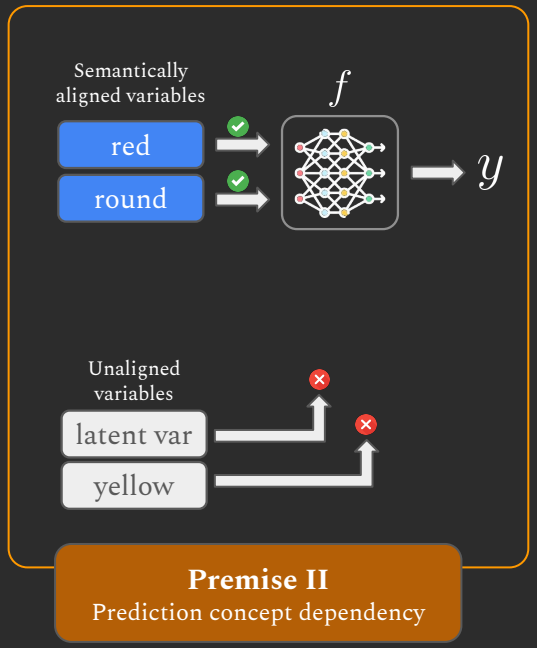
Interpretability premises



Interpretability premises

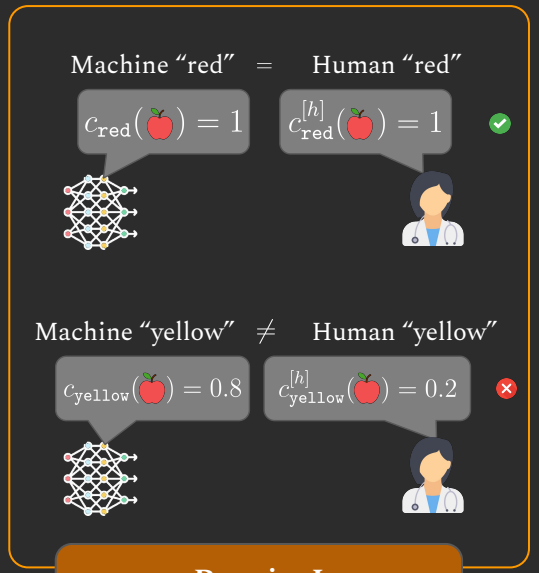


Premise I
Shared concept semantics

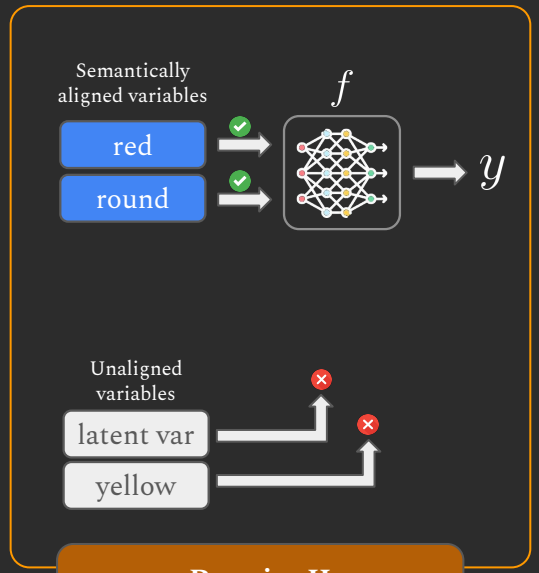


Premise II
Prediction concept dependency

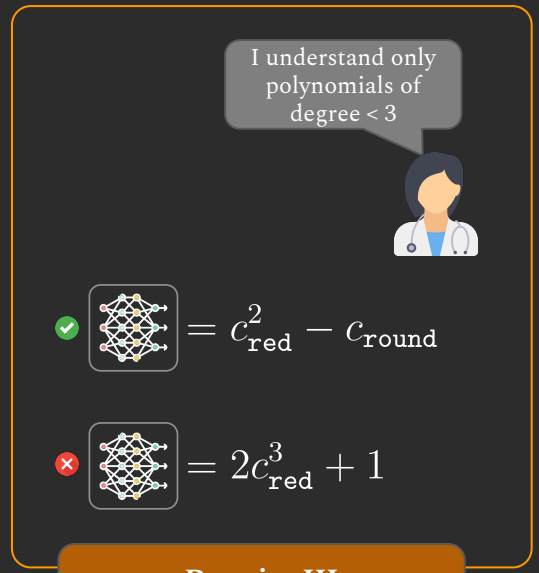
Interpretability premises



Premise I
Shared concept semantics

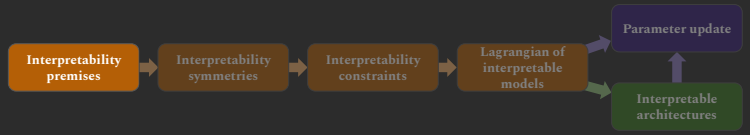


Premise II
Prediction concept dependency



Premise III
Bounded reasoning

The Standard Interpretable Model



Premises

Shared concept semantics

Prediction-concept dependency

Bounded reasoning

Symmetries

Constraints

Lagrangian

Equations of motion

Architectures

General solution

Interpretability Theory Outline

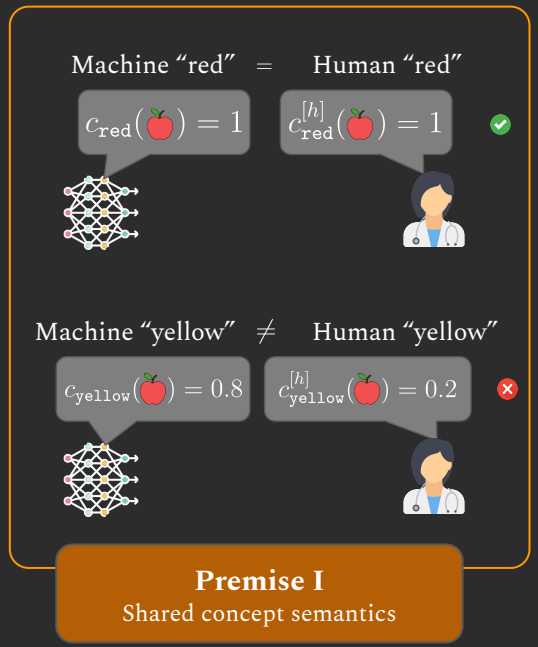
I. Introduction to interpretability (~15 min)
II. The Standard Interpretable Model (~10 min)
III. Interpretability premises (~5 min)
IV. Interpretability symmetries (~10 min)
V. Interpretability constraints (~10 min)

Q&A + 5 min break

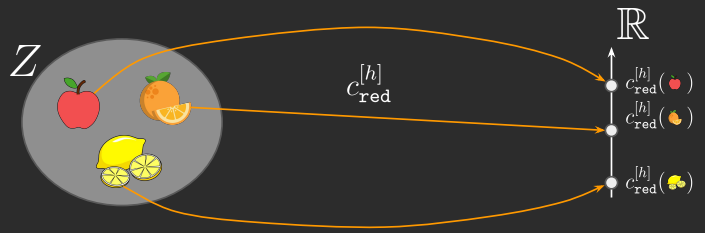
VI. Lagrangian of interpretable machine learning models (~5 min)
V. Trajectories towards interpretability (~10 min)
VI. Interpretable architectures (~10 min)

Final Q&A

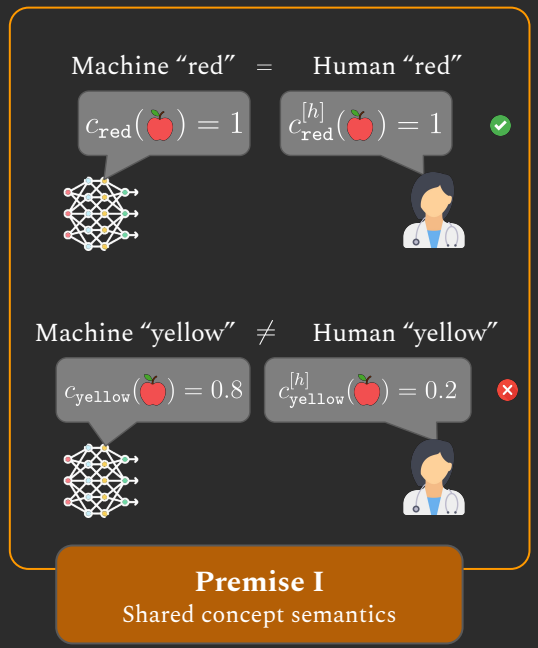
Premise I → Symmetry I



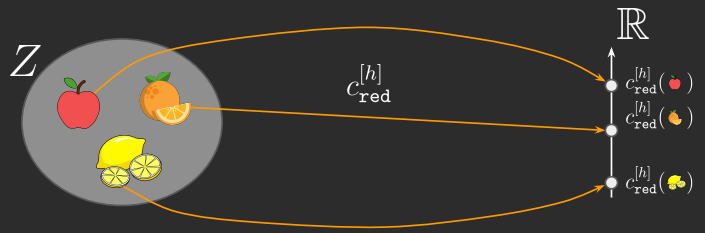
Concept (or interpretation) map:



Premise I → Symmetry I



Concept (or interpretation) map:

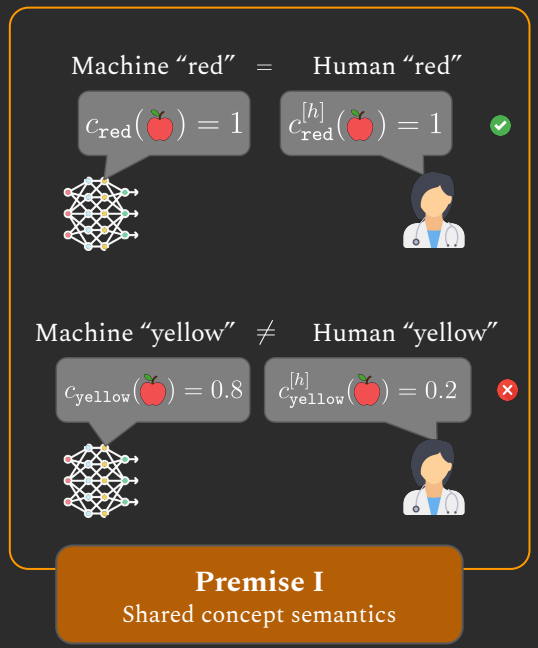


What is a general formal definition of the concept "red"?

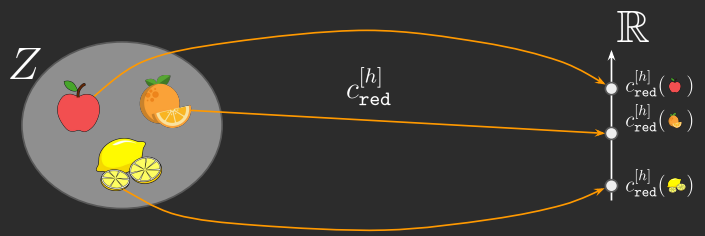
The set of all tuples (z_i, z_j) such that $c^{[h]}_{\texttt{red}}(z_i) < c^{[h]}_{\texttt{red}}(z_j)$

$$\chi(c^{[h]}_{\texttt{red}}) = \{(🍋, 🍊), (🍊, 🍎), (🍋, 🍎)\}$$

Premise I → Symmetry I



Concept (or interpretation) map:



What is a general formal definition of the concept “red”?

The set of all tuples (z_i, z_j) such that $c^{[h]}_{\texttt{red}}(z_i) < c^{[h]}_{\texttt{red}}(z_j)$

$$\chi(c^{[h]}_{\texttt{red}}) = \{(🍋, 🍊), (🍊, 🍎), (🍋, 🍎)\}$$

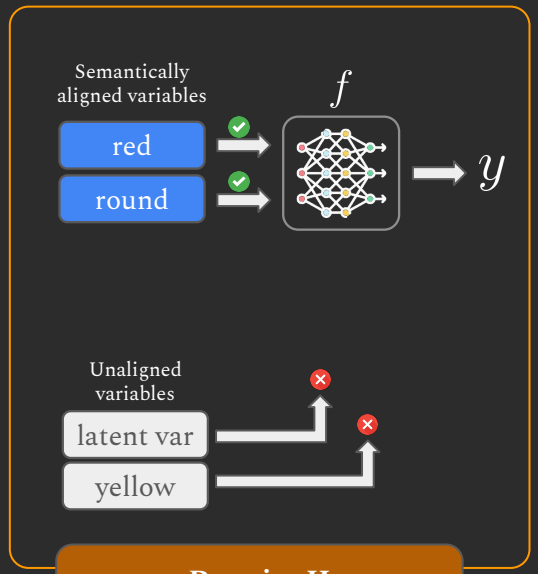
Which transformations preserve concepts?

Monotonic maps:

$$\mathfrak{G}_w = \{\mathfrak{g}_w : \mathbb{R} \to \mathbb{R} \mid \forall s_1, s_2 \in \mathbb{R}, \quad s_1 < s_2 \implies \mathfrak{g}_w(s_1) < \mathfrak{g}_w(s_2)\}$$

$$\chi(\mathfrak{g}_w \circ c^{[h]}_{\texttt{red}}) = \chi(c^{[h]}_{\texttt{red}}) = \{(🍋, 🍊), (🍊, 🍎), (🍋, 🍎)\}$$

Premise II → Symmetry II

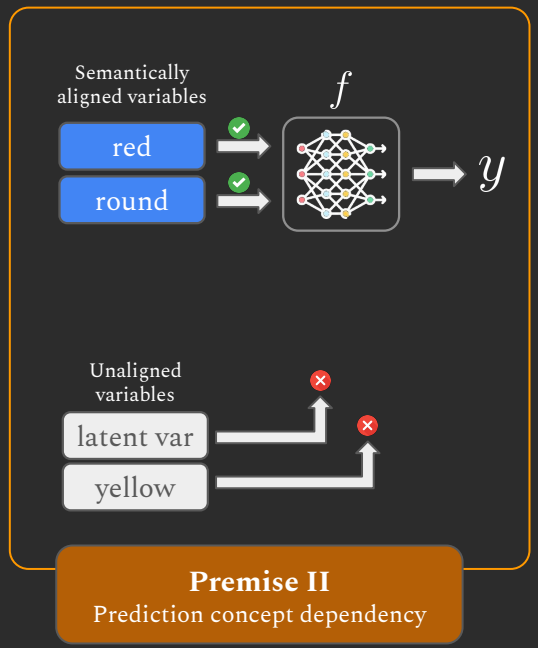


When does a function depend exclusively *on aligned concepts?*

When any change in the function output $\nabla_z f_1, \ldots, \nabla_z f_v$
can be explained by a change in aligned concepts $\nabla_z c_1, \ldots, \nabla_z c_l$

$$\text{span}\{\nabla_z f_1, \ldots, \nabla_z f_v\} \subseteq \text{span}\{\nabla_z c_1, \ldots, \nabla_z c_l\}$$

Premise II → Symmetry II



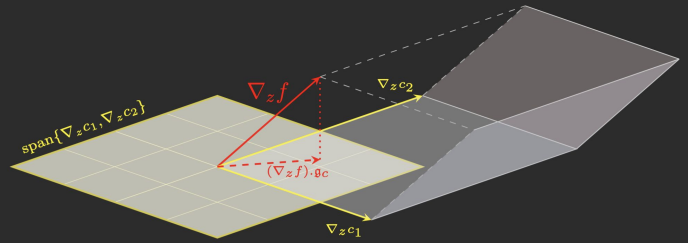
When does a function depend exclusively *on aligned concepts?*

When any change in the function output $\nabla_z f_1, \ldots, \nabla_z f_v$
can be explained by a change in aligned concepts $\nabla_z c_1, \ldots, \nabla_z c_l$

$$\operatorname{span}\{\nabla_z f_1, \ldots, \nabla_z f_v\} \subseteq \operatorname{span}\{\nabla_z c_1, \ldots, \nabla_z c_l\}$$

Which transformation guarantees prediction-concept dependency?

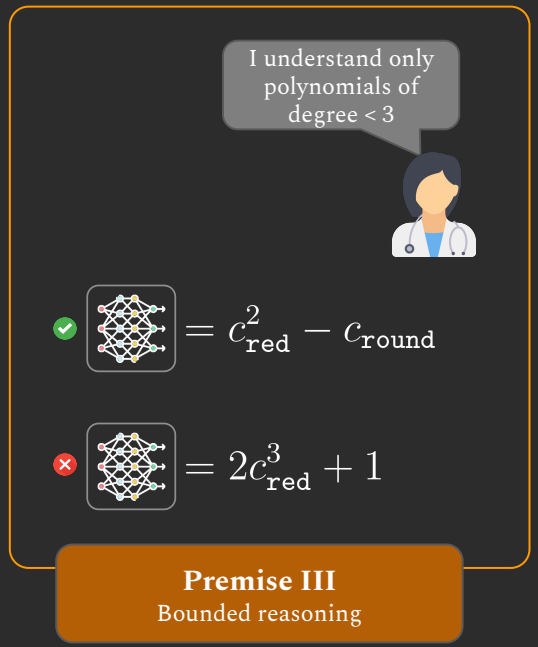
A projection into the concept subspace that leaves the output gradients invariant



$$\exists \mathfrak{g}_c \in \mathfrak{G}_c \text{ such that } \mathfrak{g}_c.(\nabla_z f)^\top = (\nabla_z f)^\top$$

$$\mathfrak{G}_c = \{\mathfrak{g}_c : Z \to Z \mid \mathfrak{g}_c^2 = \mathfrak{g}_c, \quad \operatorname{im}(\mathfrak{g}_c) \subseteq \operatorname{span}(\nabla_z c)\}$$

Premise III → Symmetry III



How can a user specify the hypothesis space?

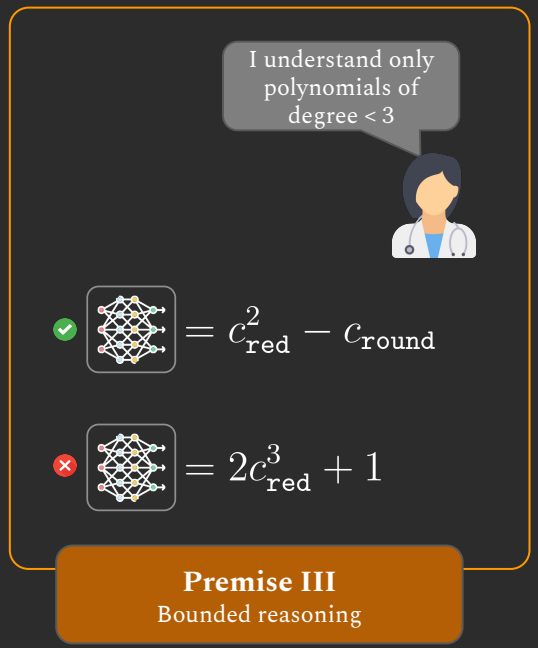
$$\{\phi \mid K^{[h]}(\phi) = \nabla_c^3 \phi = 0\}$$

$\mathcal{H}$

$2c_1^2 + 3$ $\quad c_1^2 + 3c_2$

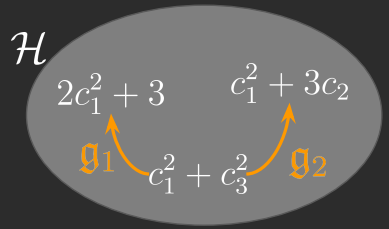
$c_1^2 + c_3^2$

Premise III → Symmetry III



How can a user specify the hypothesis space?

$$\{\phi \mid K^{[h]}(\phi) = \nabla_c^3 \phi = 0\}$$

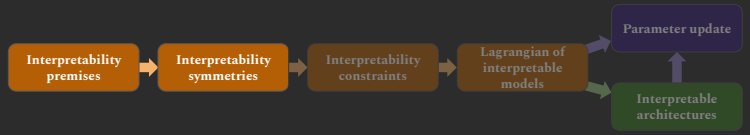


Which transformations of concept formulae preserve the hypothesis space?

All transformations that map solutions of $K^{[h]}$ to other solutions

$$K^{[h]}(\mathfrak{g}(\phi(c)), \ldots, \nabla^{(\nu)}\mathfrak{g}(\phi(c))) = \mu(c, \phi(c)) K^{[h]}(\phi, \ldots, \nabla_c^{(\nu)}\phi) = 0$$

The Standard Interpretable Model



Premises	**Shared concept semantics**	**Prediction-concept dependency**	**Bounded reasoning**
Symmetries	$\chi(\mathfrak{g}_w \circ c_w^{[h]}) = \chi(c_w^{[h]})$	$\exists \mathfrak{g}_c \in \mathfrak{G}_c$ such that $\mathfrak{g}_c.(\nabla_z f)^\top = (\nabla_z f)^\top$	$K^{[h]}(\mathfrak{g}(\phi(c)), \ldots, \nabla^{(\nu)}\mathfrak{g}(\phi(c))) = \mu(c, \phi(c))K^{[h]}(\phi, \ldots, \nabla_c^{(\nu)}\phi) = 0$
Constraints			
Lagrangian			

Equations of motion

Architectures

General solution

Interpretability Theory Outline

I. Introduction to interpretability (~15 min)

II. The Standard Interpretable Model (~10 min)

III. Interpretability premises (~5 min)

IV. Interpretability symmetries (~10 min)

V. Interpretability constraints (~10 min)

Q&A + 5 min break

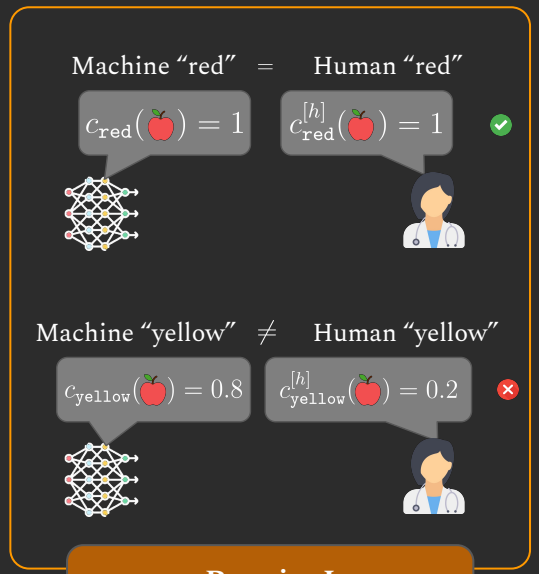
VI. Lagrangian of interpretable machine learning models (~5 min)

V. Trajectories towards interpretability (~10 min)

VI. Interpretable architectures (~10 min)

Final Q&A

Premise I → Symmetry I → Constraint I

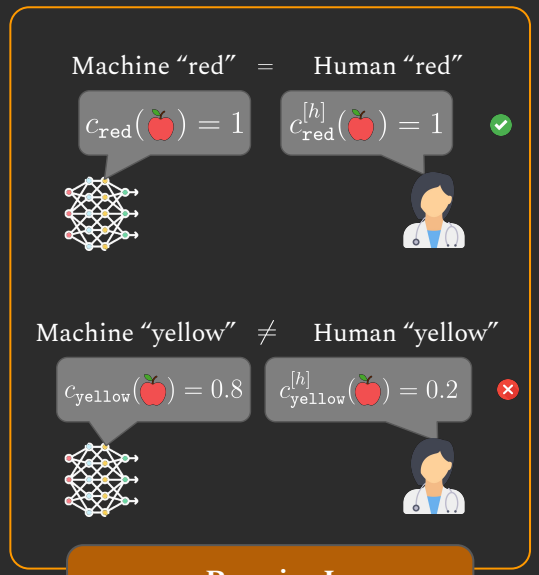


Premise I
Shared concept semantics

$$\chi(\mathfrak{g}_w \circ c_w^{[h]}) = \chi(c_w^{[h]})$$

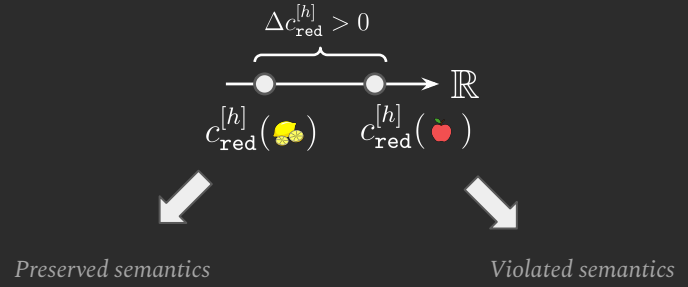
Symmetry I

Premise I → Symmetry I → Constraint I

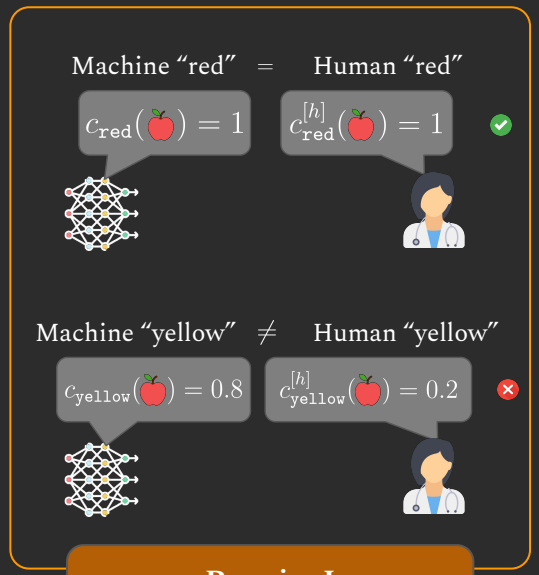


$$\chi(\mathfrak{g}_w \circ c_w^{[h]}) = \chi(c_w^{[h]})$$

Symmetry I



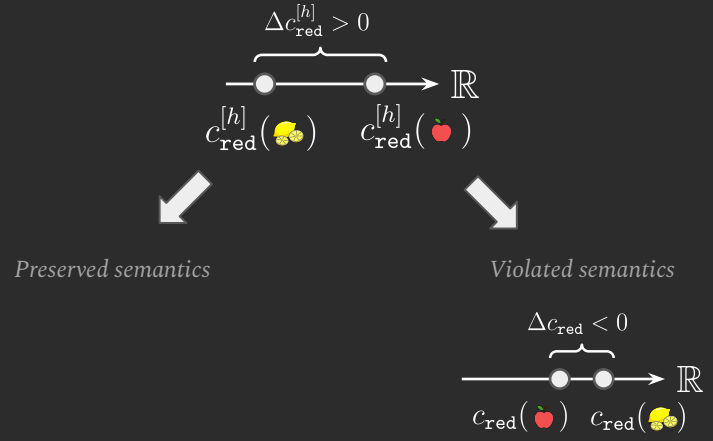
Premise I → Symmetry I → Constraint I



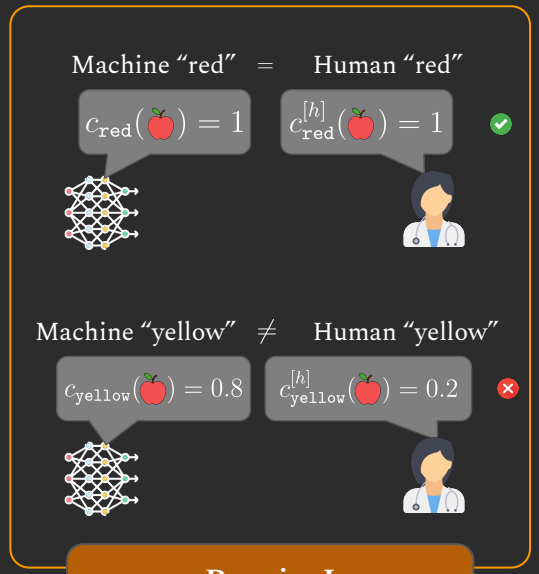
Premise I
Shared concept semantics

$$\chi(\mathfrak{g}_w \circ c_w^{[h]}) = \chi(c_w^{[h]})$$

Symmetry I

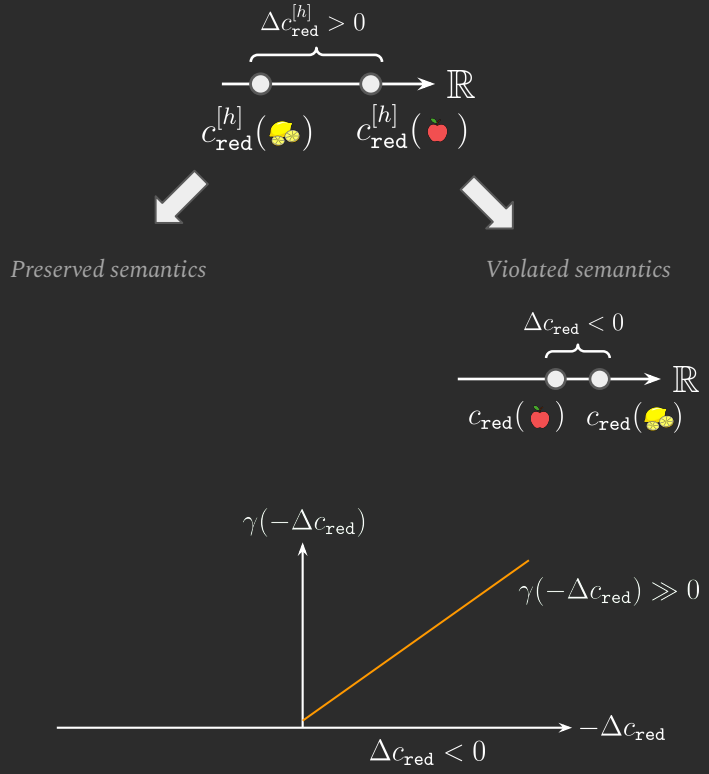


Premise I → Symmetry I → Constraint I

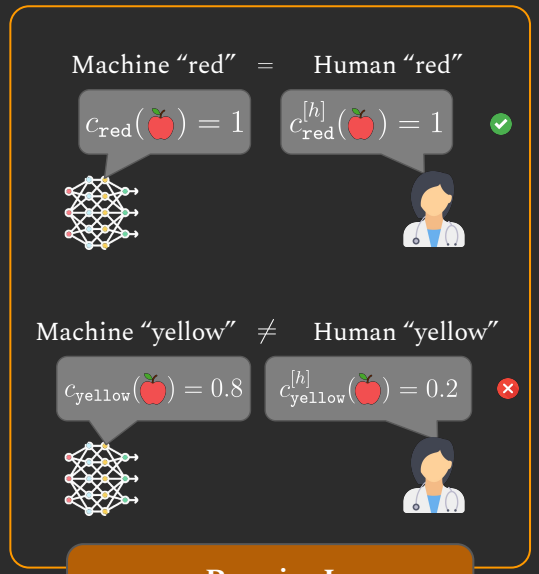


$$\chi(\mathfrak{g}_w \circ c_w^{[h]}) = \chi(c_w^{[h]})$$

Symmetry I

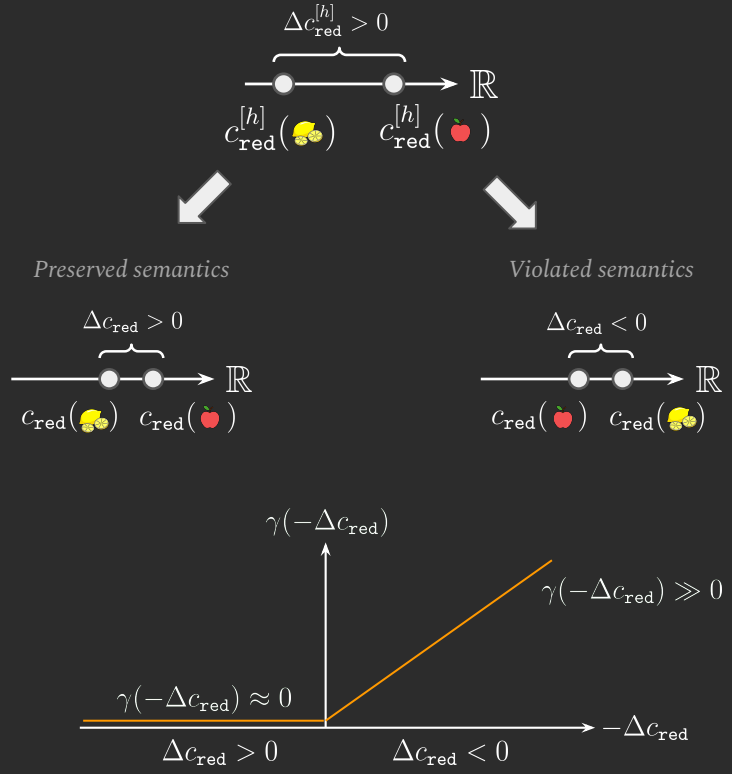


Premise I → Symmetry I → Constraint I

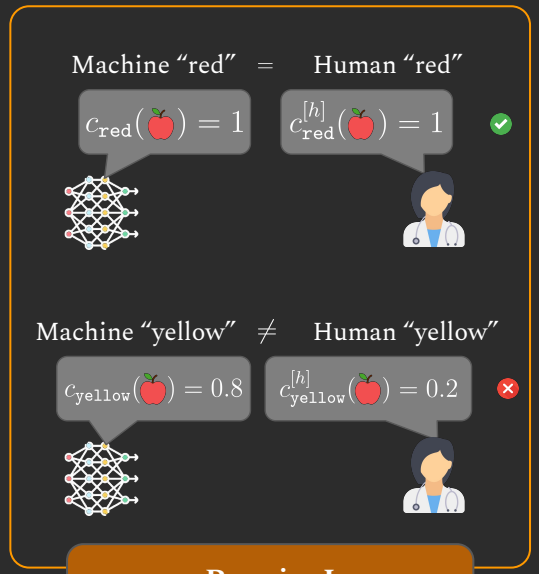


$$\chi(\mathfrak{g}_w \circ c_w^{[h]}) = \chi(c_w^{[h]})$$

Symmetry I



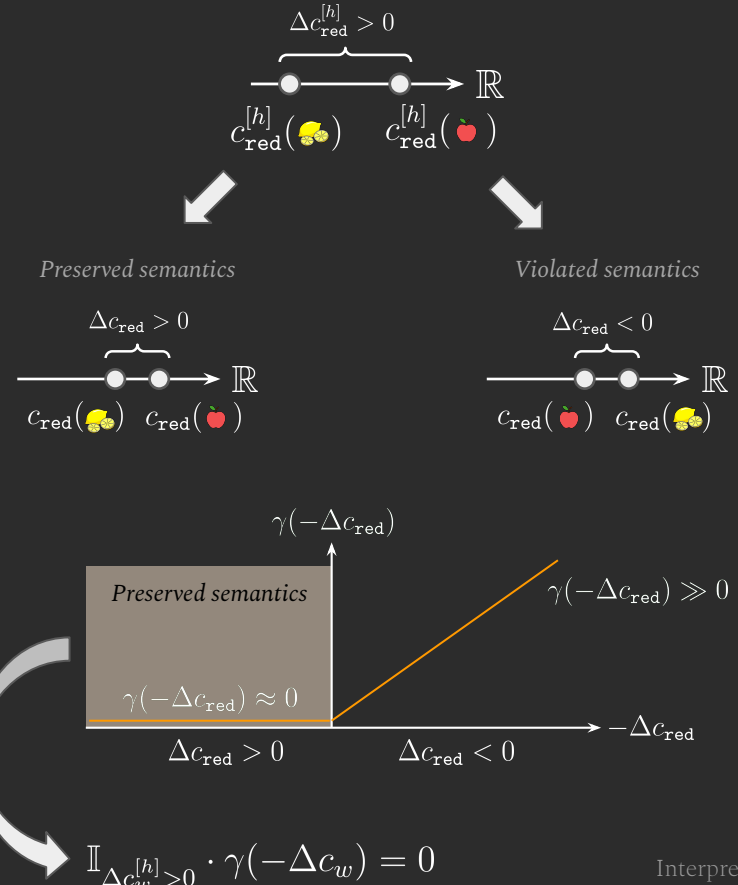
Premise I → Symmetry I → Constraint I



Premise I
Shared concept semantics

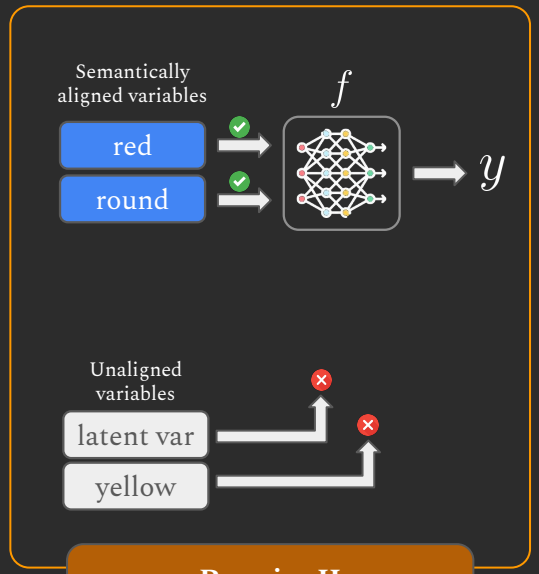
$$\chi(\mathfrak{g}_w \circ c_w^{[h]}) = \chi(c_w^{[h]})$$

Symmetry I



$$\mathbb{I}_{\Delta c_w^{[h]} > 0} \cdot \gamma(-\Delta c_w) = 0$$

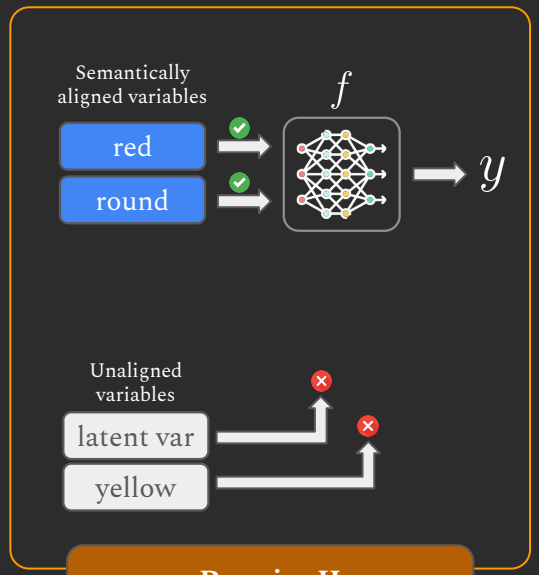
Premise II → Symmetry II → Constraint II



$$\mathfrak{g}_c \cdot (\nabla_z f)^\top = (\nabla_z f)^\top$$

Symmetry II

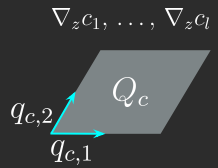
Premise II → Symmetry II → Constraint II

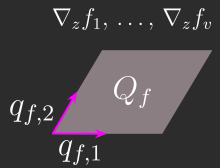


Premise II
Prediction-concept dependency

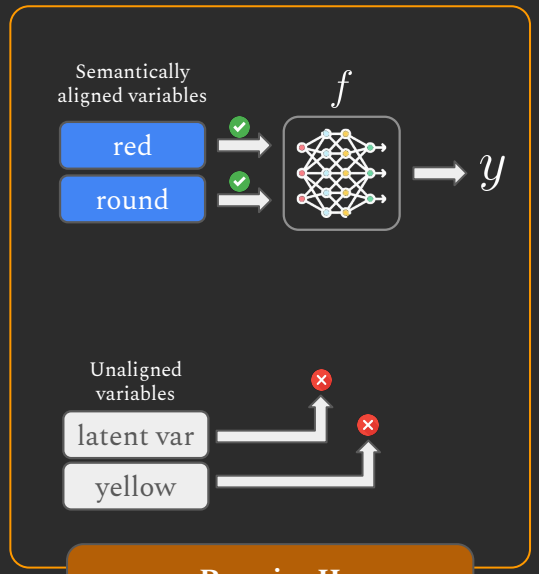
$$\mathfrak{g}_c \cdot (\nabla_z f)^\top = (\nabla_z f)^\top$$

Symmetry II





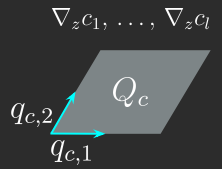
Premise II → Symmetry II → Constraint II

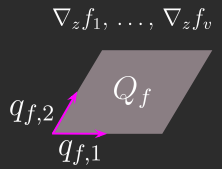


Premise II
Prediction-concept dependency

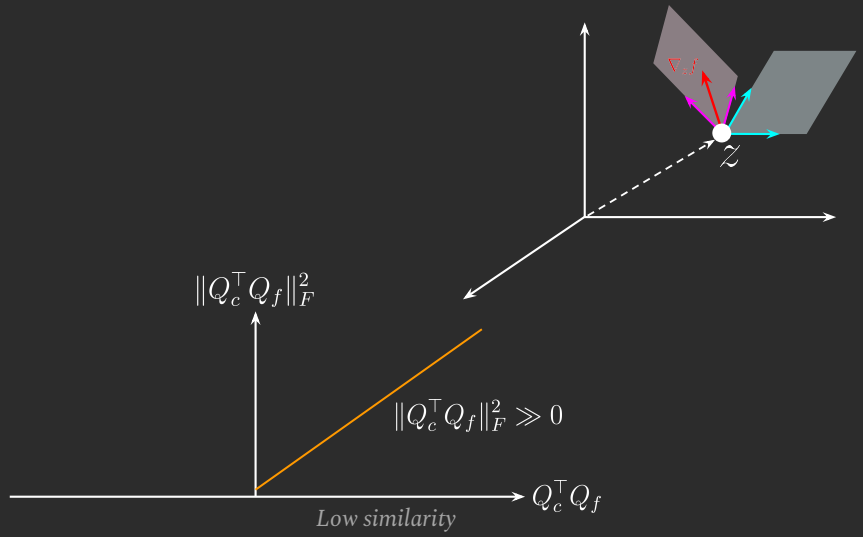
$$\mathfrak{g}_c \cdot (\nabla_z f)^\top = (\nabla_z f)^\top$$

Symmetry II

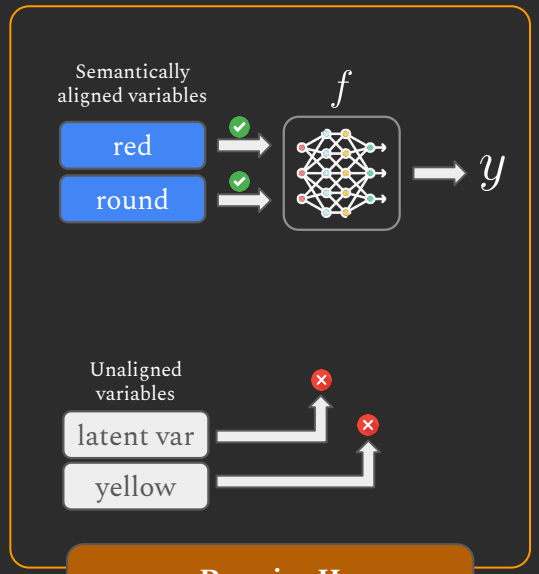




Predictions do not depend on concepts



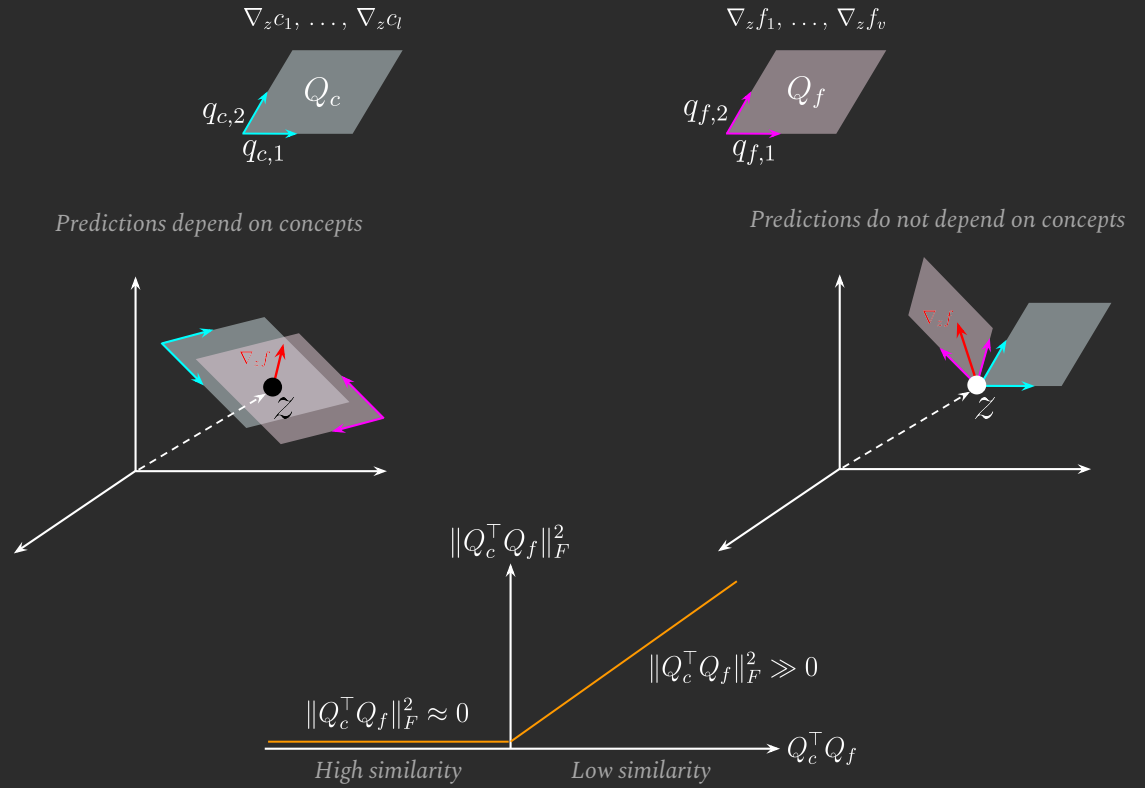
Premise II → Symmetry II → Constraint II



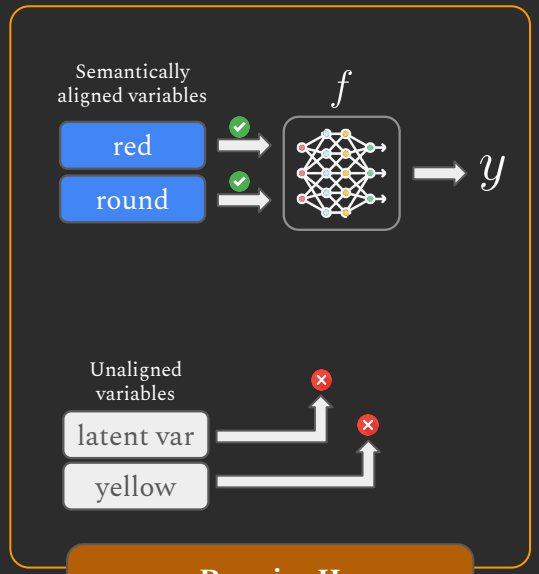
Premise II
Prediction-concept dependency

$$\mathfrak{g}_c \cdot (\nabla_z f)^\top = (\nabla_z f)^\top$$

Symmetry II



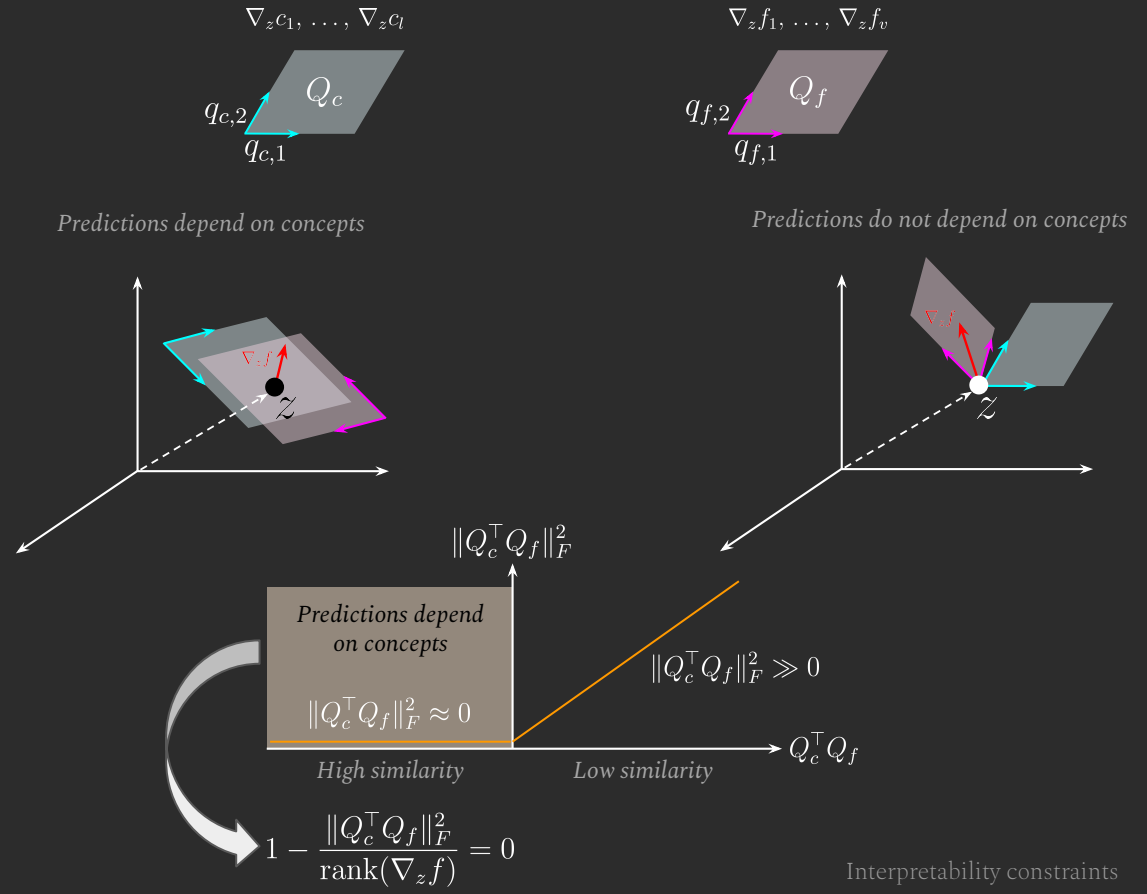
Premise II → Symmetry II → Constraint II



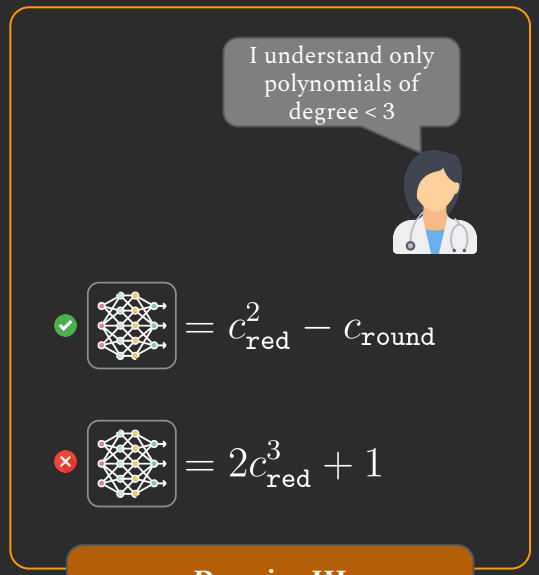
Premise II
Prediction-concept dependency

$$\mathfrak{g}_c \cdot (\nabla_z f)^\top = (\nabla_z f)^\top$$

Symmetry II



Premise III → Symmetry III → Constraint III



$$K^{[h]}(\nabla^{(\nu)}\mathfrak{g}(\phi(c))) = \mu(c,\phi(c))K^{[h]}(\nabla_c^{(\nu)}\phi) = 0$$

$$K^{[h]}(\phi, \ldots, \nabla_c^{(n)}\phi) = 0$$

The Standard Interpretable Model

	Shared concept semantics	**Prediction-concept dependency**	**Bounded reasoning**
Premises			
Symmetries	$\chi(\mathfrak{g}_w \circ c_w^{[h]}) = \chi(c_w^{[h]})$	$\exists \mathfrak{g}_c \in \mathfrak{G}_c$ such that $\mathfrak{g}_c.(\nabla_z f)^\top = (\nabla_z f)^\top$	$K^{[h]}(\mathfrak{g}(\phi(c)), \ldots, \nabla^{(\nu)}\mathfrak{g}(\phi(c))) = \mu(c, \phi(c)) K^{[h]}(\phi, \ldots, \nabla_c^{(\nu)}\phi) = 0$
Constraints	$\mathbb{I}_{\Delta c_w^{[h]} > 0} \cdot \gamma(-\Delta c_w) = 0$	$1 - \frac{\lVert Q_c^\top Q_f \rVert_F^2}{\operatorname{rank}(\nabla_z f)} = 0$	$K^{[h]}(\phi, \ldots, \nabla_c^{(n)}\phi) = 0$
Lagrangian			

Equations of motion

Architectures

General solution

The Standard Interpretable Model

	Shared concept semantics	Prediction-concept dependency	Bounded reasoning
Premises			
Symmetries	$\chi(\mathfrak{g}_w \circ c_w^{[h]}) = \chi(c_w^{[h]})$	$\exists \mathfrak{g}_c \in \mathfrak{G}_c$ such that $\mathfrak{g}_c.(\nabla_z f)^\top = (\nabla_z f)^\top$	$K^{[h]}(\mathfrak{g}(\phi(c)), \ldots, \nabla^{(\nu)}\mathfrak{g}(\phi(c))) = \mu(c, \phi(c))K^{[h]}(\phi, \ldots, \nabla_c^{(\nu)}\phi) = 0$
Constraints	$\mathbb{I}_{\Delta c_w^{[h]} > 0} \cdot \gamma(-\Delta c_w) = 0$	$1 - \frac{\|Q_c^\top Q_f\|_F^2}{\text{rank}(\nabla_z f)} = 0$	$K^{[h]}(\phi, \ldots, \nabla_c^{(n)}\phi) = 0$
Lagrangian			
Equations of motion			
Architectures			
General solution			

5′ Break

Interpretability Theory Outline

I. Introduction to interpretability (~15 min)
II. The Standard Interpretable Model (~10 min)
III. Interpretability premises (~5 min)
IV. Interpretability symmetries (~10 min)
V. Interpretability constraints (~10 min)

Q&A + 5 min break

VI. Lagrangian of interpretable machine learning models (~5 min)
V. Trajectories towards interpretability (~10 min)
VI. Interpretable architectures (~10 min)

Final Q&A

Lagrangian of interpretable machine learning models

$$\mathbb{I}_{\Delta c_w^{[h]}>0} \cdot \gamma(-\Delta c_w) = 0$$

$$1 - \frac{\|Q_c^\top Q_f\|_F^2}{\operatorname{rank}(\nabla_z f)} = 0$$

$$K^{[h]}(\phi, \ldots, \nabla_c^{(n)}\phi) = 0$$

Constraints

Lagrangian of interpretable machine learning models

$$\mathbb{I}_{\Delta c_w^{[h]}>0} \cdot \gamma(-\Delta c_w) = 0$$

$$1 - \frac{\|Q_c^\top Q_f\|_F^2}{\text{rank}(\nabla_z f)} = 0$$

$$K^{[h]}(\phi, \ldots, \nabla_c^{(n)}\phi) = 0$$

Constraints

$$\min_{\theta_f, \theta_c, \theta_\phi} \quad \mathcal{L}(f(z;\theta_f), y)$$

$$\text{s.t.} \quad \sum_{w=1}^{m} \sum_{z_i \in \mathcal{D}} \mathbb{I}_{\Delta c_w^{[h]}(z,z_i)>0} \cdot \gamma(-\Delta c_w(z, z_i; \theta_c)) = 0 \qquad \text{(Constraint 1)}$$

$$\left(1 - \frac{\|Q_c^\top Q_f\|_F^2}{\text{rank}(\nabla_z f)}\right) = 0 \qquad \text{(Constraint 2)}$$

$$K^{[h]}(\phi(c(z;\theta_c);\theta_\phi), \ldots, \nabla_c^{(n)}\phi(c(z;\theta_c);\theta_\phi)) = 0 \qquad \text{(Constraint 3)}$$

Constrained optimisation problem

Lagrangian of interpretable machine learning models

$$\mathbb{I}_{\Delta c_w^{[h]}>0} \cdot \gamma(-\Delta c_w) = 0$$

$$1 - \frac{\|Q_c^\top Q_f\|_F^2}{\operatorname{rank}(\nabla_z f)} = 0$$

$$K^{[h]}(\phi, \ldots, \nabla_c^{(n)}\phi) = 0$$

Constraints

$$\min_{\theta_f, \theta_c, \theta_\phi} \quad \mathcal{L}(f(z;\theta_f), y)$$

$$\text{s.t.} \quad \sum_{w=1}^{m} \sum_{z_i \in \mathcal{D}} \mathbb{I}_{\Delta c_w^{[h]}(z,z_i)>0} \cdot \gamma(-\Delta c_w(z, z_i; \theta_c)) = 0 \qquad \text{(Constraint 1)}$$

$$\left(1 - \frac{\|Q_c^\top Q_f\|_F^2}{\operatorname{rank}(\nabla_z f)}\right) = 0 \qquad \text{(Constraint 2)}$$

$$K^{[h]}(\phi(c(z;\theta_c);\theta_\phi), \ldots, \nabla_c^{(n)}\phi(c(z;\theta_c);\theta_\phi)) = 0 \qquad \text{(Constraint 3)}$$

Constrained optimisation problem

$$\begin{aligned}\max_{\lambda_i>0} \min_{\theta_f, \theta_\phi, \theta_c} L(\theta_{f,c,\phi}, \lambda_i, \mathcal{D}, y, K^{[h]}) = & \mathcal{L}(f(z;\theta_f), y) \\ & + \lambda_1 \sum_{w=1}^{m} \sum_{z_i \in \mathcal{D}} \mathbb{I}_{\Delta c_w^{[h]}(z,z_i)>0} \cdot \gamma(-\Delta c_w(z, z_i; \theta_c)) \\ & + \lambda_2 \left(1 - \frac{\|Q_c^\top Q_f\|_F^2}{\operatorname{rank}(\nabla_z f)}\right) \\ & + \lambda_3 K^{[h]}(\phi(c(z;\theta_c);\theta_\phi), \ldots, \nabla_c^{(n)}\phi(c(z;\theta_c);\theta_\phi))\end{aligned}$$

Unconstrained optimisation problem

Lagrangian of interpretable machine learning models

$$\mathbb{I}_{\Delta c_w^{[h]}>0} \cdot \gamma(-\Delta c_w) = 0$$

$$1 - \frac{\|Q_c^\top Q_f\|_F^2}{\text{rank}(\nabla_z f)} = 0$$

$$K^{[h]}(\phi, \ldots, \nabla_c^{(n)}\phi) = 0$$

Constraints

$$\min_{\theta_f, \theta_c, \theta_\phi} \quad \mathcal{L}(f(z;\theta_f), y)$$

$$\text{s.t.} \quad \sum_{w=1}^{m} \sum_{z_i \in \mathcal{D}} \mathbb{I}_{\Delta c_w^{[h]}(z,z_i)>0} \cdot \gamma(-\Delta c_w(z, z_i; \theta_c)) = 0 \qquad \text{(Constraint 1)}$$

$$\left(1 - \frac{\|Q_c^\top Q_f\|_F^2}{\text{rank}(\nabla_z f)}\right) = 0 \qquad \text{(Constraint 2)}$$

$$K^{[h]}(\phi(c(z;\theta_c);\theta_\phi), \ldots, \nabla_c^{(n)}\phi(c(z;\theta_c);\theta_\phi)) = 0 \qquad \text{(Constraint 3)}$$

Constrained optimisation problem

$$\max_{\lambda_i>0} \min_{\theta_f, \theta_\phi, \theta_c} L(\theta_{f,c,\phi}, \lambda_i, \mathcal{D}, y, K^{[h]}) = \mathcal{L}(f(z;\theta_f), y)$$
$$+ \lambda_1 \sum_{w=1}^{m} \sum_{z_i \in \mathcal{D}} \mathbb{I}_{\Delta c_w^{[h]}(z,z_i)>0} \cdot \gamma(-\Delta c_w(z, z_i; \theta_c))$$
$$+ \lambda_2 \left(1 - \frac{\|Q_c^\top Q_f\|_F^2}{\text{rank}(\nabla_z f)}\right)$$
$$+ \lambda_3 K^{[h]}(\phi(c(z;\theta_c);\theta_\phi), \ldots, \nabla_c^{(n)}\phi(c(z;\theta_c);\theta_\phi))$$

Unconstrained optimisation problem

$$L(\theta_{f,c,\phi}, \mathcal{D}, y, K^{[h]}) = T - V = \underbrace{\sum_{i \in \{f,c,\phi\}} \frac{1}{2} m(\partial_t \theta_i)^T (\partial_t \theta_i)}_{\text{parameter dynamics}} - \underbrace{\mathcal{L}(f(z;\theta_f), y)}_{\text{prediction error}} - \underbrace{\lambda_1 \sum_{w=1}^{m} \sum_{z_i \in \mathcal{D}} \mathbb{I}_{\Delta c_w^{[h]}(z,z_i)>0} \cdot \gamma(-\Delta c_w(z, z_i; \theta_c))}_{\text{Constraint I}} - \underbrace{\lambda_2 \left(1 - \frac{\|Q_c^\top Q_f\|_F^2}{\text{rank}(\nabla_z f)}\right)}_{\text{Constraint II}} - \underbrace{K^{[h]}(\phi(c(z;\theta_c);\theta_\phi), \ldots, \nabla_c^{(n)}\phi(c(z;\theta_c);\theta_\phi))}_{\text{Constraint III}}$$

Lagrangian

The Standard Interpretable Model

	Shared concept semantics	Prediction-concept dependency	Bounded reasoning
Premises			
Symmetries	$\chi(\mathfrak{g}_w \circ c_w^{[h]}) = \chi(c_w^{[h]})$	$\exists \mathfrak{g}_c \in \mathfrak{G}_c$ such that $\mathfrak{g}_c.(\nabla_z f)^\top = (\nabla_z f)^\top$	$K^{[h]}(\mathfrak{g}(\phi(c)), \ldots, \nabla^{(\nu)}\mathfrak{g}(\phi(c))) = \mu(c, \phi(c)) K^{[h]}(\phi, \ldots, \nabla_c^{(\nu)}\phi) = 0$
Constraints	$\mathbb{I}_{\Delta c_w^{[h]} > 0} \cdot \gamma(-\Delta c_w) = 0$	$1 - \frac{\|Q_c^\top Q_f\|_F^2}{\operatorname{rank}(\nabla_z f)} = 0$	$K^{[h]}(\phi, \ldots, \nabla_c^{(n)}\phi) = 0$
Lagrangian	$L(\theta_{f,c,\phi}, \mathcal{D}, y, K^{[h]}) = T - V = \underbrace{\sum_{i \in \{f,c,\phi\}} \frac{1}{2} m (\partial_t \theta_i)^T (\partial_t \theta_i)}_{\text{parameter dynamics}} - \underbrace{\mathcal{L}(f(z;\theta_f), y)}_{\text{prediction error}} - \underbrace{\lambda_1 \sum_{w=1}^{m} \sum_{z_i \in \mathcal{D}} \mathbb{I}_{\Delta c_w^{[h]}(z,z_i) > 0} \cdot \gamma(-\Delta c_w(z, z_i; \theta_c))}_{\text{Constraint I}} - \underbrace{\lambda_2 \left(1 - \frac{\|Q_c^\top Q_f\|_F^2}{\operatorname{rank}(\nabla_z f)}\right)}_{\text{Constraint II}} - \underbrace{K^{[h]}(\phi(c(z;\theta_c);\theta_\phi), \ldots, \nabla_c^{(n)}\phi(c(z;\theta_c);\theta_\phi))}_{\text{Constraint III}}$		

Equations of motion

Architectures

General solution

Interpretability Theory Outline

I. Introduction to interpretability (~15 min)
II. The Standard Interpretable Model (~10 min)
III. Interpretability premises (~5 min)
IV. Interpretability symmetries (~10 min)
V. Interpretability constraints (~10 min)

Q&A + 5 min break

VI. Lagrangian of interpretable machine learning models (~5 min)
V. Trajectories towards interpretability (~10 min)
VI. Interpretable architectures (~10 min)

Final Q&A

Trajectories towards interpretability

$$L(\mathcal{D}, \theta) = \underbrace{T(\partial_t \theta)}_{\text{parameter dynamics}} - \underbrace{V(\mathcal{D}, \theta)}_{\text{objective function}}$$

Lagrangian

$$S = \iint L(\mathcal{D}, \theta)\, \mathrm{d}x\, \mathrm{d}t$$

Action

Principle of least action

$$\delta S = 0 \iff \frac{d}{dt}\left(\frac{\partial L}{\partial(\partial_t \theta)}\right) - \frac{\partial L}{\partial \theta} = 0$$

Euler-Lagrange equations

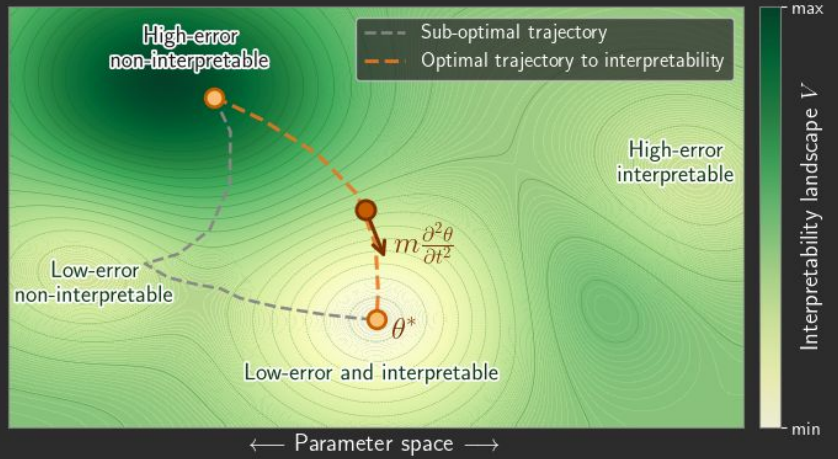
$$m\frac{\partial^2 \theta}{\partial t^2} = \nabla_\theta V(\mathcal{D}, \theta)$$

Equations of motion

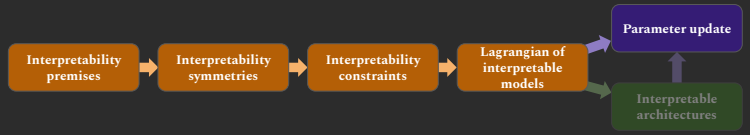
Discretization

$$\theta_{t+1} = \theta_t + (\theta_t - \theta_{t-1}) - \frac{(\Delta t)^2}{m}\nabla_\theta V(\mathcal{D}, \theta)$$

Parameter update



The Standard Interpretable Model



Premises	**Shared concept semantics**	**Prediction-concept dependency**	**Bounded reasoning**
Symmetries	$\chi(\mathfrak{g}_w \circ c_w^{[h]}) = \chi(c_w^{[h]})$	$\exists \mathfrak{g}_c \in \mathfrak{G}_c$ such that $\mathfrak{g}_c.(\nabla_z f)^\top = (\nabla_z f)^\top$	$K^{[h]}(\mathfrak{g}(\phi(c)), \ldots, \nabla^{(\nu)}\mathfrak{g}(\phi(c))) = \mu(c, \phi(c))K^{[h]}(\phi, \ldots, \nabla_c^{(\nu)}\phi) = 0$
Constraints	$\mathbb{I}_{\Delta c_w^{[h]}>0} \cdot \gamma(-\Delta c_w) = 0$	$1 - \frac{\|Q_c^\top Q_f\|_F^2}{\text{rank}(\nabla_z f)} = 0$	$K^{[h]}(\phi, \ldots, \nabla_c^{(n)}\phi) = 0$

Lagrangian

$$L(\theta_{f,c,\phi}, \mathcal{D}, y, K^{[h]}) = T - V = \underbrace{\sum_{i\in\{f,c,\phi\}} \frac{1}{2} m(\partial_t \theta_i)^T (\partial_t \theta_i)}_{\text{parameter dynamics}} - \underbrace{\mathcal{L}(f(z;\theta_f), y)}_{\text{prediction error}} - \underbrace{\lambda_1 \sum_{w=1}^{m} \sum_{z_i \in \mathcal{D}} \mathbb{I}_{\Delta c_w^{[h]}(z,z_i)>0} \cdot \gamma(-\Delta c_w(z, z_i; \theta_c))}_{\text{Constraint I}} - \underbrace{\lambda_2 \left(1 - \frac{\|Q_c^\top Q_f\|_F^2}{\text{rank}(\nabla_z f)}\right)}_{\text{Constraint II}} - \underbrace{K^{[h]}(\phi(c(z;\theta_c);\theta_\phi), \ldots, \nabla_c^{(n)}\phi(c(z;\theta_c);\theta_\phi))}_{\text{Constraint III}}$$

Equations of motion

$$m\frac{\partial^2 \theta_f}{\partial t^2} = -\nabla_{\theta_f}\mathcal{L}(f(z;\theta_f), y) - \lambda_2 \left(1 - \frac{\|Q_c^\top Q_f\|_F^2}{\text{rank}(\nabla_z f)}\right)$$

$$m\frac{\partial^2 \theta_c}{\partial t^2} = -\lambda_1 \sum_{\substack{w=1,\ldots,m \\ z_i \in \mathcal{D}}} \mathbb{I}_{\Delta c_w^{[h]}>0} \nabla_{\theta_c} \gamma(-\Delta c_w(z, z_i; \theta_c)) - \lambda_2 \nabla_{\theta_c}\left(1 - \frac{\|Q_c^\top Q_f\|_F^2}{\text{rank}(\nabla_z f)}\right) - \lambda_3 \nabla_{\theta_c} K^{[h]}(\phi(c;\theta_\phi), \ldots, \nabla_c^{(n)}\phi(c;\theta_\phi))$$

$$m\frac{\partial^2 \theta_\phi}{\partial t^2} = -\lambda_3 \nabla_{\theta_\phi} K^{[h]}(\phi(c;\theta_\phi), \ldots, \nabla_c^{(n)}\phi(c;\theta_\phi))$$

Architectures

General solution

Interpretability Theory Outline

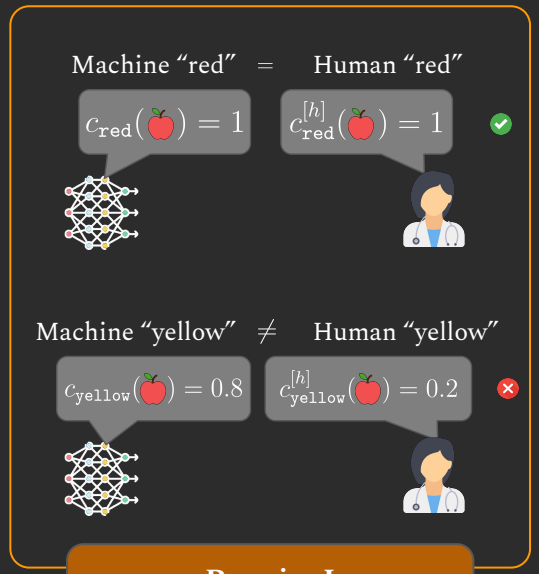
I. Introduction to interpretability (~15 min)
II. The Standard Interpretable Model (~10 min)
III. Interpretability premises (~5 min)
IV. Interpretability symmetries (~10 min)
V. Interpretability constraints (~10 min)

Q&A + 5 min break

VI. Lagrangian of interpretable machine learning models (~5 min)
V. Trajectories towards interpretability (~10 min)
VI. Interpretable architectures (~10 min)

Final Q&A

Premise I → Symmetry I → Constraint I → Architecture I

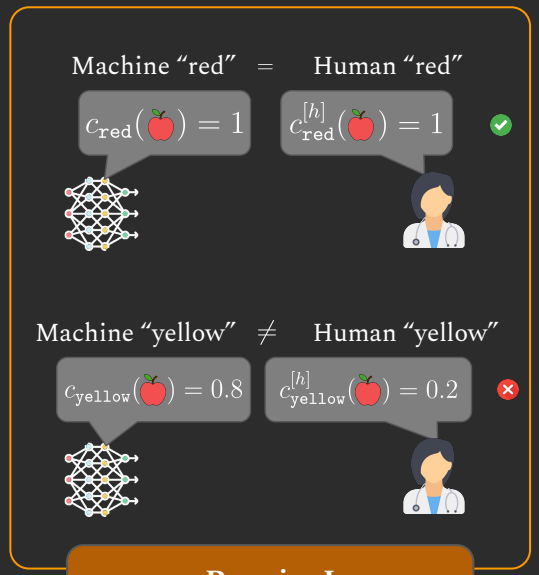


Premise I
Shared concept semantics

$$\mathbb{I}_{\Delta c_w^{[h]} > 0} \cdot \gamma(-\Delta c_w) = 0$$

Constraint I

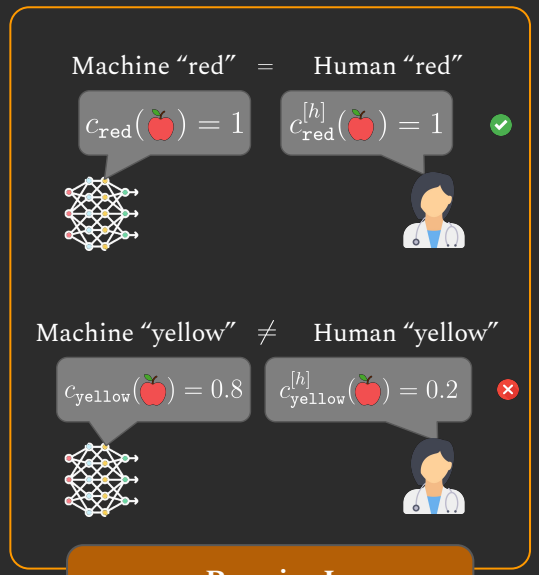
Premise I → Symmetry I → Constraint I → Architecture I



$$\mathbb{I}_{\Delta c_w^{[h]} > 0} \cdot \gamma(-\Delta c_w) = 0$$

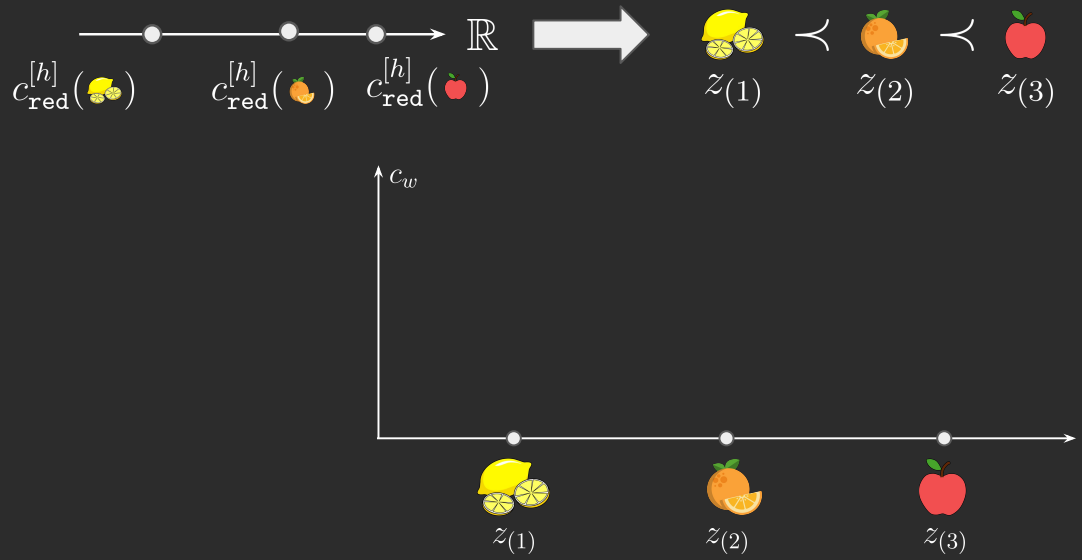
Constraint I

Premise I → Symmetry I → Constraint I → Architecture I

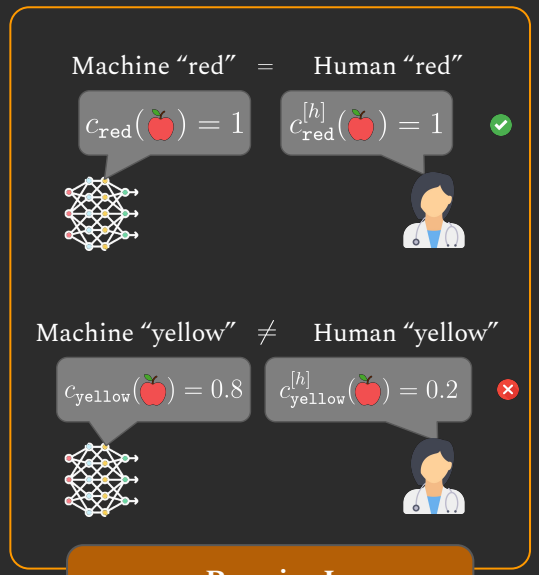


$$\mathbb{I}_{\Delta c_w^{[h]} > 0} \cdot \gamma(-\Delta c_w) = 0$$

Constraint I

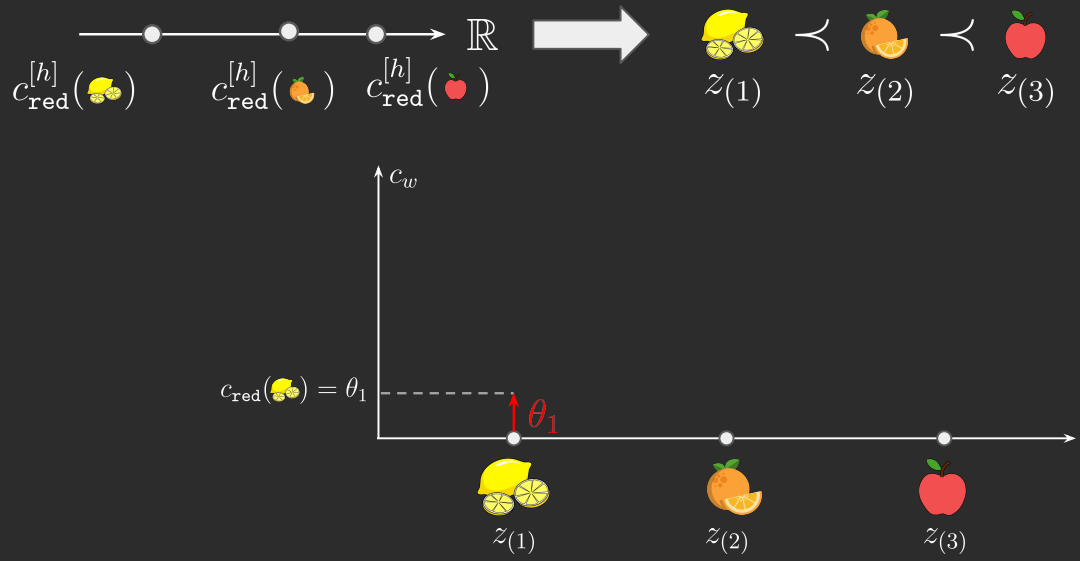


Premise I → Symmetry I → Constraint I → Architecture I

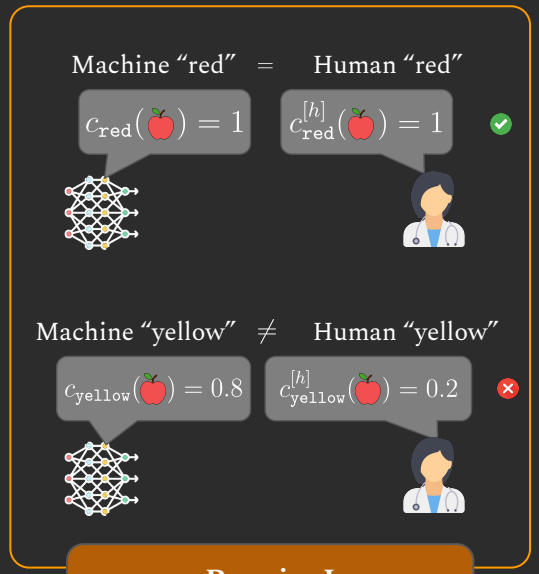


$$\mathbb{I}_{\Delta c_w^{[h]} > 0} \cdot \gamma(-\Delta c_w) = 0$$

Constraint I

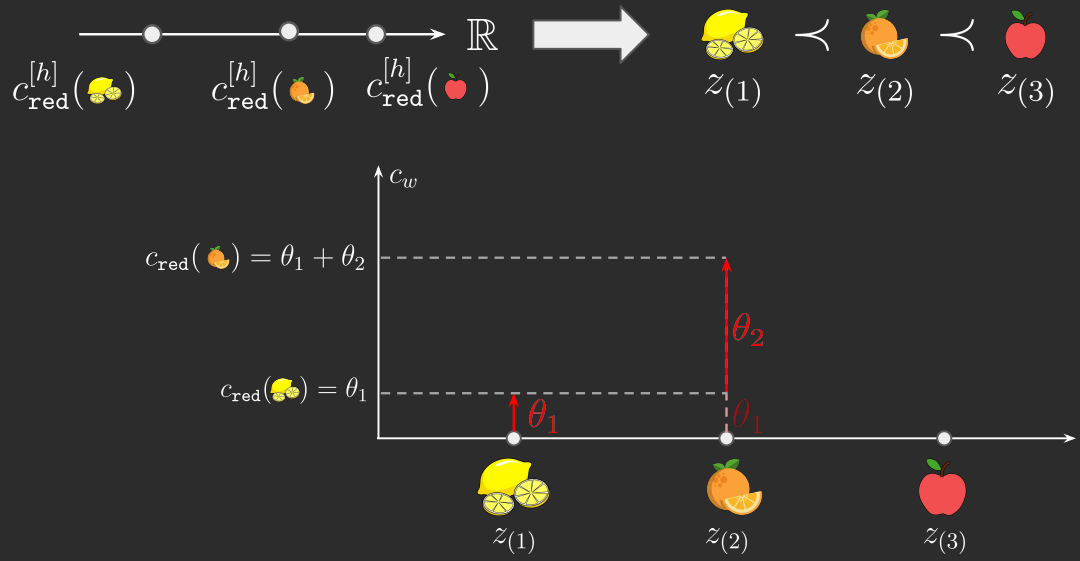


Premise I → Symmetry I → Constraint I → Architecture I

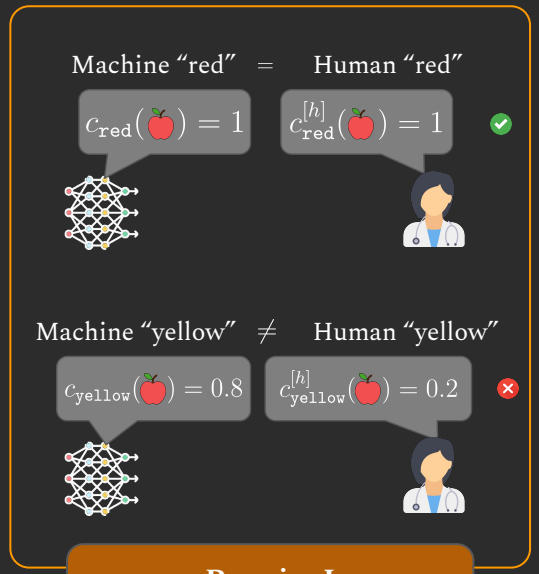


$$\mathbb{I}_{\Delta c_w^{[h]} > 0} \cdot \gamma(-\Delta c_w) = 0$$

Constraint I

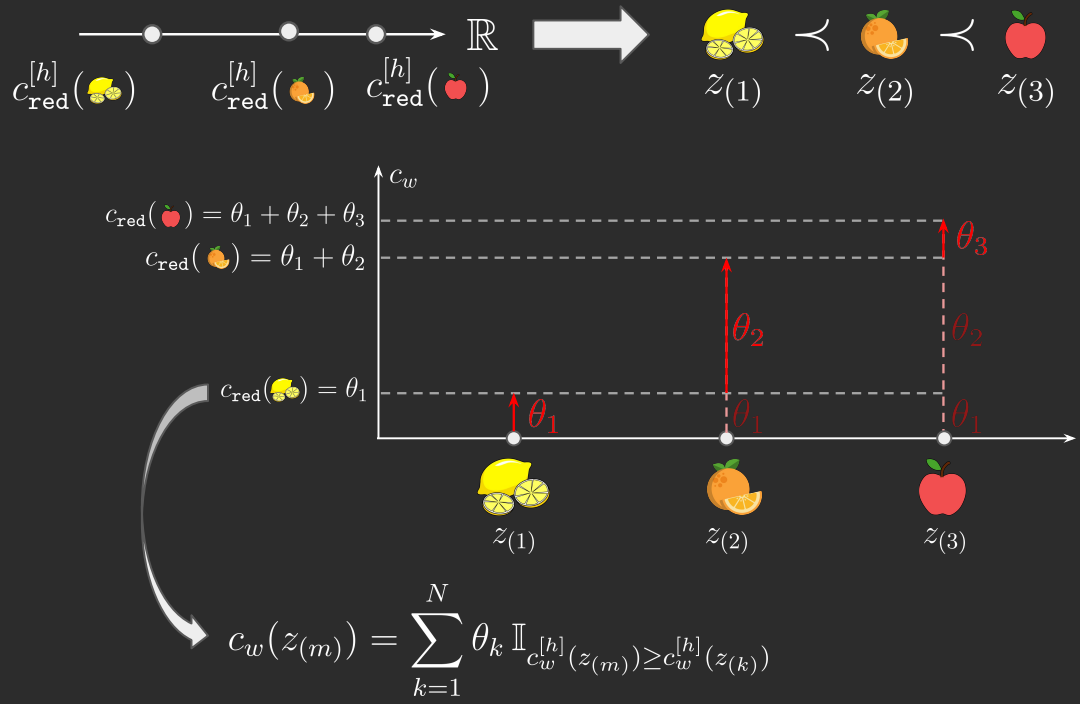


Premise I → Symmetry I → Constraint I → Architecture I

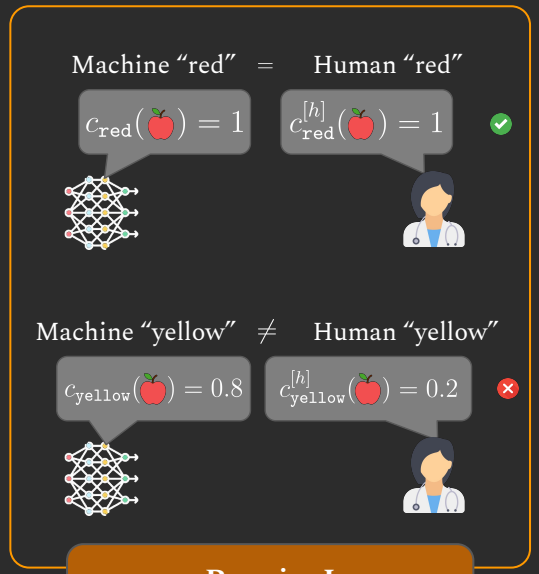


$$\mathbb{I}_{\Delta c_w^{[h]} > 0} \cdot \gamma(-\Delta c_w) = 0$$

Constraint I



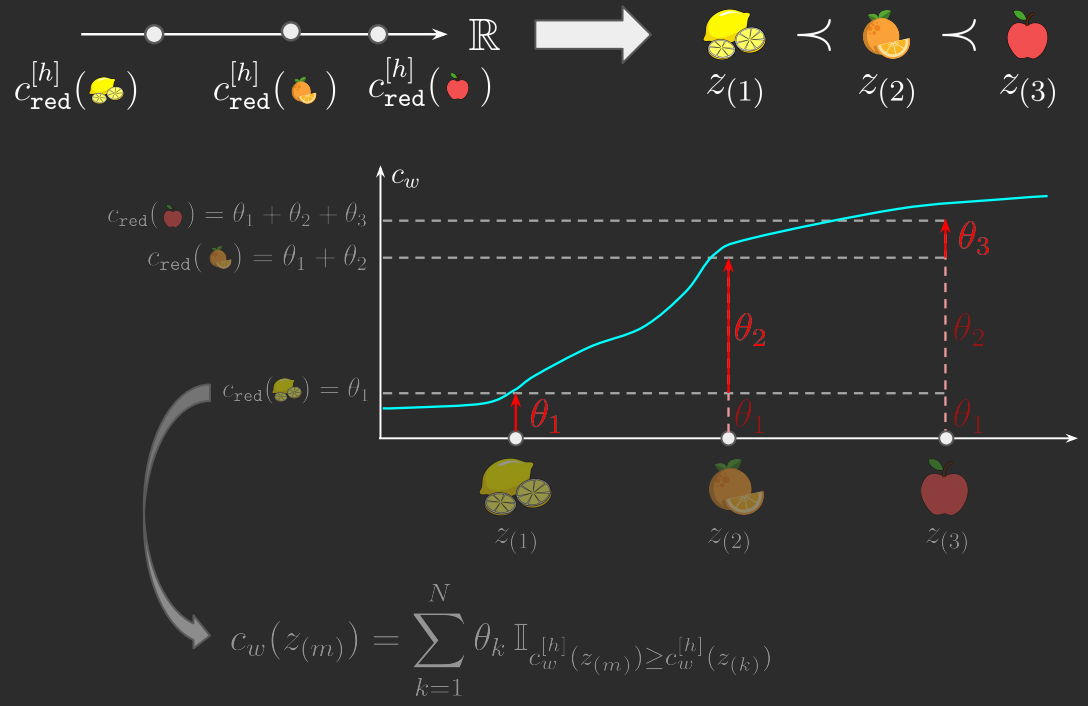
Premise I → Symmetry I → Constraint I → Architecture I



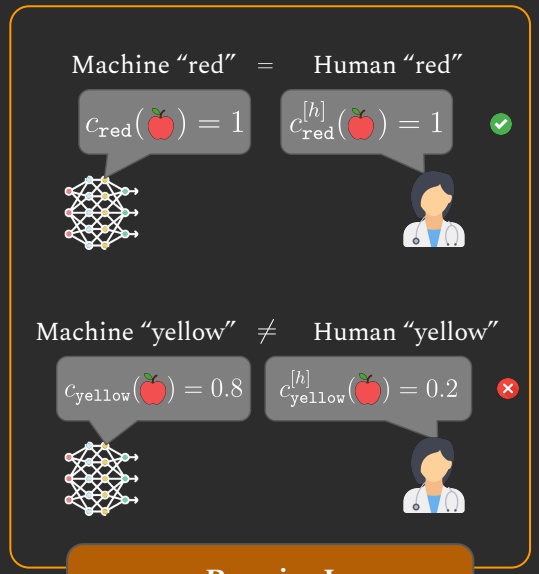
Premise I
Shared concept semantics

$$\mathbb{I}_{\Delta c_w^{[h]} > 0} \cdot \gamma(-\Delta c_w) = 0$$

Constraint I

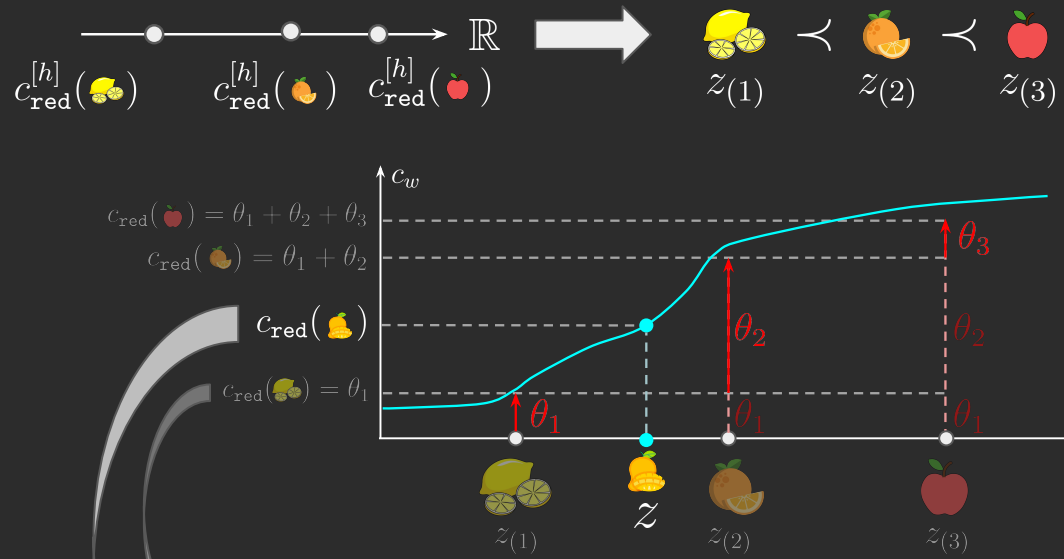


Premise I → Symmetry I → Constraint I → Architecture I



$$\mathbb{I}_{\Delta c_w^{[h]} > 0} \cdot \gamma(-\Delta c_w) = 0$$

Constraint I

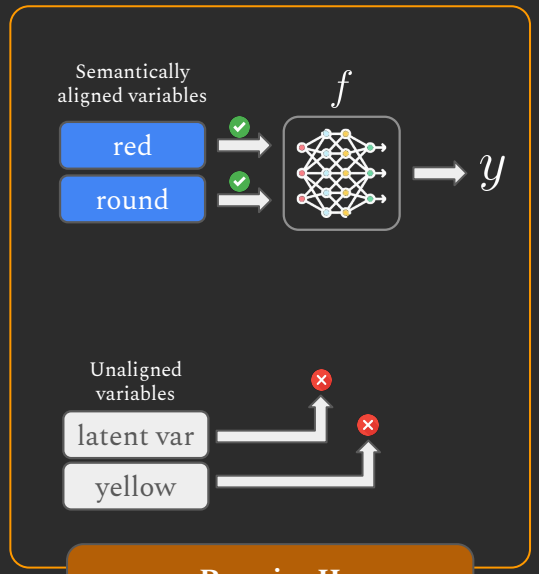


$$c_w(z_{(m)}) = \sum_{k=1}^{N} \theta_k \, \mathbb{I}_{c_w^{[h]}(z_{(m)}) \geq c_w^{[h]}(z_{(k)})}$$

$$c_w(z) = \sum_{i=1}^{N} \beta_i(z) \left(\sum_{k=1}^{N} \theta_k \, \mathbb{I}_{c_w^{[h]}(z_i) \geq c_w^{[h]}(z_k)} \right)$$

"Instance-based model"

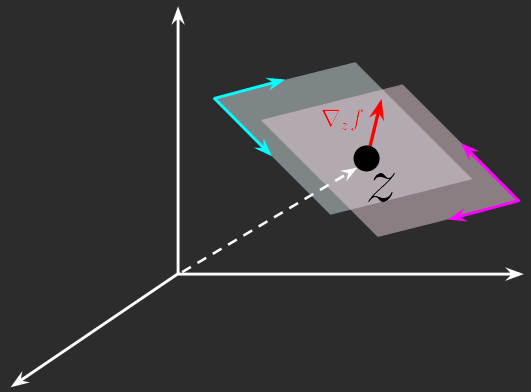
Premise II → Symmetry II → Constraint II → Architecture II



Premise II
Prediction-concept dependency

$$1 - \frac{\|Q_c^\top Q_f\|_F^2}{\operatorname{rank}(\nabla_z f)} = 0$$

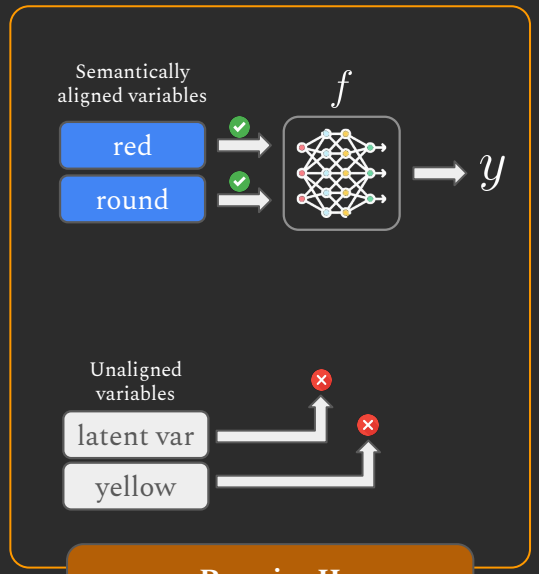
Constraint II



$$f = \phi(c_1, \ldots, c_l)$$

"Concept bottleneck"

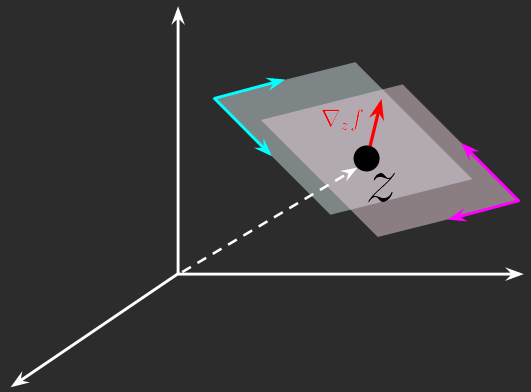
Premise II → Symmetry II → Constraint II → Architecture II



Premise II
Prediction-concept dependency

$$1 - \frac{\|Q_c^\top Q_f\|_F^2}{\operatorname{rank}(\nabla_z f)} = 0$$

Constraint II



$$f = \phi(c_1, \ldots, c_l)$$

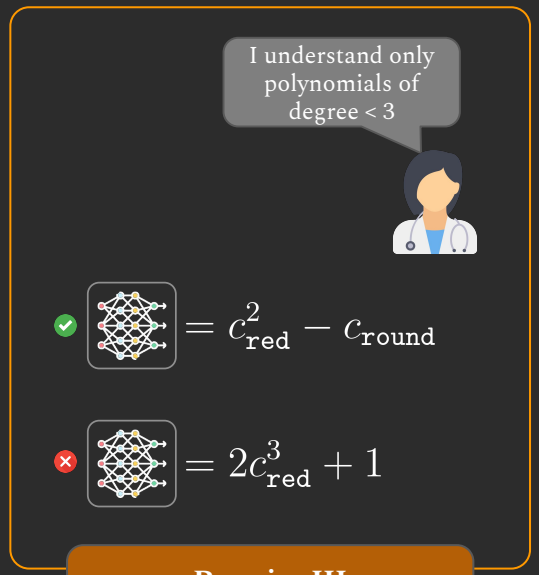
"Concept bottleneck"

Proof using the chain rule

$$\nabla_z(\phi \circ c) = \frac{\partial \phi}{\partial c} \nabla_z c$$

$$f = \phi(c) \iff \forall z \in Z,\ \nabla_z f = \frac{\partial \phi}{\partial c} \nabla_z c$$

Premise III → Symmetry III → Constraint III → Architecture III



$$K^{[h]}(\phi, \ldots, \nabla_c^{(n)}\phi) = 0$$

Constraint III

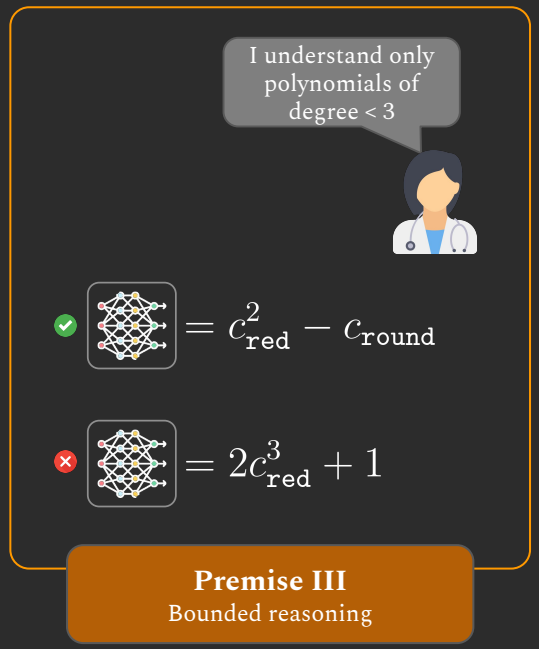
$$K^{[h]}(\phi, \ldots, \nabla_c^{(n)}\phi) = 0$$

Solve the differential equation

$$\phi_\theta(c) = \sum_{k=1}^{n} \theta_k \psi_k(c), \quad \phi_\theta(c) \in \ker(K^{[h]})$$

"General solution"

Premise III → Symmetry III → Constraint III → Architecture III



$$K^{[h]}(\phi, \ldots, \nabla_c^{(n)}\phi) = 0$$

Constraint III

$$K^{[h]}(\phi, \ldots, \nabla_c^{(n)}\phi) = 0$$

Solve the differential equation

$$\phi_\theta(c) = \sum_{k=1}^{n} \theta_k \psi_k(c), \quad \phi_\theta(c) \in \ker(K^{[h]})$$

"General solution"

Case	PDE $K^{[h]}$	1st Order PDE System	General Solution
Linear	$\nabla_c^2\phi = 0$	$\begin{cases}\nabla_c\phi = u_2 \\ \nabla_c u_2 = 0\end{cases}$	$\phi(c) = \theta_1^\top c + \theta_0$
Exponential	$\nabla_c\phi - \theta^\top\phi = 0$	$\nabla_c(\ln\phi) = \theta^\top$	$\phi(c) = e^{\theta^\top c + \theta_0}$
Piece-wise Constant	$\nabla_c\phi = \sum_i \theta_i \delta(c - c_i)$	$\nabla_c\phi = \sum_i \theta_i \delta(c - c_i)$	$\phi(c) = \sum_i \theta_i \sum_{j=1}^{l} H(c_j - c_{i,j}) + \theta_0$
Periodic	$\nabla_c^2\phi + \omega^2\phi = 0$	$\begin{cases}\nabla_c\phi = u_2 \\ \nabla_c u_2 = -\omega^2\phi I_{m\times m}\end{cases}$	$\phi(c) = \theta_1 \cos(k^\top c) + \theta_0 \sin(k^\top c)$

Interpretable architectures

$$c_w(z) = \sum_{i=1}^{N} \beta_i(z) \left(\sum_{k=1}^{N} \theta_k \mathbb{I}_{c_w^{[h]}(z_i) \geq c_w^{[h]}(z_k)} \right) \qquad f = \phi(c_1, \ldots, c_l) \qquad \phi_\theta(c) = \sum_{k=1}^{n} \theta_k \psi_k(c), \quad \phi_\theta(c) \in \ker(K^{[h]})$$

Architectures

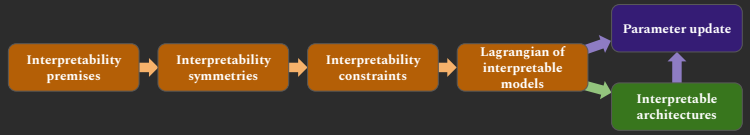
$$\underbrace{f = \phi_\theta^{[h]}}_{\text{Constraint II}} \left(\left[\underbrace{\sum_{i=1}^{N} \beta_i(z) \left(\sum_{k=1}^{N} \theta_k \mathbb{I}_{c_w^{[h]}(z_i) \geq c_w^{[h]}(z_k)} \right)}_{\text{Constraint I}} \right]_{w=1}^{l} \right), \quad \underbrace{\phi_\theta^{[h]} \in \ker(K^{[h]})}_{\text{Constraint III}}$$

General solution

$$L(\theta_{c,\phi}, \mathcal{D}, y, K^{[h]}) = T - \mathcal{L} \left(\phi_\theta^{[h]} \left(\left[\sum_{i=1}^{N} \beta_i(z) \left(\sum_{k=1}^{N} \theta_k \mathbb{I}_{c_w^{[h]}(z_i) \geq c_w^{[h]}(z_k)} \right) \right]_{w=1}^{l} \right), y \right), \ \phi_\theta^{[h]} \in \ker(K^{[h]})$$

Lagrangian

The Standard Interpretable Model

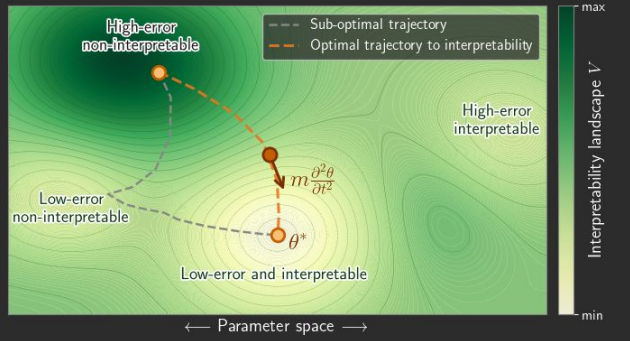


	Shared concept semantics	Prediction-concept dependency	Bounded reasoning
Premises			
Symmetries	$\chi(\mathfrak{g}_w \circ c_w^{[h]}) = \chi(c_w^{[h]})$	$\exists \mathfrak{g}_c \in \mathfrak{G}_c$ such that $\mathfrak{g}_c.(\nabla_z f)^\top = (\nabla_z f)^\top$	$K^{[h]}(\mathfrak{g}(\phi(c)), \ldots, \nabla^{(\nu)}\mathfrak{g}(\phi(c))) = \mu(c, \phi(c))K^{[h]}(\phi, \ldots, \nabla_c^{(\nu)}\phi) = 0$
Constraints	$\mathbb{I}_{\Delta c_w^{[h]}>0} \cdot \gamma(-\Delta c_w) = 0$	$1 - \frac{\|Q_c^\top Q_f\|_F^2}{\text{rank}(\nabla_z f)} = 0$	$K^{[h]}(\phi, \ldots, \nabla_c^{(n)}\phi) = 0$
Lagrangian	$L(\theta_{f,c,\phi}, \mathcal{D}, y, K^{[h]}) = T - V = \underbrace{\sum_{i\in\{f,c,\phi\}} \frac{1}{2} m(\partial_t \theta_i)^T(\partial_t \theta_i)}_{\text{parameter dynamics}} - \underbrace{\mathcal{L}(f(z;\theta_f), y)}_{\text{prediction error}} - \underbrace{\lambda_1 \sum_{w=1}^{m} \sum_{z_i \in \mathcal{D}} \mathbb{I}_{\Delta c_w^{[h]}(z,z_i)>0} \cdot \gamma(-\Delta c_w(z, z_i; \theta_c))}_{\text{Constraint I}} - \underbrace{\lambda_2 \left(1 - \frac{\|Q_c^\top Q_f\|_F^2}{\text{rank}(\nabla_z f)}\right)}_{\text{Constraint II}} - \underbrace{K^{[h]}(\phi(c(z;\theta_c);\theta_\phi), \ldots, \nabla_c^{(n)}\phi(c(z;\theta_c);\theta_\phi))}_{\text{Constraint III}}$		
Equations of motion	$m\frac{\partial^2 \theta_f}{\partial t^2} = -\nabla_{\theta_f}\mathcal{L}(f(z;\theta_f), y) - \lambda_2 \left(1 - \frac{\|Q_c^\top Q_f\|_F^2}{\text{rank}(\nabla_z f)}\right)$	$m\frac{\partial^2 \theta_c}{\partial t^2} = -\lambda_1 \sum_{\substack{w=1,\ldots,m \\ z_i \in \mathcal{D}}} \mathbb{I}_{\Delta c_w^{[h]}>0} \nabla_{\theta_c} \gamma(-\Delta c_w(z, z_i; \theta_c)) - \lambda_2 \nabla_{\theta_c}\left(1 - \frac{\|Q_c^\top Q_f\|_F^2}{\text{rank}(\nabla_z f)}\right) - \lambda_3 \nabla_{\theta_c} K^{[h]}(\phi(c;\theta_\phi), \ldots, \nabla_c^{(n)}\phi(c;\theta_\phi))$	$m\frac{\partial^2 \theta_\phi}{\partial t^2} = -\lambda_3 \nabla_{\theta_\phi} K^{[h]}(\phi(c;\theta_\phi), \ldots, \nabla_c^{(n)}\phi(c;\theta_\phi))$
Architectures	$c_w(z) = \sum_{i=1}^{N} \beta_i(z) \left(\sum_{k=1}^{N} \theta_k \mathbb{I}_{c_w^{[h]}(z_i) \geq c_w^{[h]}(z_k)}\right)$	$f = \phi(c_1, \ldots, c_l)$	$\phi_\theta(c) = \sum_{k=1}^{n} \theta_k \psi_k(c), \quad \phi_\theta(c) \in \ker(K^{[h]})$
General solution		$\underbrace{f = \phi_\theta^{[h]}}_{\text{Constraint II}} \left(\left[\underbrace{\sum_{i=1}^{N} \beta_i(z) \left(\sum_{k=1}^{N} \theta_k \mathbb{I}_{c_w^{[h]}(z_i) \geq c_w^{[h]}(z_k)}\right)}_{\text{Constraint I}} \right]_{w=1}^{l} \right), \quad \underbrace{\phi_\theta^{[h]} \in \ker(K^{[h]})}_{\text{Constraint III}}$	

Interpretability "in the loss"

$$L(\theta_{f,c,\phi}, \mathcal{D}, y, K^{[h]}) = T - V = \underbrace{\sum_{i\in\{f,c,\phi\}} \frac{1}{2} m(\partial_t \theta_i)^T (\partial_t \theta_i)}_{\text{parameter dynamics}} - \underbrace{\mathcal{L}(f(z;\theta_f), y)}_{\text{prediction error}} - \underbrace{\lambda_1 \sum_{w=1}^{m} \sum_{z_i \in \mathcal{D}} \mathbb{I}_{\Delta c_w^{[h]}(z,z_i)>0} \cdot \gamma(-\Delta c_w(z, z_i; \theta_c))}_{\text{Constraint I}} - \underbrace{\lambda_2 \left(1 - \frac{\|Q_c^\top Q_f\|_F^2}{\text{rank}(\nabla_z f)}\right)}_{\text{Constraint II}} - \underbrace{K^{[h]}(\phi(c(z;\theta_c);\theta_\phi), \ldots, \nabla_c^{(m)}\phi(c(z;\theta_c);\theta_\phi))}_{\text{Constraint III}}$$

Lagrangian #1

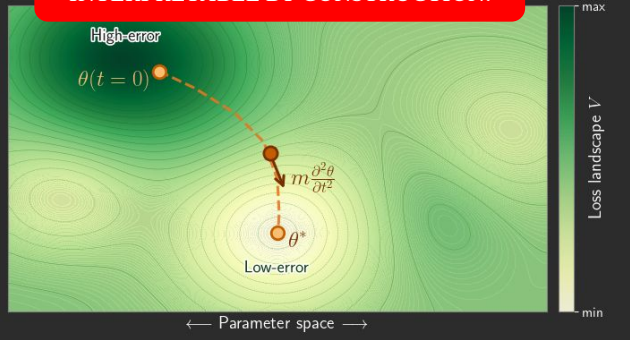


Interpretability "in the architecture"

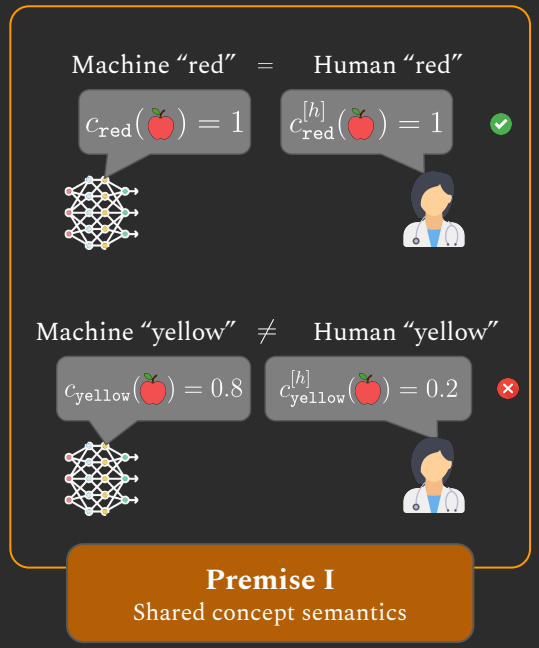
$$L(\theta_{c,\phi}, \mathcal{D}, y, K^{[h]}) = T - \mathcal{L}\left(\phi_\theta^{[h]}\left(\left[\sum_{i=1}^{N} \beta_i(z)\left(\sum_{k=1}^{N} \theta_k \mathbb{I}_{c_w^{[h]}(z_i) \geq c_w^{[h]}(z_k)}\right)\right]_{w=1}^{l}\right), y\right), \quad \phi_\theta^{[h]} \in \ker(K^{[h]})$$

Lagrangian #2

THE MODEL IS ALWAYS PERFECTLY INTERPRETABLE BY CONSTRUCTION!

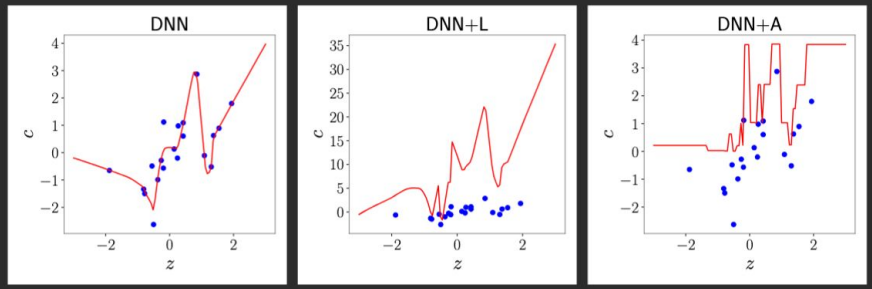


Empirical validation: symmetry I

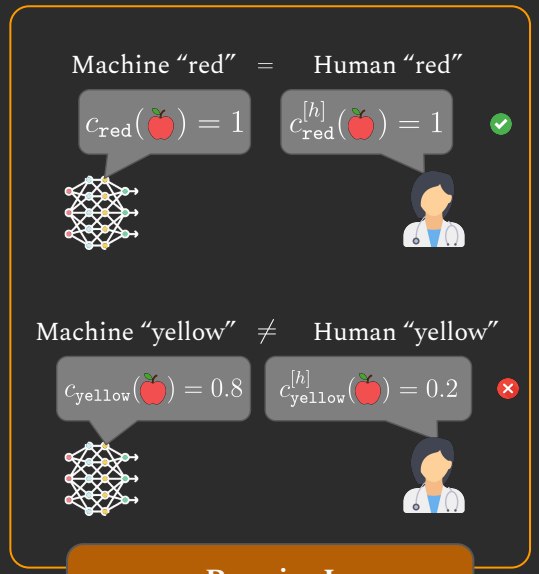


$$\mathbb{I}_{\Delta c_w^{[h]} > 0} \cdot \gamma(-\Delta c_w) = 0$$

Constraint I



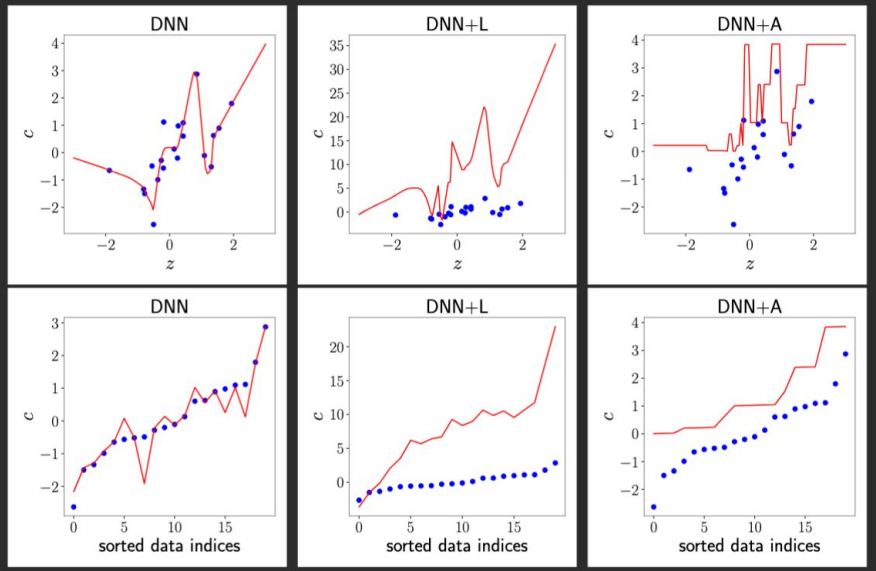
Empirical validation: symmetry I



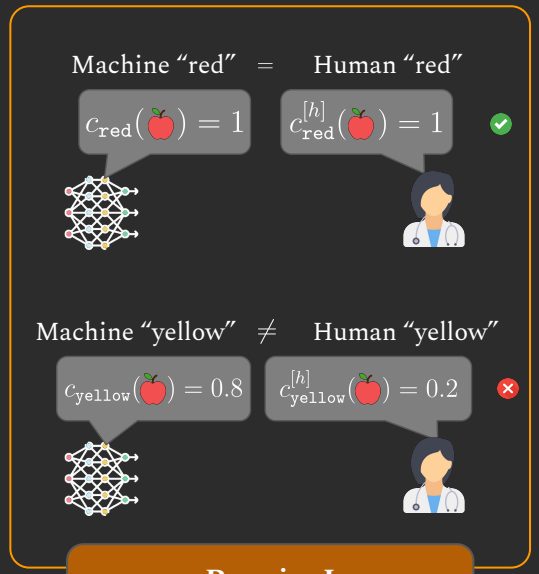
Premise I
Shared concept semantics

$$\mathbb{I}_{\Delta c_w^{[h]} > 0} \cdot \gamma(-\Delta c_w) = 0$$

Constraint I



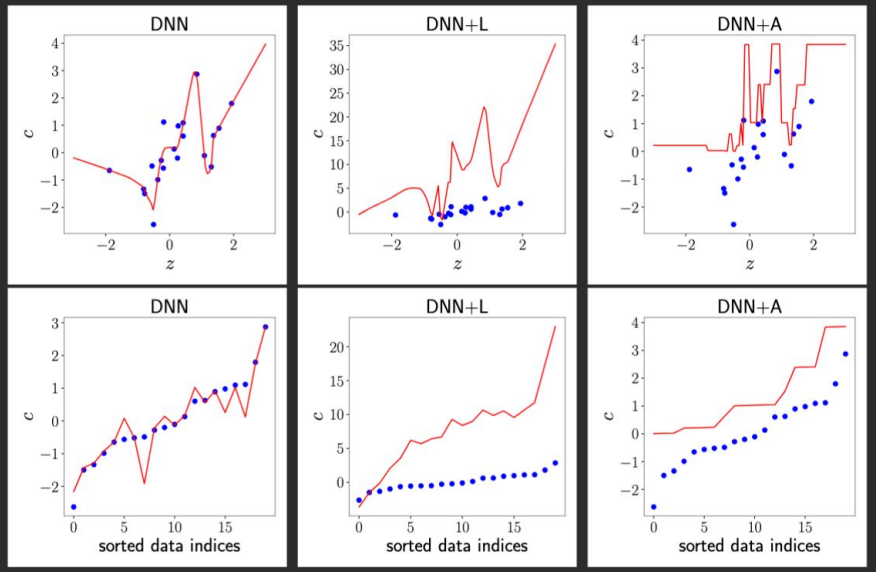
Empirical validation: symmetry I

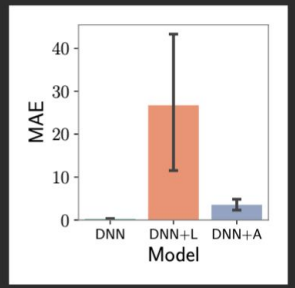


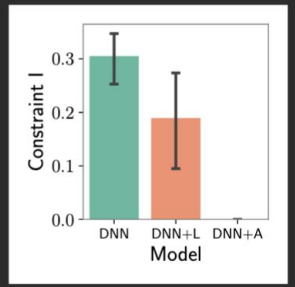
Premise I
Shared concept semantics

$$\mathbb{I}_{\Delta c_w^{[h]} > 0} \cdot \gamma(-\Delta c_w) = 0$$

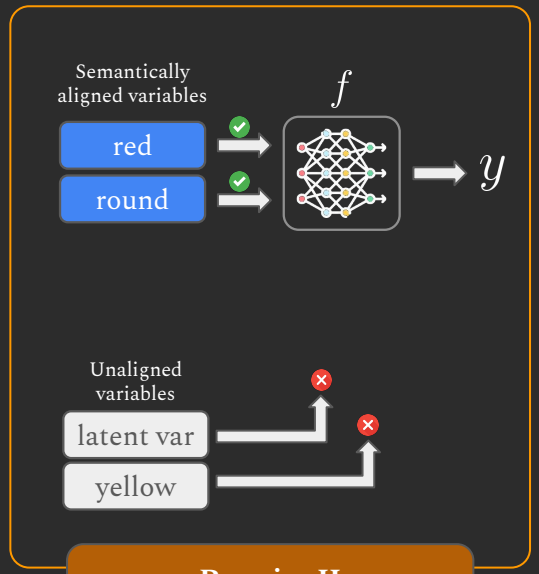
Constraint I







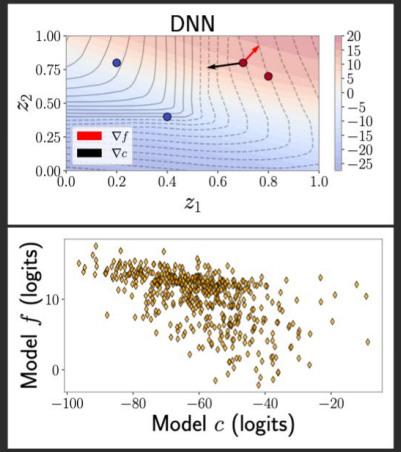
Empirical validation: symmetry II



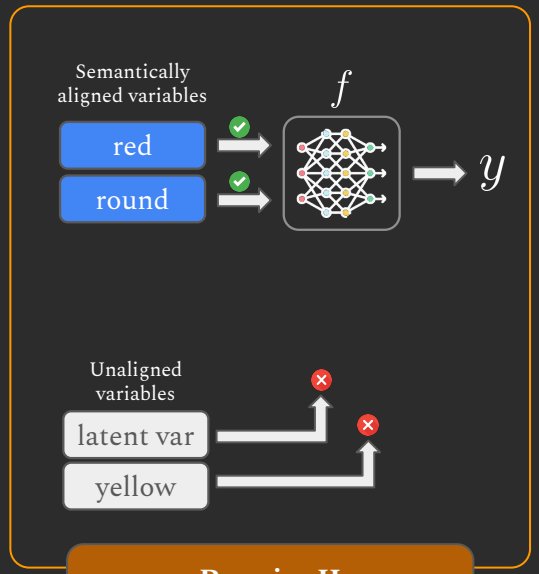
Premise II
Prediction-concept dependency

$$1 - \frac{\|Q_c^\top Q_f\|_F^2}{\operatorname{rank}(\nabla_z f)} = 0$$

Constraint II



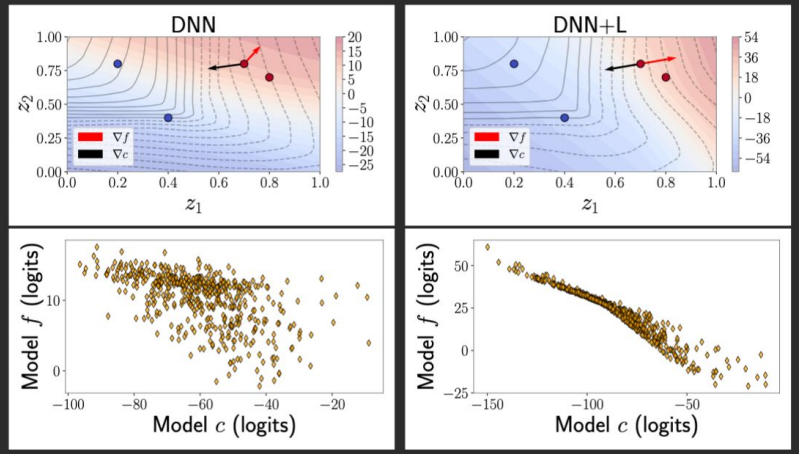
Empirical validation: symmetry II



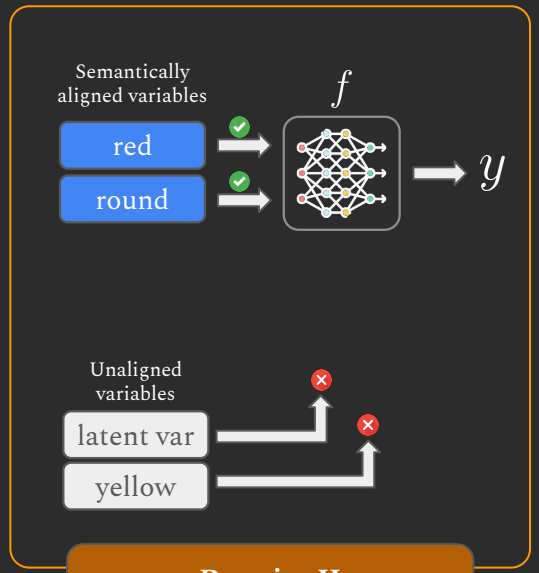
Premise II
Prediction-concept dependency

$$1 - \frac{\|Q_c^\top Q_f\|_F^2}{\operatorname{rank}(\nabla_z f)} = 0$$

Constraint II



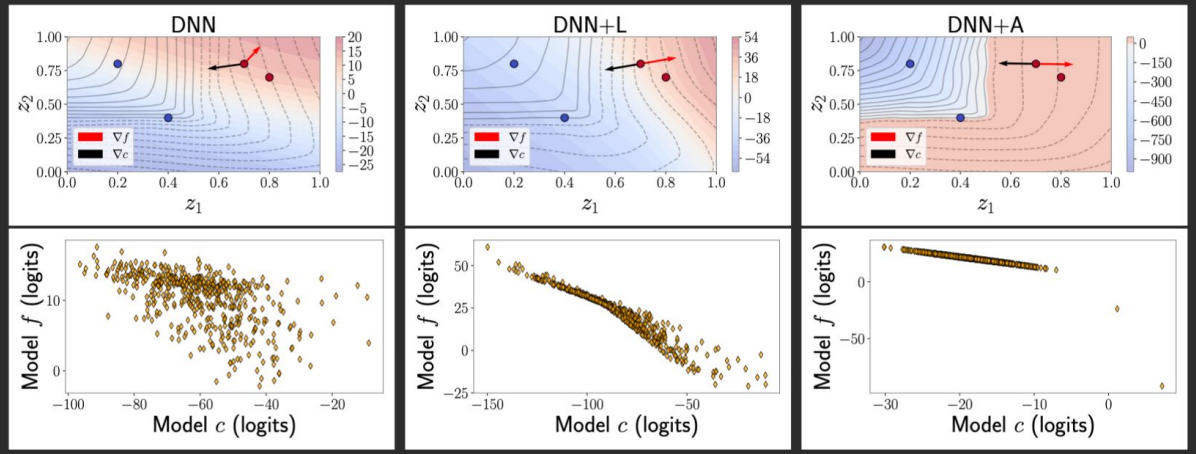
Empirical validation: symmetry II



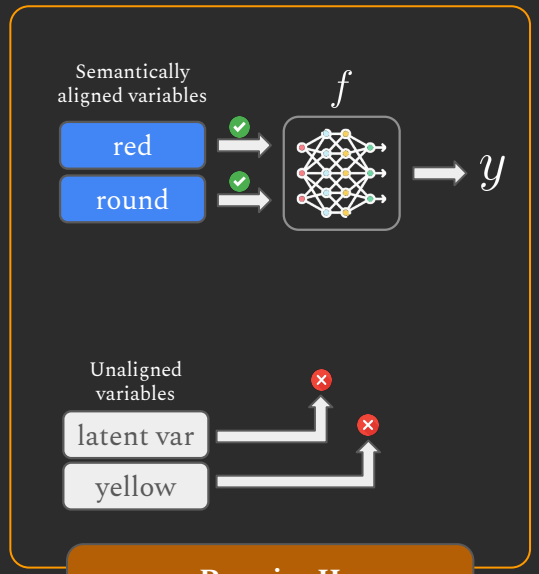
Premise II
Prediction-concept dependency

$$1 - \frac{\|Q_c^\top Q_f\|_F^2}{\mathrm{rank}(\nabla_z f)} = 0$$

Constraint II



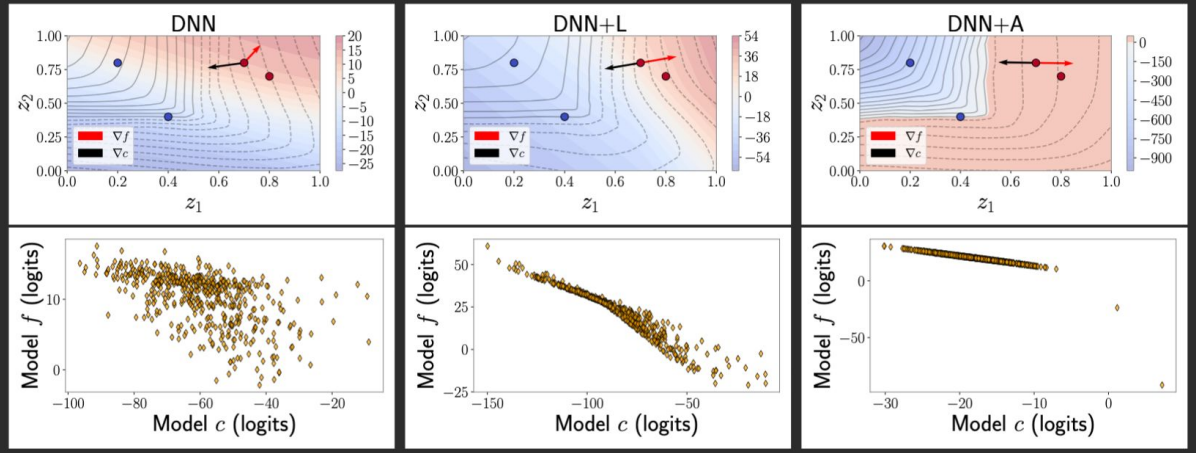
Empirical validation: symmetry II

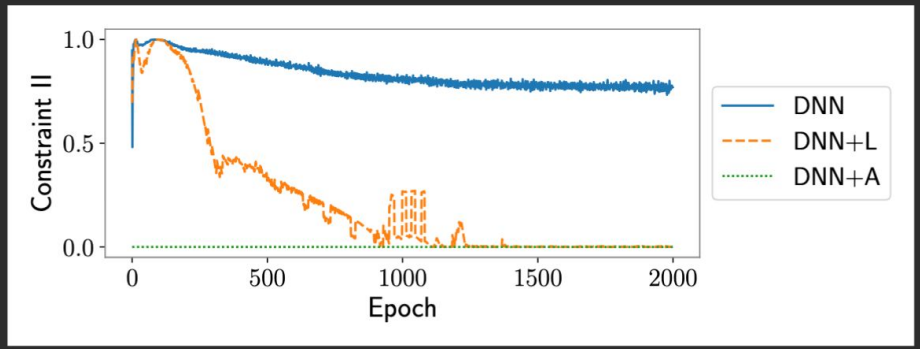


Premise II
Prediction-concept dependency

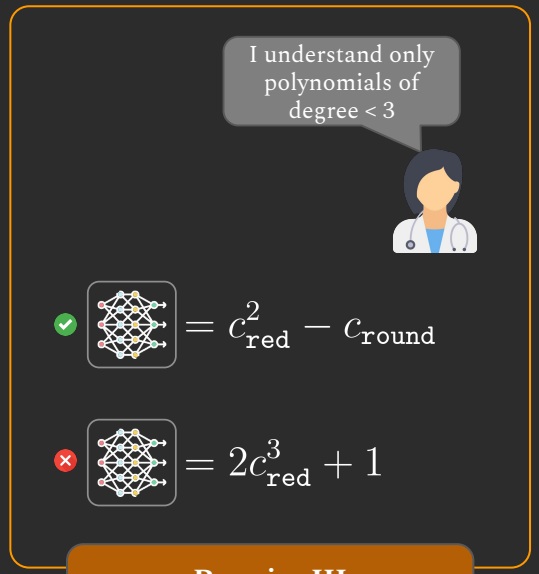
$$1 - \frac{\|Q_c^\top Q_f\|_F^2}{\operatorname{rank}(\nabla_z f)} = 0$$

Constraint II



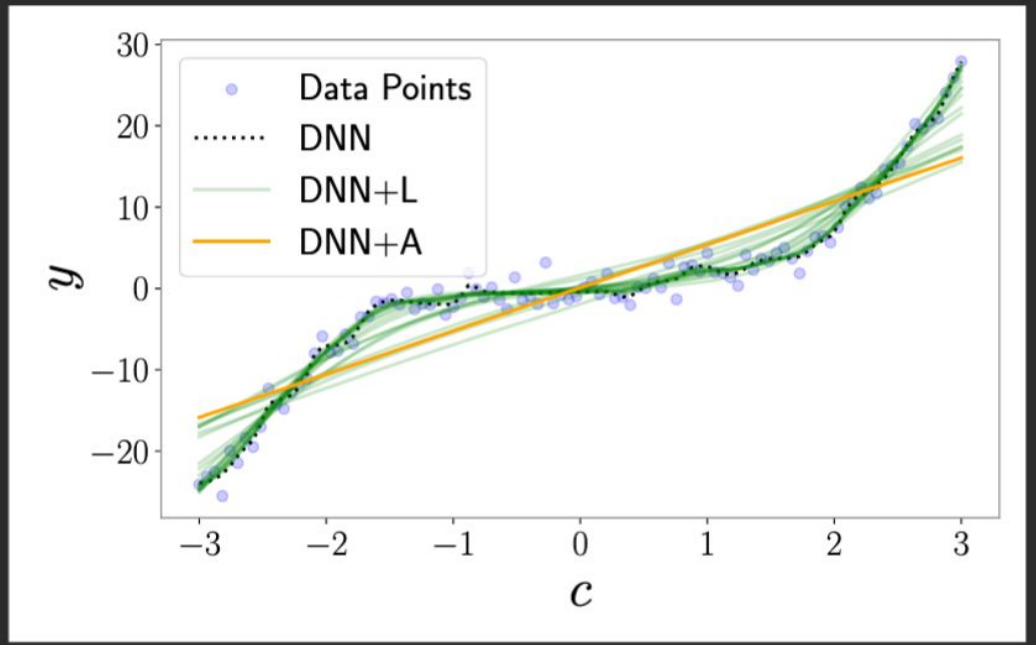


Empirical validation: symmetry III

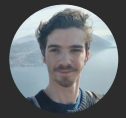


$$K^{[h]}(\phi, \ldots, \nabla_c^{(n)}\phi) = 0$$

Constraint III



The Standard Interpretable Model



Giovanni De Felice

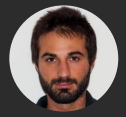
Università della Svizzera Italiana (CH)

Mateo Espinosa Zarlenga

University of Oxford (UK)

Mateja Jamnik

University of Cambridge (UK)

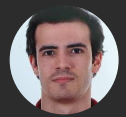


Francesco Giannini

Università di Pisa (IT)

Filippo Bonchi

Università di Pisa (IT)



Ruggero Noris

Politecnico di Torino (IT)

arXiv link

What's next?

Theory

- Definitions
- Research method

Code

- Model design
- Benchmarks

Methods

- Models
- Metrics